Chapter 8

NP and Computational Intractability

We now arrive at a major transition point in the book. Up until now, we've developed efficient algorithms for a wide range of problems and have even made some progress on informally categorizing the problems that admit efficient solutions—for example, problems expressible as minimum cuts in a graph, or problems that allow a dynamic programming formulation. But although we've often paused to take note of other problems that we don't see how to solve, we haven't yet made any attempt to actually quantify or characterize the range of problems that *can't be solved efficiently.*

Back when we were first laying out the fundamental definitions, we settled on polynomial time as our working notion of efficiency. One advantage of using a concrete definition like this, as we noted earlier, is that it gives us the opportunity to prove mathematically that certain problems cannot be solved by polynomial-time—and hence "efficient"—algorithms.

When people began investigating computational complexity in earnest, there was some initial progress in proving that certain *extremely hard* problems cannot be solved by efficient algorithms. But for many of the most fundamental discrete computational problems—arising in optimization, artificial intelligence, combinatorics, logic, and elsewhere—the question was too difficult to resolve, and it has remained open since then: We do not know of polynomial-time algorithms for these problems, and we cannot prove that no polynomial-time algorithm exists.

In the face of this formal ambiguity, which becomes increasingly hardened as years pass, people working in the study of complexity have made significant progress. A large class of problems in this "gray area" has been characterized, and it has been proved that they are equivalent in the following sense: a polynomial-time algorithm for any one of them would imply the existence of a

polynomial-time algorithm for all of them. These are the *NP-complete problems*, a name that will make more sense as we proceed a little further. There are literally thousands of NP-complete problems, arising in numerous areas, and the class seems to contain a large fraction of the fundamental problems whose complexity we can't resolve. So the formulation of NP-completeness, and the proof that all these problems are equivalent, is a powerful thing: it says that all these open questions are really a *single* open question, a single type of complexity that we don't yet fully understand.

From a pragmatic point of view, NP-completeness essentially means "computationally hard for all practical purposes, though we can't prove it." Discovering that a problem is NP-complete provides a compelling reason to stop searching for an efficient algorithm—you might as well search for an efficient algorithm for any of the famous computational problems already known to be NP-complete, for which many people have tried and failed to find efficient algorithms.

8.1 Polynomial-Time Reductions

Our plan is to explore the space of computationally hard problems, eventually arriving at a mathematical characterization of a large class of them. Our basic technique in this exploration is to compare the relative difficulty of different problems; we'd like to formally express statements like, "Problem X is at least as hard as problem Y." We will formalize this through the notion of *reduction*: we will show that a particular problem X is at least as hard as some other problem Y by arguing that, if we had a "black box" capable of solving X, then we could also solve Y. (In other words, X is powerful enough to let us solve Y.)

To make this precise, we add the assumption that X can be solved in polynomial time directly to our model of computation. Suppose we had a *black box* that could solve instances of a problem X; if we write down the input for an instance of X, then in a single step, the black box will return the correct answer. We can now ask the following question:

> *($*) Can arbitrary instances of problem Y be solved using a polynomial number of standard computational steps, plus a polynomial number of calls to a black box that solves problem X?*

If the answer to this question is yes, then we write $Y \leq_P X$; we read this as "Y is polynomial-time reducible to X," or "X is at least as hard as Y (with respect to polynomial time)." Note that in this definition, we still pay for the time it takes to write down the input to the black box solving X, and to read the answer that the black box provides.

This formulation of reducibility is very natural. When we ask about reductions to a problem X, it is as though we've supplemented our computational model with a piece of specialized hardware that solves instances of X in a single step. We can now explore the question: How much extra power does this piece of hardware give us?

An important consequence of our definition of $\leq_P$ is the following. Suppose $Y \leq_P X$ and there actually *exists* a polynomial-time algorithm to solve X. Then our specialized black box for X is actually not so valuable; we can replace it with a polynomial-time algorithm for X. Consider what happens to our algorithm for problem Y that involved a polynomial number of steps plus a polynomial number of calls to the black box. It now becomes an algorithm that involves a polynomial number of steps, plus a polynomial number of calls to a subroutine that runs in polynomial time; in other words, it has become a polynomial-time algorithm. We have therefore proved the following fact.

(8.1) *Suppose $Y \leq_P X$. If X can be solved in polynomial time, then Y can be solved in polynomial time.*

We've made use of precisely this fact, implicitly, at a number of earlier points in the book. Recall that we solved the Bipartite Matching Problem using a polynomial amount of preprocessing plus the solution of a single instance of the Maximum-Flow Problem. Since the Maximum-Flow Problem can be solved in polynomial time, we concluded that Bipartite Matching could as well. Similarly, we solved the foreground/background Image Segmentation Problem using a polynomial amount of preprocessing plus the solution of a single instance of the Minimum-Cut Problem, with the same consequences. Both of these can be viewed as direct applications of (8.1). Indeed, (8.1) summarizes a great way to design polynomial-time algorithms for new problems: by reduction to a problem we already know how to solve in polynomial time.

In this chapter, however, we will be using (8.1) to establish the computational *intractability* of various problems. We will be engaged in the somewhat subtle activity of relating the tractability of problems even when we don't know how to solve *either* of them in polynomial time. For this purpose, we will really be using the contrapositive of (8.1), which is sufficiently valuable that we'll state it as a separate fact.

(8.2) *Suppose $Y \leq_P X$. If Y cannot be solved in polynomial time, then X cannot be solved in polynomial time.*

Statement (8.2) is transparently equivalent to (8.1), but it emphasizes our overall plan: If we have a problem Y that is known to be hard, and we show

that $Y \leq_P X$, then the hardness has "spread" to X; X must be hard or else it could be used to solve Y.

In reality, given that we don't actually know whether the problems we're studying can be solved in polynomial time or not, we'll be using $\leq_P$ to establish relative levels of difficulty among problems.

With this in mind, we now establish some reducibilities among an initial collection of fundamental hard problems.

A First Reduction: Independent Set and Vertex Cover

The Independent Set Problem, which we introduced as one of our five representative problems in Chapter 1, will serve as our first prototypical example of a hard problem. We don't know a polynomial-time algorithm for it, but we also don't know how to prove that none exists.

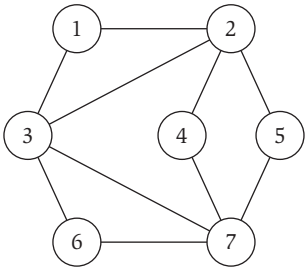

Figure 8.1 A graph whose largest independent set has size 4, and whose smallest vertex cover has size 3.

Let's review the formulation of Independent Set, because we're going to add one wrinkle to it. Recall that in a graph $G = (V, E)$, we say a set of nodes $S \subseteq V$ is *independent* if no two nodes in S are joined by an edge. It is easy to find small independent sets in a graph (for example, a single node forms an independent set); the hard part is to find a large independent set, since you need to build up a large collection of nodes without ever including two neighbors. For example, the set of nodes $\{3, 4, 5\}$ is an independent set of size 3 in the graph in Figure 8.1, while the set of nodes $\{1, 4, 5, 6\}$ is a larger independent set.

In Chapter 1, we posed the problem of finding the *largest* independent set in a graph G. For purposes of our current exploration in terms of reducibility, it will be much more convenient to work with problems that have yes/no answers only, and so we phrase Independent Set as follows.

> *Given a graph G and a number k, does G contain an independent set of size at least k?*

In fact, from the point of view of polynomial-time solvability, there is not a significant difference between the *optimization version* of the problem (find the maximum size of an independent set) and the *decision version* (decide, yes or no, whether G has an independent set of size at least a given k). Given a method to solve the optimization version, we automatically solve the decision version (for any k) as well. But there is also a slightly less obvious converse to this: If we can solve the decision version of Independent Set for every k, then we can also find a maximum independent set. For given a graph G on n nodes, we simply solve the decision version of Independent Set for each k; the largest k for which the answer is "yes" is the size of the largest independent set in G. (And using binary search, we need only solve the decision version

for $O(\log n)$ different values of k.) This simple equivalence between decision and optimization will also hold in the problems we discuss below.

Now, to illustrate our basic strategy for relating hard problems to one another, we consider another fundamental graph problem for which no efficient algorithm is known: *Vertex Cover*. Given a graph $G = (V, E)$, we say that a set of nodes $S \subseteq V$ is a *vertex cover* if every edge $e \in E$ has at least one end in S. Note the following fact about this use of terminology: In a vertex cover, the vertices do the "covering," and the edges are the objects being "covered." Now, it is easy to find large vertex covers in a graph (for example, the full vertex set is one); the hard part is to find small ones. We formulate the Vertex Cover Problem as follows.

> *Given a graph G and a number k, does G contain a vertex cover of size at most k?*

For example, in the graph in Figure 8.1, the set of nodes $\{1, 2, 6, 7\}$ is a vertex cover of size 4, while the set $\{2, 3, 7\}$ is a vertex cover of size 3.

We don't know how to solve either Independent Set or Vertex Cover in polynomial time; but what can we say about their relative difficulty? We now show that they are equivalently hard, by establishing that Independent Set $\leq_P$ Vertex Cover and also that Vertex Cover $\leq_P$ Independent Set. This will be a direct consequence of the following fact.

(8.3) *Let $G = (V, E)$ be a graph. Then S is an independent set if and only if its complement $V - S$ is a vertex cover.*

Proof. First, suppose that S is an independent set. Consider an arbitrary edge $e = (u, v)$. Since S is independent, it cannot be the case that both u and v are in S; so one of them must be in $V - S$. It follows that every edge has at least one end in $V - S$, and so $V - S$ is a vertex cover.

Conversely, suppose that $V - S$ is a vertex cover. Consider any two nodes u and v in S. If they were joined by edge e, then neither end of e would lie in $V - S$, contradicting our assumption that $V - S$ is a vertex cover. It follows that no two nodes in S are joined by an edge, and so S is an independent set. ■

Reductions in each direction between the two problems follow immediately from (8.3).

(8.4) Independent Set $\leq_P$ Vertex Cover.

Proof. If we have a black box to solve Vertex Cover, then we can decide whether G has an independent set of size at least k by asking the black box whether G has a vertex cover of size at most $n - k$. ■

(8.5) Vertex Cover $\leq_P$ Independent Set.

Proof. If we have a black box to solve Independent Set, then we can decide whether G has a vertex cover of size at most k by asking the black box whether G has an independent set of size at least $n - k$. ■

To sum up, this type of analysis illustrates our plan in general: although we don't know how to solve either Independent Set or Vertex Cover efficiently, (8.4) and (8.5) tell us how we could solve either given an efficient solution to the other, and hence these two facts establish the relative levels of difficulty of these problems.

We now pursue this strategy for a number of other problems.

Reducing to a More General Case: Vertex Cover to Set Cover

Independent Set and Vertex Cover represent two different genres of problems. Independent Set can be viewed as a *packing problem*: The goal is to "pack in" as many vertices as possible, subject to conflicts (the edges) that try to prevent one from doing this. Vertex Cover, on the other hand, can be viewed as a *covering problem*: The goal is to parsimoniously "cover" all the edges in the graph using as few vertices as possible.

Vertex Cover is a covering problem phrased specifically in the language of graphs; there is a more general covering problem, *Set Cover*, in which you seek to cover an arbitrary set of objects using a collection of smaller sets. We can phrase Set Cover as follows.

> *Given a set U of n elements, a collection $S_1, \ldots, S_m$ of subsets of U, and a number k, does there exist a collection of at most k of these sets whose union is equal to all of U?*

Imagine, for example, that we have m available pieces of software, and a set U of n *capabilities* that we would like our system to have. The i^{th} piece of software includes the set $S_i \subseteq U$ of capabilities. In the Set Cover Problem, we seek to include a small number of these pieces of software on our system, with the property that our system will then have all n capabilities.

Figure 8.2 shows a sample instance of the Set Cover Problem: The ten circles represent the elements of the underlying set U, and the seven ovals and polygons represent the sets $S_1, S_2, \ldots, S_7$. In this instance, there is a collection

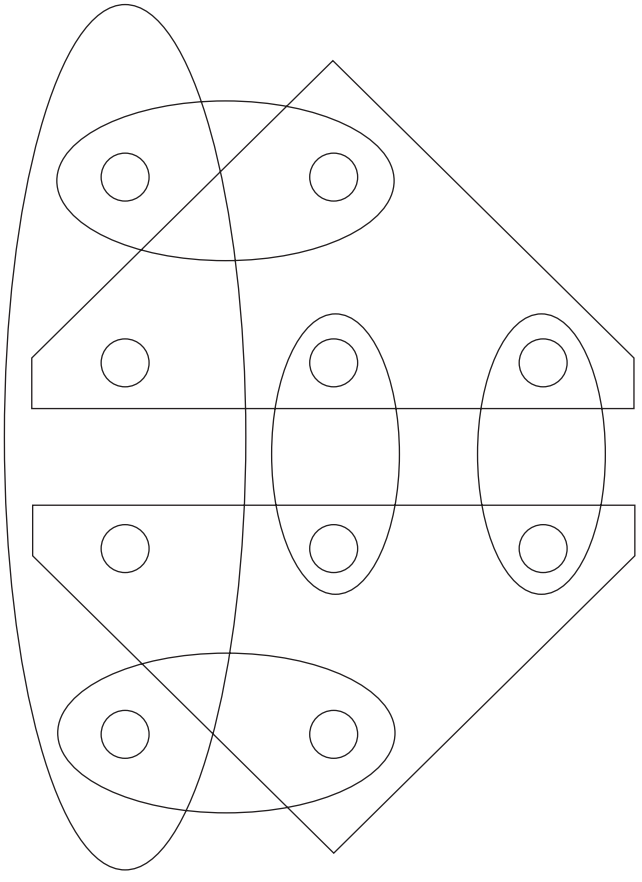

Figure 8.2 An instance of the Set Cover Problem.

of three of the sets whose union is equal to all of U: We can choose the tall thin oval on the left, together with the two polygons.

Intuitively, it feels like Vertex Cover is a special case of Set Cover: in the latter case, we are trying to cover an arbitrary set using arbitrary subsets, while in the former case, we are specifically trying to cover edges of a graph using sets of edges incident to vertices. In fact, we can show the following reduction.

(8.6) Vertex Cover $\leq_P$ Set Cover.

Proof. Suppose we have access to a black box that can solve Set Cover, and consider an arbitrary instance of Vertex Cover, specified by a graph $G = (V, E)$ and a number k. How can we use the black box to help us?

Our goal is to cover the edges in E, so we formulate an instance of Set Cover in which the ground set U is equal to E. Each time we pick a vertex in the Vertex Cover Problem, we cover all the edges incident to it; thus, for each vertex $i \in V$, we add a set $S_i \subseteq U$ to our Set Cover instance, consisting of all the edges in G incident to i.

We now claim that U can be covered with at most k of the sets $S_1, \ldots, S_n$ if and only if G has a vertex cover of size at most k. This can be proved very easily. For if $S_{i_1}, \ldots, S_{i_\ell}$ are $\ell \leq k$ sets that cover U, then every edge in G is incident to one of the vertices $i_1, \ldots, i_\ell$, and so the set $\{i_1, \ldots, i_\ell\}$ is a vertex cover in G of size $\ell \leq k$. Conversely, if $\{i_1, \ldots, i_\ell\}$ is a vertex cover in G of size $\ell \leq k$, then the sets $S_{i_1}, \ldots, S_{i_\ell}$ cover U.

Thus, given our instance of Vertex Cover, we formulate the instance of Set Cover described above, and pass it to our black box. We answer yes if and only if the black box answers yes.

(You can check that the instance of Set Cover pictured in Figure 8.2 is actually the one you'd get by following the reduction in this proof, starting from the graph in Figure 8.1.) ■

Here is something worth noticing, both about this proof and about the previous reductions in (8.4) and (8.5). Although the definition of $\leq_P$ allows us to issue many calls to our black box for Set Cover, we issued only one. Indeed, our algorithm for Vertex Cover consisted simply of encoding the problem as a single instance of Set Cover and then using the answer to this instance as our overall answer. This will be true of essentially all the reductions that we consider; they will consist of establishing $Y \leq_P X$ by transforming our instance of Y to a single instance of X, invoking our black box for X on this instance, and reporting the box's answer as our answer for the instance of Y.

Just as Set Cover is a natural generalization of Vertex Cover, there is a natural generalization of Independent Set as a packing problem for arbitrary sets. Specifically, we define the *Set Packing Problem* as follows.

> *Given a set U of n elements, a collection $S_1, \ldots, S_m$ of subsets of U, and a number k, does there exist a collection of at least k of these sets with the property that no two of them intersect?*

In other words, we wish to "pack" a large number of sets together, with the constraint that no two of them are overlapping.

As an example of where this type of issue might arise, imagine that we have a set U of n non-sharable *resources*, and a set of m software processes. The i^{th} process requires the set $S_i \subseteq U$ of resources in order to run. Then the Set Packing Problem seeks a large collection of these processes that can be run

simultaneously, with the property that none of their resource requirements overlap (i.e., represent a conflict).

There is a natural analogue to (8.6), and its proof is almost the same as well; we will leave the details as an exercise.

(8.7) Independent Set $\leq_P$ Set Packing.

8.2 Reductions via "Gadgets": The Satisfiability Problem

We now introduce a somewhat more abstract set of problems, which are formulated in Boolean notation. As such, they model a wide range of problems in which we need to set decision variables so as to satisfy a given set of constraints; such formalisms are common, for example, in artificial intelligence. After introducing these problems, we will relate them via reduction to the graph- and set-based problems that we have been considering thus far.

The SAT and 3-SAT Problems

Suppose we are given a set X of *n Boolean variables* $x_1, \ldots, x_n$; each can take the value 0 or 1 (equivalently, "false" or "true"). By a *term* over X, we mean one of the variables x_i or its negation $\overline{x_i}$. Finally, a *clause* is simply a disjunction of distinct terms

$$t_1 \vee t_2 \vee \cdots \vee t_\ell.$$

(Again, each $t_i \in \{x_1, x_2, \ldots, x_n, \overline{x_1}, \ldots, \overline{x_n}\}$.) We say the clause has length ℓ if it contains ℓ terms.

We now formalize what it means for an assignment of values to satisfy a collection of clauses. A *truth assignment* for X is an assignment of the value 0 or 1 to each x_i; in other words, it is a function $\nu : X \to \{0, 1\}$. The assignment ν implicitly gives $\overline{x_i}$ the opposite truth value from x_i. An assignment *satisfies* a clause C if it causes C to evaluate to 1 under the rules of Boolean logic; this is equivalent to requiring that at least one of the terms in C should receive the value 1. An assignment satisfies a collection of clauses $C_1, \ldots, C_k$ if it causes all of the C_i to evaluate to 1; in other words, if it causes the conjunction

$$C_1 \wedge C_2 \wedge \cdots \wedge C_k$$

to evaluate to 1. In this case, we will say that ν is a *satisfying assignment* with respect to $C_1, \ldots, C_k$; and that the set of clauses $C_1, \ldots, C_k$ is *satisfiable*.

Here is a simple example. Suppose we have the three clauses

$$(x_1 \vee \overline{x_2}), (\overline{x_1} \vee \overline{x_3}), (x_2 \vee \overline{x_3}).$$

Then the truth assignment v that sets all variables to 1 is not a satisfying assignment, because it does not satisfy the second of these clauses; but the truth assignment v' that sets all variables to 0 is a satisfying assignment.

We can now state the *Satisfiability Problem,* also referred to as *SAT*:

Given a set of clauses $C_1, \ldots, C_k$ over a set of variables $X = \{x_1, \ldots, x_n\}$, does there exist a satisfying truth assignment?

There is a special case of SAT that will turn out to be equivalently difficult and is somewhat easier to think about; this is the case in which all clauses contain exactly three terms (corresponding to distinct variables). We call this problem *3-Satisfiability*, or *3-SAT*:

Given a set of clauses $C_1, \ldots, C_k$, each of length 3, over a set of variables $X = \{x_1, \ldots, x_n\}$, does there exist a satisfying truth assignment?

Satisfiability and 3-Satisfiability are really fundamental combinatorial search problems; they contain the basic ingredients of a hard computational problem in very "bare-bones" fashion. We have to make n independent decisions (the assignments for each x_i) so as to satisfy a set of constraints. There are several ways to satisfy each constraint in isolation, but we have to arrange our decisions so that all constraints are satisfied simultaneously.

Reducing 3-SAT to Independent Set

We now relate the type of computational hardness embodied in SAT and 3-SAT to the superficially different sort of hardness represented by the search for independent sets and vertex covers in graphs. Specifically, we will show that 3-SAT $\leq_P$ Independent Set. The difficulty in proving a thing like this is clear; 3-SAT is about setting Boolean variables in the presence of constraints, while Independent Set is about selecting vertices in a graph. To solve an instance of 3-SAT using a black box for Independent Set, we need a way to encode all these Boolean constraints in the nodes and edges of a graph, so that satisfiability corresponds to the existence of a large independent set.

Doing this illustrates a general principle for designing complex reductions $Y \leq_P X$: building "gadgets" out of components in problem X to represent what is going on in problem Y.

(8.8) 3-SAT $\leq_P$ Independent Set.

Proof. We have a black box for Independent Set and want to solve an instance of 3-SAT consisting of variables $X = \{x_1, \ldots, x_n\}$ and clauses $C_1, \ldots, C_k$.

The key to thinking about the reduction is to realize that there are two conceptually distinct ways of thinking about an instance of 3-SAT.

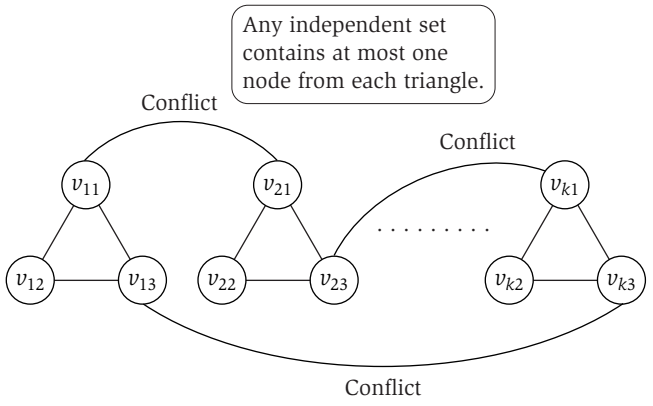

Figure 8.3 The reduction from 3-SAT to Independent Set.

- One way to picture the 3-SAT instance was suggested earlier: You have to make an independent 0/1 decision for each of the n variables, and you succeed if you manage to achieve one of three ways of satisfying each clause.
- A different way to picture the same 3-SAT instance is as follows: You have to choose one term from each clause, and then find a truth assignment that causes all these terms to evaluate to 1, thereby satisfying all clauses. So you succeed if you can select a term from each clause in such a way that no two selected terms "conflict"; we say that two terms *conflict* if one is equal to a variable x_i and the other is equal to its negation $\overline{x_i}$. If we avoid conflicting terms, we can find a truth assignment that makes the selected terms from each clause evaluate to 1.

Our reduction will be based on this second view of the 3-SAT instance; here is how we encode it using independent sets in a graph. First, construct a graph $G = (V, E)$ consisting of $3k$ nodes grouped into k triangles as shown in Figure 8.3. That is, for $i = 1, 2, \ldots, k$, we construct three vertices v_{i1}, v_{i2}, v_{i3} joined to one another by edges. We give each of these vertices a *label*; v_{ij} is labeled with the j^{th} term from the clause C_i of the 3-SAT instance.

Before proceeding, consider what the independent sets of size k look like in this graph: Since two vertices cannot be selected from the same triangle, they consist of all ways of choosing one vertex from each of the triangles. This is implementing our goal of choosing a term in each clause that will evaluate to 1; but we have so far not prevented ourselves from choosing two terms that conflict.

We encode conflicts by adding some more edges to the graph: For each pair of vertices whose labels correspond to terms that conflict, we add an edge between them. Have we now destroyed all the independent sets of size k, or does one still exist? It's not clear; it depends on whether we can still select one node from each triangle so that no conflicting pairs of vertices are chosen. But this is precisely what the 3-SAT instance required.

Let's claim, precisely, that the original 3-SAT instance is satisfiable if and only if the graph G we have constructed has an independent set of size at least k. First, if the 3-SAT instance is satisfiable, then each triangle in our graph contains at least one node whose label evaluates to 1. Let S be a set consisting of one such node from each triangle. We claim S is independent; for if there were an edge between two nodes $u, v \in S$, then the labels of u and v would have to conflict; but this is not possible, since they both evaluate to 1.

Conversely, suppose our graph G has an independent set S of size at least k. Then, first of all, the size of S is exactly k, and it must consist of one node from each triangle. Now, we claim that there is a truth assignment ν for the variables in the 3-SAT instance with the property that the labels of all nodes in S evaluate to 1. Here is how we could construct such an assignment ν. For each variable x_i, if neither x_i nor $\overline{x_i}$ appears as a label of a node in S, then we arbitrarily set $\nu(x_i) = 1$. Otherwise, exactly one of x_i or $\overline{x_i}$ appears as a label of a node in S; for if one node in S were labeled x_i and another were labeled $\overline{x_i}$, then there would be an edge between these two nodes, contradicting our assumption that S is an independent set. Thus, if x_i appears as a label of a node in S, we set $\nu(x_i) = 1$, and otherwise we set $\nu(x_i) = 0$. By constructing ν in this way, all labels of nodes in S will evaluate to 1.

Since G has an independent set of size at least k if and only if the original 3-SAT instance is satisfiable, the reduction is complete. ■

Some Final Observations: Transitivity of Reductions

We've now seen a number of different hard problems, of various flavors, and we've discovered that they are closely related to one another. We can infer a number of additional relationships using the following fact: $\leq_P$ is a *transitive* relation.

(8.9) *If $Z \leq_P Y$, and $Y \leq_P X$, then $Z \leq_P X$.*

Proof. Given a black box for X, we show how to solve an instance of Z; essentially, we just compose the two algorithms implied by $Z \leq_P Y$ and $Y \leq_P X$. We run the algorithm for Z using a black box for Y; but each time the black box for Y is called, we *simulate* it in a polynomial number of steps using the algorithm that solves instances of Y using a black box for X. ■

Transitivity can be quite useful. For example, since we have proved

$$\text{3-SAT} \leq_P \text{Independent Set} \leq_P \text{Vertex Cover} \leq_P \text{Set Cover},$$

we can conclude that 3-SAT $\leq_P$ Set Cover.

8.3 Efficient Certification and the Definition of NP

Reducibility among problems was the first main ingredient in our study of computational intractability. The second ingredient is a characterization of the class of problems that we are dealing with. Combining these two ingredients, together with a powerful theorem of Cook and Levin, will yield some surprising consequences.

Recall that in Chapter 1, when we first encountered the Independent Set Problem, we asked: Can we say anything *good* about it, from a computational point of view? And, indeed, there was something: If a graph does contain an independent set of size at least k, then we could give you an easy proof of this fact by exhibiting such an independent set. Similarly, if a 3-SAT instance is satisfiable, we can prove this to you by revealing the satisfying assignment. It may be an enormously difficult task to actually *find* such an assignment; but if we've done the hard work of finding one, it's easy for you to plug it into the clauses and check that they are all satisfied.

The issue here is the contrast between *finding* a solution and *checking* a proposed solution. For Independent Set or 3-SAT, we do not know a polynomial-time algorithm to find solutions; but *checking* a proposed solution to these problems can be easily done in polynomial time. To see that this is not an entirely trivial issue, consider the problem we'd face if we had to prove that a 3-SAT instance was *not* satisfiable. What "evidence" could we show that would convince you, in polynomial time, that the instance was unsatisfiable?

Problems and Algorithms

This will be the crux of our characterization; we now proceed to formalize it. The input to a computational problem will be encoded as a finite binary string s. We denote the length of a string s by $|s|$. We will identify a decision problem X with the *set* of strings on which the answer is "yes." An algorithm A for a decision problem receives an input string s and returns the value "yes" or "no"—we will denote this returned value by $A(s)$. We say that A *solves* the problem X if for all strings s, we have $A(s) = \texttt{yes}$ if and only if $s \in X$.

As always, we say that A has a *polynomial running time* if there is a polynomial function $p(\cdot)$ so that for every input string s, the algorithm A terminates on s in at most $O(p(|s|))$ steps. Thus far in the book, we have been concerned with problems solvable in polynomial time. In the notation

above, we can express this as the set $\mathcal{P}$ of all problems X for which there exists an algorithm A with a polynomial running time that solves X.

Efficient Certification

Now, how should we formalize the idea that a solution to a problem can be *checked* efficiently, independently of whether it can be solved efficiently? A "checking algorithm" for a problem X has a different structure from an algorithm that actually seeks to solve the problem; in order to "check" a solution, we need the input string s, as well as a separate "certificate" string t that contains the evidence that s is a "yes" instance of X.

Thus we say that B is an *efficient certifier* for a problem X if the following properties hold.

- B is a polynomial-time algorithm that takes two input arguments s and t.
- There is a polynomial function p so that for every string s, we have $s \in X$ if and only if there exists a string t such that $|t| \leq p(|s|)$ and $B(s,t) = \texttt{yes}$.

It takes some time to really think through what this definition is saying. One should view an efficient certifier as approaching a problem X from a "managerial" point of view. It will not actually try to decide whether an input s belongs to X on its own. Rather, it is willing to efficiently evaluate proposed "proofs" t that s belongs to X—provided they are not too long—and it is a correct algorithm in the weak sense that s belongs to X if and only if there exists a proof that will convince it.

An efficient certifier B can be used as the core component of a "brute-force" algorithm for a problem X: On an input s, try all strings t of length $\leq p(|s|)$, and see if $B(s,t) = \texttt{yes}$ for any of these strings. But the existence of B does not provide us with any clear way to design an efficient algorithm that actually solves X; after all, it is still up to us to *find* a string t that will cause $B(s,t)$ to say "yes," and there are exponentially many possibilities for t.

NP: A Class of Problems

We define $\mathcal{NP}$ to be the set of all problems for which there exists an efficient certifier.[1] Here is one thing we can observe immediately.

(8.10) $\mathcal{P} \subseteq \mathcal{NP}$.

[1] The act of searching for a string t that will cause an efficient certifier to accept the input s is often viewed as a *nondeterministic search* over the space of possible proofs t; for this reason, $\mathcal{NP}$ was named as an acronym for "nondeterministic polynomial time."

Proof. Consider a problem $X \in \mathcal{P}$; this means that there is a polynomial-time algorithm A that solves X. To show that $X \in \mathcal{NP}$, we must show that there is an efficient certifier B for X.

This is very easy; we design B as follows. When presented with the input pair (s, t), the certifier B simply returns the value $A(s)$. (Think of B as a very "hands-on" manager that ignores the proposed proof t and simply solves the problem on its own.) Why is B an efficient certifier for X? Clearly it has polynomial running time, since A does. If a string $s \in X$, then for every t of length at most $p(|s|)$, we have $B(s, t) = \texttt{yes}$. On the other hand, if $s \notin X$, then for every t of length at most $p(|s|)$, we have $B(s, t) = \texttt{no}$. ■

We can easily check that the problems introduced in the first two sections belong to $\mathcal{NP}$: it is a matter of determining how an efficient certifier for each of them will make use of a "certificate" string t. For example:

- For the 3-Satisfiability Problem, the certificate t is an assignment of truth values to the variables; the certifier B evaluates the given set of clauses with respect to this assignment.
- For the Independent Set Problem, the certificate t is the identity of a set of at least k vertices; the certifier B checks that, for these vertices, no edge joins any pair of them.
- For the Set Cover Problem, the certificate t is a list of k sets from the given collection; the certifier checks that the union of these sets is equal to the underlying set U.

Yet we cannot prove that any of these problems require more than polynomial time to solve. Indeed, we cannot prove that there is any problem in $\mathcal{NP}$ that does not belong to $\mathcal{P}$. So in place of a concrete theorem, we can only ask a question:

(8.11) *Is there a problem in $\mathcal{NP}$ that does not belong to $\mathcal{P}$? Does $\mathcal{P} = \mathcal{NP}$?*

The question of whether $\mathcal{P} = \mathcal{NP}$ is fundamental in the area of algorithms, and it is one of the most famous problems in computer science. The general belief is that $\mathcal{P} \neq \mathcal{NP}$—and this is taken as a working hypothesis throughout the field—but there is not a lot of hard technical evidence for it. It is more based on the sense that $\mathcal{P} = \mathcal{NP}$ would be too amazing to be true. How could there be a general transformation from the task of *checking* a solution to the much harder task of actually *finding* a solution? How could there be a general means for designing efficient algorithms, powerful enough to handle all these hard problems, that we have somehow failed to discover? More generally, a huge amount of effort has gone into failed attempts at designing polynomial-time algorithms for hard problems in $\mathcal{NP}$; perhaps the most natural explanation

for this consistent failure is that these problems simply cannot be solved in polynomial time.

8.4 NP-Complete Problems

In the absence of progress on the $\mathcal{P} = \mathcal{NP}$ question, people have turned to a related but more approachable question: What are the hardest problems in $\mathcal{NP}$? Polynomial-time reducibility gives us a way of addressing this question and gaining insight into the structure of $\mathcal{NP}$.

Arguably the most natural way to define a "hardest" problem X is via the following two properties: (i) $X \in \mathcal{NP}$; and (ii) for all $Y \in \mathcal{NP}$, $Y \leq_P X$. In other words, we require that every problem in $\mathcal{NP}$ can be reduced to X. We will call such an X an *NP-complete* problem.

The following fact helps to further reinforce our use of the term *hardest*.

(8.12) *Suppose X is an NP-complete problem. Then X is solvable in polynomial time if and only if $\mathcal{P} = \mathcal{NP}$.*

Proof. Clearly, if $\mathcal{P} = \mathcal{NP}$, then X can be solved in polynomial time since it belongs to $\mathcal{NP}$. Conversely, suppose that X can be solved in polynomial time. If Y is any other problem in $\mathcal{NP}$, then $Y \leq_P X$, and so by (8.1), it follows that Y can be solved in polynomial time. Hence $\mathcal{NP} \subseteq \mathcal{P}$; combined with (8.10), we have the desired conclusion. ■

A crucial consequence of (8.12) is the following: If there is *any* problem in $\mathcal{NP}$ that cannot be solved in polynomial time, then no NP-complete problem can be solved in polynomial time.

Circuit Satisfiability: A First NP-Complete Problem

Our definition of NP-completeness has some very nice properties. But before we get too carried away in thinking about this notion, we should stop to notice something: it is not at all obvious that NP-complete problems should even *exist*. Why couldn't there exist two incomparable problems X' and X'', so that there is no $X \in \mathcal{NP}$ with the property that $X' \leq_P X$ and $X'' \leq_P X$? Why couldn't there exist an infinite sequence of problems $X_1, X_2, X_3, \ldots$ in $\mathcal{NP}$, each strictly harder than the previous one? To prove a problem is NP-complete, one must show how it could encode *any* problem in $\mathcal{NP}$. This is a much trickier matter than what we encountered in Sections 8.1 and 8.2, where we sought to encode specific, individual problems in terms of others.

In 1971, Cook and Levin independently showed how to do this for very natural problems in $\mathcal{NP}$. Maybe the most natural problem choice for a first NP-complete problem is the following *Circuit Satisfiability Problem.*

To specify this problem, we need to make precise what we mean by a *circuit*. Consider the standard Boolean operators that we used to define the Satisfiability Problem: $\wedge$ (AND), $\vee$ (OR), and $\neg$ (NOT). Our definition of a circuit is designed to represent a physical circuit built out of gates that implement these operators. Thus we define a circuit K to be a labeled, directed acyclic graph such as the one shown in the example of Figure 8.4.

- The *sources* in K (the nodes with no incoming edges) are labeled either with one of the constants 0 or 1, or with the name of a distinct variable. The nodes of the latter type will be referred to as the *inputs* to the circuit.
- Every other node is labeled with one of the Boolean operators $\wedge$, $\vee$, or $\neg$; nodes labeled with $\wedge$ or $\vee$ will have two incoming edges, and nodes labeled with $\neg$ will have one incoming edge.
- There is a single node with no outgoing edges, and it will represent the *output:* the result that is computed by the circuit.

A circuit computes a function of its inputs in the following natural way. We imagine the edges as "wires" that carry the 0/1 value at the node they emanate from. Each node v other than the sources will take the values on its incoming edge(s) and apply the Boolean operator that labels it. The result of this $\wedge$, $\vee$, or $\neg$ operation will be passed along the edge(s) leaving v. The overall value computed by the circuit will be the value computed at the output node.

For example, consider the circuit in Figure 8.4. The leftmost two sources are preassigned the values 1 and 0, and the next three sources constitute the

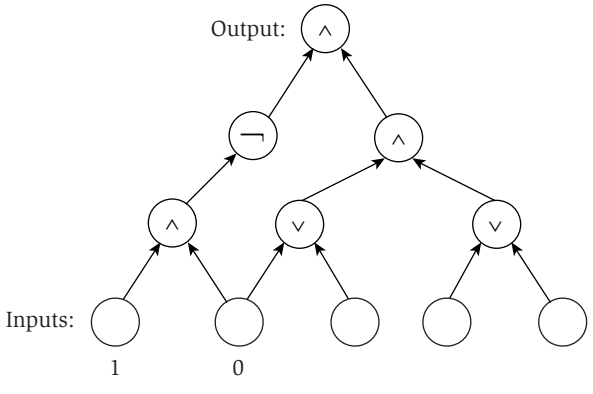

Figure 8.4 A circuit with three inputs, two additional sources that have assigned truth values, and one output.

inputs. If the inputs are assigned the values 1, 0, 1 from left to right, then we get values 0, 1, 1 for the gates in the second row, values 1, 1 for the gates in the third row, and the value 1 for the output.

Now, the Circuit Satisfiability Problem is the following. We are given a circuit as input, and we need to decide whether there is an assignment of values to the inputs that causes the output to take the value 1. (If so, we will say that the given circuit is *satisfiable*, and a *satisfying assignment* is one that results in an output of 1.) In our example, we have just seen—via the assignment 1, 0, 1 to the inputs—that the circuit in Figure 8.4 is satisfiable.

We can view the theorem of Cook and Levin as saying the following.

(8.13) Circuit Satisfiability *is NP-complete.*

As discussed above, the proof of (8.13) requires that we consider an arbitrary problem X in $\mathcal{NP}$, and show that $X \leq_P$ Circuit Satisfiability. We won't describe the proof of (8.13) in full detail, but it is actually not so hard to follow the basic idea that underlies it. We use the fact that any algorithm that takes a fixed number n of bits as input and produces a yes/no answer can be represented by a circuit of the type we have just defined: This circuit is equivalent to the algorithm in the sense that its output is 1 on precisely the inputs for which the algorithm outputs `yes`. Moreover, if the algorithm takes a number of steps that is polynomial in n, then the circuit has polynomial size. This transformation from an algorithm to a circuit is the part of the proof of (8.13) that we won't go into here, though it is quite natural given the fact that algorithms implemented on physical computers can be reduced to their operations on an underlying set of $\wedge$, $\vee$, and $\neg$ gates. (Note that fixing the number of input bits is important, since it reflects a basic distinction between algorithms and circuits: an algorithm typically has no trouble dealing with different inputs of varying lengths, but a circuit is structurally hard-coded with the size of the input.)

How should we use this relationship between algorithms and circuits? We are trying to show that $X \leq_P$ Circuit Satisfiability—that is, given an input s, we want to decide whether $s \in X$ using a black box that can solve instances of Circuit Satisfiability. Now, all we know about X is that it has an efficient certifier $B(\cdot, \cdot)$. So to determine whether $s \in X$, for some specific input s of length n, we need to answer the following question: Is there a t of length $p(n)$ so that $B(s, t) =$ `yes`?

We will answer this question by appealing to a black box for Circuit Satisfiability as follows. Since we only care about the answer for a specific input s, we view $B(\cdot, \cdot)$ as an algorithm on $n + p(n)$ bits (the input s and the

certificate t), and we convert it to a polynomial-size circuit K with $n + p(n)$ sources. The first n sources will be hard-coded with the values of the bits in s, and the remaining $p(n)$ sources will be labeled with variables representing the bits of t; these latter sources will be the inputs to K.

Now we simply observe that $s \in X$ if and only if there is a way to set the input bits to K so that the circuit produces an output of 1—in other words, if and only if K is satisfiable. This establishes that $X \leq_P$ *Circuit Satisfiability*, and completes the proof of (8.13).

An Example To get a better sense for what's going on in the proof of (8.13), we consider a simple, concrete example. Suppose we have the following problem.

Given a graph G, does it contain a two-node independent set?

Note that this problem belongs to $\mathcal{NP}$. Let's see how an instance of this problem can be solved by constructing an equivalent instance of Circuit Satisfiability.

Following the proof outline above, we first consider an efficient certifier for this problem. The input s is a graph on n nodes, which will be specified by $\binom{n}{2}$ bits: For each pair of nodes, there will be a bit saying whether there is an edge joining this pair. The certificate t can be specified by n bits: For each node, there will be a bit saying whether this node belongs to the proposed independent set. The efficient certifier now needs to check two things: that at least two of the bits in t are set to 1, and that no two bits in t are both set to 1 if they form the two ends of an edge (as determined by the corresponding bit in s).

Now, for the specific input length n corresponding to the s that we are interested in, we construct an equivalent circuit K. Suppose, for example, that we are interested in deciding the answer to this problem for a graph G on the three nodes u, v, w, in which v is joined to both u and w. This means that we are concerned with an input of length $n = 3$. Figure 8.5 shows a circuit that is equivalent to an efficient certifier for our problem on arbitrary three-node graphs. (Essentially, the right-hand side of the circuit checks that at least two nodes have been selected, and the left-hand side checks that we haven't selected both ends of any edge.) We encode the edges of G as constants in the first three sources, and we leave the remaining three sources (representing the choice of nodes to put in the independent set) as variables. Now observe that this instance of Circuit Satisfiability is satisfiable, by the assignment $1, 0, 1$ to the inputs. This corresponds to choosing nodes u and w, which indeed form a two-node independent set in our three-node graph G.

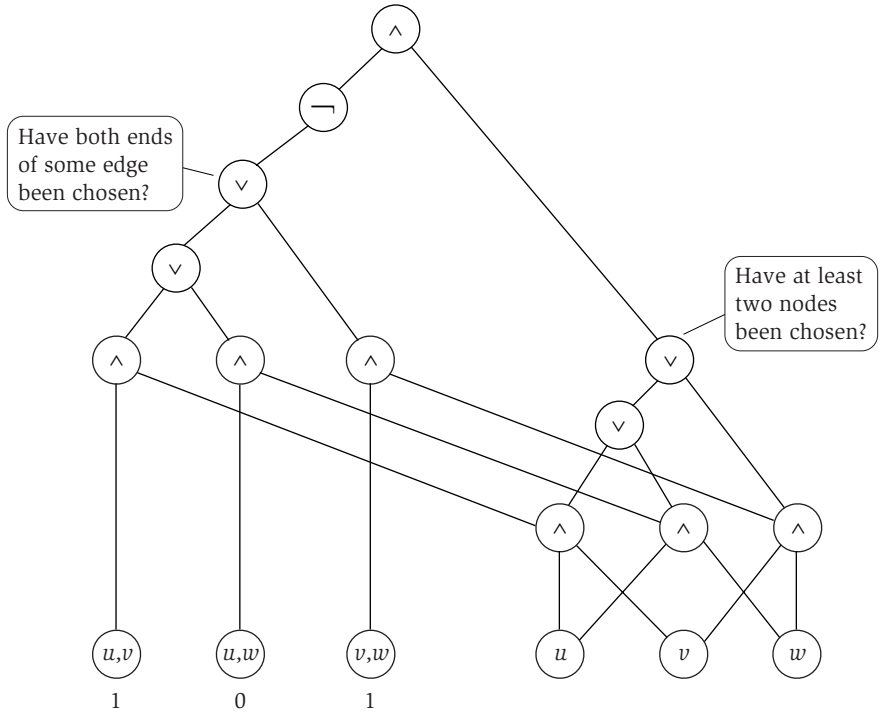

Figure 8.5 A circuit to verify whether a 3-node graph contains a 2-node independent set.

Proving Further Problems NP-Complete

Statement (8.13) opens the door to a much fuller understanding of hard problems in $\mathcal{NP}$: Once we have our hands on a first NP-complete problem, we can discover many more via the following simple observation.

(8.14) *If Y is an NP-complete problem, and X is a problem in $\mathcal{NP}$ with the property that $Y \leq_P X$, then X is NP-complete.*

Proof. Since $X \in \mathcal{NP}$, we need only verify property (ii) of the definition. So let Z be any problem in $\mathcal{NP}$. We have $Z \leq_P Y$, by the NP-completeness of Y, and $Y \leq_P X$ by assumption. By (8.9), it follows that $Z \leq_P X$. ■

So while proving (8.13) required the hard work of considering any possible problem in $\mathcal{NP}$, proving further problems NP-complete only requires a reduction from a single problem already known to be NP-complete, thanks to (8.14).

In earlier sections, we have seen a number of reductions among some basic hard problems. To establish their NP-completeness, we need to connect Circuit Satisfiability to this set of problems. The easiest way to do this is by relating it to the problem it most closely resembles, 3-Satisfiability.

(8.15) 3-Satisfiability *is NP-complete.*

Proof. Clearly 3-Satisfiability is in $\mathcal{NP}$, since we can verify in polynomial time that a proposed truth assignment satisfies the given set of clauses. We will prove that it is NP-complete via the reduction Circuit Satisfiability $\leq_P$ 3-SAT.

Given an arbitrary instance of Circuit Satisfiability, we will first construct an equivalent instance of SAT in which each clause contains *at most* three variables. Then we will convert this SAT instance to an equivalent one in which each clause has *exactly* three variables. This last collection of clauses will thus be an instance of 3-SAT, and hence will complete the reduction.

So consider an arbitrary circuit K. We associate a variable x_v with each node v of the circuit, to encode the truth value that the circuit holds at that node. Now we will define the clauses of the SAT problem. First we need to encode the requirement that the circuit computes values correctly at each gate from the input values. There will be three cases depending on the three types of gates.

- If node v is labeled with $\neg$, and its only entering edge is from node u, then we need to have $x_v = \overline{x_u}$. We guarantee this by adding two clauses $(x_v \vee x_u)$, and $(\overline{x_v} \vee \overline{x_u})$.
- If node v is labeled with $\vee$, and its two entering edges are from nodes u and w, we need to have $x_v = x_u \vee x_w$. We guarantee this by adding the following clauses: $(x_v \vee \overline{x_u})$, $(x_v \vee \overline{x_w})$, and $(\overline{x_v} \vee x_u \vee x_w)$.
- If node v is labeled with $\wedge$, and its two entering edges are from nodes u and w, we need to have $x_v = x_u \wedge x_w$. We guarantee this by adding the following clauses: $(\overline{x_v} \vee x_u)$, $(\overline{x_v} \vee x_w)$, and $(x_v \vee \overline{x_u} \vee \overline{x_w})$.

Finally, we need to guarantee that the constants at the sources have their specified values, and that the output evaluates to 1. Thus, for a source v that has been labeled with a constant value, we add a clause with the single variable x_v or $\overline{x_v}$, which forces x_v to take the designated value. For the output node o, we add the single-variable clause x_o, which requires that o take the value 1. This concludes the construction.

It is not hard to show that the SAT instance we just constructed is equivalent to the given instance of Circuit Satisfiability. To show the equivalence, we need to argue two things. First suppose that the given circuit K is satisfiable. The satisfying assignment to the circuit inputs can be propagated to create

values at all nodes in K (as we did in the example of Figure 8.4). This set of values clearly satisfies the SAT instance we constructed.

To argue the other direction, we suppose that the SAT instance we constructed is satisfiable. Consider a satisfying assignment for this instance, and look at the values of the variables corresponding to the circuit K's inputs. We claim that these values constitute a satisfying assignment for the circuit K. To see this, simply note that the SAT clauses ensure that the values assigned to all nodes of K are the same as what the circuit computes for these nodes. In particular, a value of 1 will be assigned to the output, and so the assignment to inputs satisfies K.

Thus we have shown how to create a SAT instance that is equivalent to the Circuit Satisfiability Problem. But we are not quite done, since our goal was to create an instance of 3-SAT, which requires that all clauses have length exactly 3—in the instance we constructed, some clauses have lengths of 1 or 2. So to finish the proof, we need to convert this instance of SAT to an equivalent instance in which each clause has exactly three variables.

To do this, we create four new variables: z_1, z_2, z_3, z_4. The idea is to ensure that in any satisfying assignment, we have $z_1 = z_2 = 0$, and we do this by adding the clauses $(\overline{z_i} \vee z_3 \vee z_4)$, $(\overline{z_i} \vee \overline{z_3} \vee z_4)$, $(\overline{z_i} \vee z_3 \vee \overline{z_4})$, and $(\overline{z_i} \vee \overline{z_3} \vee \overline{z_4})$ for each of $i = 1$ and $i = 2$. Note that there is no way to satisfy all these clauses unless we set $z_1 = z_2 = 0$.

Now consider a clause in the SAT instance we constructed that has a single term t (where the term t can be either a variable or the negation of a variable). We replace each such term by the clause $(t \vee z_1 \vee z_2)$. Similarly, we replace each clause that has two terms, say, $(t \vee t')$, with the clause $(t \vee t' \vee z_1)$. The resulting 3-SAT formula is clearly equivalent to the SAT formula with at most three variables in each clause, and this finishes the proof. ■

Using this NP-completeness result, and the sequence of reductions

$$\text{3-SAT} \leq_P \text{Independent Set} \leq_P \text{Vertex Cover} \leq_P \text{Set Cover},$$

summarized earlier, we can use (8.14) to conclude the following.

(8.16) *All of the following problems are NP-complete:* Independent Set, Set Packing, Vertex Cover, *and* Set Cover.

Proof. Each of these problems has the property that it is in $\mathcal{NP}$ and that 3-SAT (and hence Circuit Satisfiability) can be reduced to it. ■

General Strategy for Proving New Problems NP-Complete

For most of the remainder of this chapter, we will take off in search of further NP-complete problems. In particular, we will discuss further genres of hard computational problems and prove that certain examples of these genres are NP-complete. As we suggested initially, there is a very practical motivation in doing this: since it is widely believed that $\mathcal{P} \neq \mathcal{NP}$, the discovery that a problem is NP-complete can be taken as a strong indication that it cannot be solved in polynomial time.

Given a new problem X, here is the basic strategy for proving it is NP-complete.

1. Prove that $X \in \mathcal{NP}$.
2. Choose a problem Y that is known to be NP-complete.
3. Prove that $Y \leq_P X$.

We noticed earlier that most of our reductions $Y \leq_P X$ consist of transforming a given instance of Y into a *single* instance of X with the same answer. This is a particular way of using a black box to solve X; in particular, it requires only a single invocation of the black box. When we use this style of reduction, we can refine the strategy above to the following outline of an NP-completeness proof.

1. Prove that $X \in \mathcal{NP}$.
2. Choose a problem Y that is known to be NP-complete.
3. Consider an arbitrary instance s_Y of problem Y, and show how to construct, in polynomial time, an instance s_X of problem X that satisfies the following properties:
   (a) If s_Y is a "yes" instance of Y, then s_X is a "yes" instance of X.
   (b) If s_X is a "yes" instance of X, then s_Y is a "yes" instance of Y.
   In other words, this establishes that s_Y and s_X have the same answer.

There has been research aimed at understanding the distinction between polynomial-time reductions with this special structure—asking the black box a single question and using its answer verbatim—and the more general notion of polynomial-time reduction that can query the black box multiple times. (The more restricted type of reduction is known as a *Karp reduction*, while the more general type is known as a *Cook reduction* and also as a *polynomial-time Turing reduction*.) We will not be pursuing this distinction further here.

8.5 Sequencing Problems

Thus far we have seen problems that (like Independent Set and Vertex Cover) have involved searching over subsets of a collection of objects; we have also

seen problems that (like 3-SAT) have involved searching over 0/1 settings to a collection of variables. Another type of computationally hard problem involves searching over the set of all *permutations* of a collection of objects.

The Traveling Salesman Problem

Probably the most famous such sequencing problem is the *Traveling Salesman Problem*. Consider a salesman who must visit n cities labeled $v_1, v_2, \ldots, v_n$. The salesman starts in city v_1, his home, and wants to find a *tour*—an order in which to visit all the other cities and return home. His goal is to find a tour that causes him to travel as little total distance as possible.

To formalize this, we will take a very general notion of distance: for each ordered pair of cities (v_i, v_j), we will specify a nonnegative number $d(v_i, v_j)$ as the distance from v_i to v_j. We will not require the distance to be symmetric (so it may happen that $d(v_i, v_j) \neq d(v_j, v_i)$), nor will we require it to satisfy the triangle inequality (so it may happen that $d(v_i, v_j)$ plus $d(v_j, v_k)$ is actually less than the "direct" distance $d(v_i, v_k)$). The reason for this is to make our formulation as general as possible. Indeed, Traveling Salesman arises naturally in many applications where the points are not cities and the traveler is not a salesman. For example, people have used Traveling Salesman formulations for problems such as planning the most efficient motion of a robotic arm that drills holes in n points on the surface of a VLSI chip; or for serving I/O requests on a disk; or for sequencing the execution of n software modules to minimize the context-switching time.

Thus, given the set of distances, we ask: Order the cities into a *tour* $v_{i_1}, v_{i_2}, \ldots, v_{i_n}$, with $i_1 = 1$, so as to minimize the total distance $\sum_j d(v_{i_j}, v_{i_{j+1}}) + d(v_{i_n}, v_{i_1})$. The requirement $i_1 = 1$ simply "orients" the tour so that it starts at the home city, and the terms in the sum simply give the distance from each city on the tour to the next one. (The last term in the sum is the distance required for the salesman to return home at the end.)

Here is a decision version of the Traveling Salesman Problem.

Given a set of distances on n cities, and a bound D, is there a tour of length at most D?

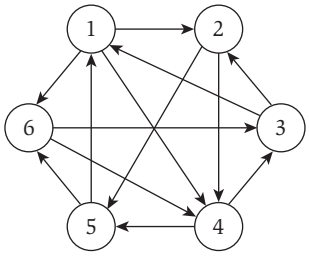

Figure 8.6 A directed graph containing a Hamiltonian cycle.

The Hamiltonian Cycle Problem

The Traveling Salesman Problem has a natural graph-based analogue, which forms one of the fundamental problems in graph theory. Given a directed graph $G = (V, E)$, we say that a cycle C in G is a *Hamiltonian cycle* if it visits each vertex exactly once. In other words, it constitutes a "tour" of all the vertices, with no repetitions. For example, the directed graph pictured in Figure 8.6 has

several Hamiltonian cycles; one visits the nodes in the order 1, 6, 4, 3, 2, 5, 1, while another visits the nodes in the order 1, 2, 4, 5, 6, 3, 1.

The Hamiltonian Cycle Problem is then simply the following:

Given a directed graph G, does it contain a Hamiltonian cycle?

Proving *Hamiltonian Cycle* is NP-Complete

We now show that both these problems are NP-complete. We do this by first establishing the NP-completeness of Hamiltonian Cycle, and then proceeding to reduce from Hamiltonian Cycle to Traveling Salesman.

(8.17) Hamiltonian Cycle *is NP-complete.*

Proof. We first show that Hamiltonian Cycle is in $\mathcal{NP}$. Given a directed graph $G = (V, E)$, a certificate that there is a solution would be the ordered list of the vertices on a Hamiltonian cycle. We could then check, in polynomial time, that this list of vertices does contain each vertex exactly once, and that each consecutive pair in the ordering is joined by an edge; this would establish that the ordering defines a Hamiltonian cycle.

We now show that 3-SAT $\leq_P$ Hamiltonian Cycle. Why are we reducing from 3-SAT? Essentially, faced with Hamiltonian Cycle, we really have no idea *what* to reduce from; it's sufficiently different from all the problems we've seen so far that there's no real basis for choosing. In such a situation, one strategy is to go back to 3-SAT, since its combinatorial structure is very basic. Of course, this strategy guarantees at least a certain level of complexity in the reduction, since we need to encode variables and clauses in the language of graphs.

So consider an arbitrary instance of 3-SAT, with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_k$. We must show how to solve it, given the ability to detect Hamiltonian cycles in directed graphs. As always, it helps to focus on the essential ingredients of 3-SAT: We can set the values of the variables however we want, and we are given three chances to satisfy each clause.

We begin by describing a graph that contains 2^n different Hamiltonian cycles that correspond very naturally to the 2^n possible truth assignments to the variables. After this, we will add nodes to model the constraints imposed by the clauses.

We construct n paths $P_1, \ldots, P_n$, where P_i consists of nodes $v_{i1}, v_{i2}, \ldots, v_{ib}$ for a quantity b that we take to be somewhat larger than the number of clauses k; say, $b = 3k + 3$. There are edges from v_{ij} to $v_{i,j+1}$ and in the other direction from $v_{i,j+1}$ to v_{ij}. Thus P_i can be traversed "left to right," from v_{i1} to v_{ib}, or "right to left," from v_{ib} to v_{i1}.

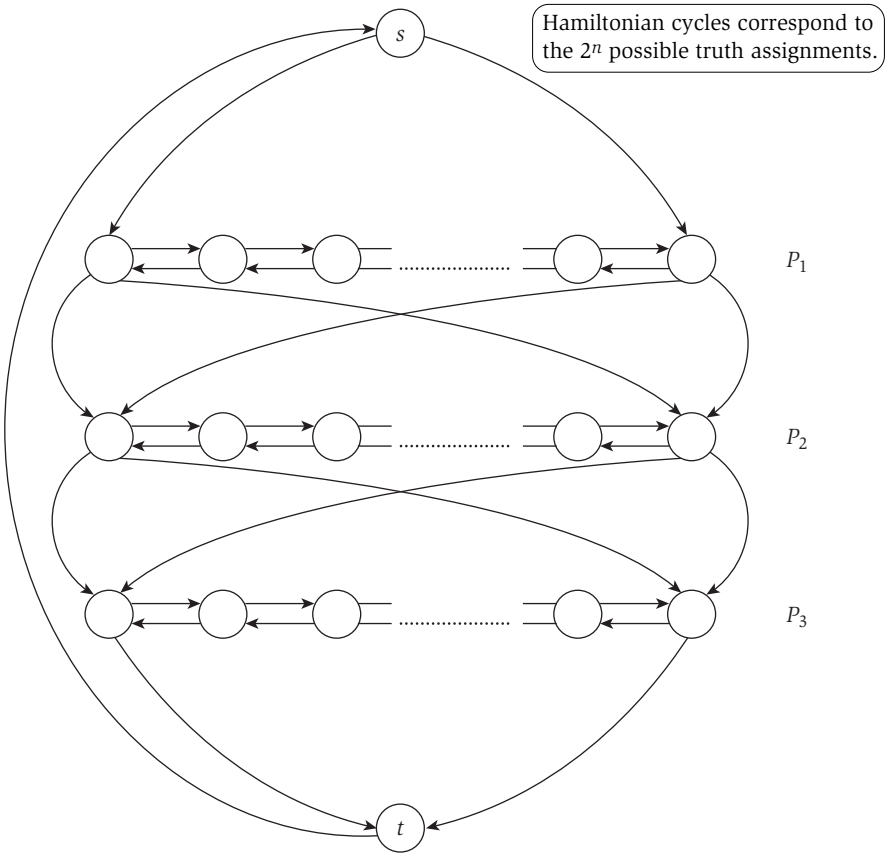

Figure 8.7 The reduction from 3-SAT to Hamiltonian Cycle: part 1.

We hook these paths together as follows. For each $i = 1, 2, \ldots, n-1$, we define edges from v_{i1} to $v_{i+1,1}$ and to $v_{i+1,b}$. We also define edges from v_{ib} to $v_{i+1,1}$ and to $v_{i+1,b}$. We add two extra nodes s and t; we define edges from s to v_{11} and v_{1b}; from v_{n1} and v_{nb} to t; and from t to s.

The construction up to this point is pictured in Figure 8.7. It's important to pause here and consider what the Hamiltonian cycles in our graph look like. Since only one edge leaves t, we know that any Hamiltonian cycle $\mathcal{C}$ must use the edge (t, s). After entering s, the cycle $\mathcal{C}$ can then traverse P_1 either left to right or right to left; regardless of what it does here, it can then traverse P_2 either left to right or right to left; and so forth, until it finishes traversing P_n and enters t. In other words, there are exactly 2^n different Hamiltonian cycles, and they correspond to the n independent choices of how to traverse each P_i.

This naturally models the n independent choices of how to set each variables $x_1, \ldots, x_n$ in the 3-SAT instance. Thus we will identify each Hamiltonian cycle uniquely with a truth assignment as follows: If $\mathcal{C}$ traverses P_i left to right, then x_i is set to 1; otherwise, x_i is set to 0.

Now we add nodes to model the clauses; the 3-SAT instance will turn out to be satisfiable if and only if any Hamiltonian cycle survives. Let's consider, as a concrete example, a clause

$$C_1 = x_1 \vee \overline{x_2} \vee x_3.$$

In the language of Hamiltonian cycles, this clause says, "The cycle should traverse P_1 left to right; or it should traverse P_2 right to left; or it should traverse P_3 left to right." So we add a node c_1, as in Figure 8.8, that does just this. (Note that certain edges have been eliminated from this drawing, for the sake of clarity.) For some value of ℓ, node c_1 will have edges *from* $v_{1\ell}$, $v_{2,\ell+1}$, and $v_{3\ell}$; it will have edges *to* $v_{1,\ell+1}$, $v_{2,\ell}$, and $v_{3,\ell+1}$. Thus it can be easily spliced into any Hamiltonian cycle that traverses P_1 left to right by visiting node c_1 between $v_{1\ell}$ and $v_{1,\ell+1}$; similarly, c_1 can be spliced into any Hamiltonian cycle that traverses P_2 right to left, or P_3 left to right. It cannot be spliced into a Hamiltonian cycle that does not do any of these things.

More generally, we will define a node c_j for each clause C_j. We will reserve node positions $3j$ and $3j+1$ in each path P_i for variables that participate in clause C_j. Suppose clause C_j contains a term t. Then if $t = x_i$, we will add edges $(v_{i,3j}, c_j)$ and $(c_j, v_{i,3j+1})$; if $t = \overline{x_i}$, we will add edges $(v_{i,3j+1}, c_j)$ and $(c_j, v_{i,3j})$.

This completes the construction of the graph G. Now, following our generic outline for NP-completeness proofs, we claim that the 3-SAT instance is satisfiable if and only if G has a Hamiltonian cycle.

First suppose there is a satisfying assignment for the 3-SAT instance. Then we define a Hamiltonian cycle following our informal plan above. If x_i is assigned 1 in the satisfying assignment, then we traverse the path P_i left to right; otherwise we traverse P_i right to left. For each clause C_j, since it is satisfied by the assignment, there will be at least one path P_i in which we will be going in the "correct" direction relative to the node c_j, and we can splice it into the tour there via edges incident on $v_{i,3j}$ and $v_{i,3j+1}$.

Conversely, suppose that there is a Hamiltonian cycle $\mathcal{C}$ in G. The crucial thing to observe is the following. If $\mathcal{C}$ enters a node c_j on an edge from $v_{i,3j}$, it must depart on an edge to $v_{i,3j+1}$. For if not, then $v_{i,3j+1}$ will have only one unvisited neighbor left, namely, $v_{i,3j+2}$, and so the tour will not be able to visit this node and still maintain the Hamiltonian property. Symmetrically, if it enters from $v_{i,3j+1}$, it must depart immediately to $v_{i,3j}$. Thus, for each node c_j,

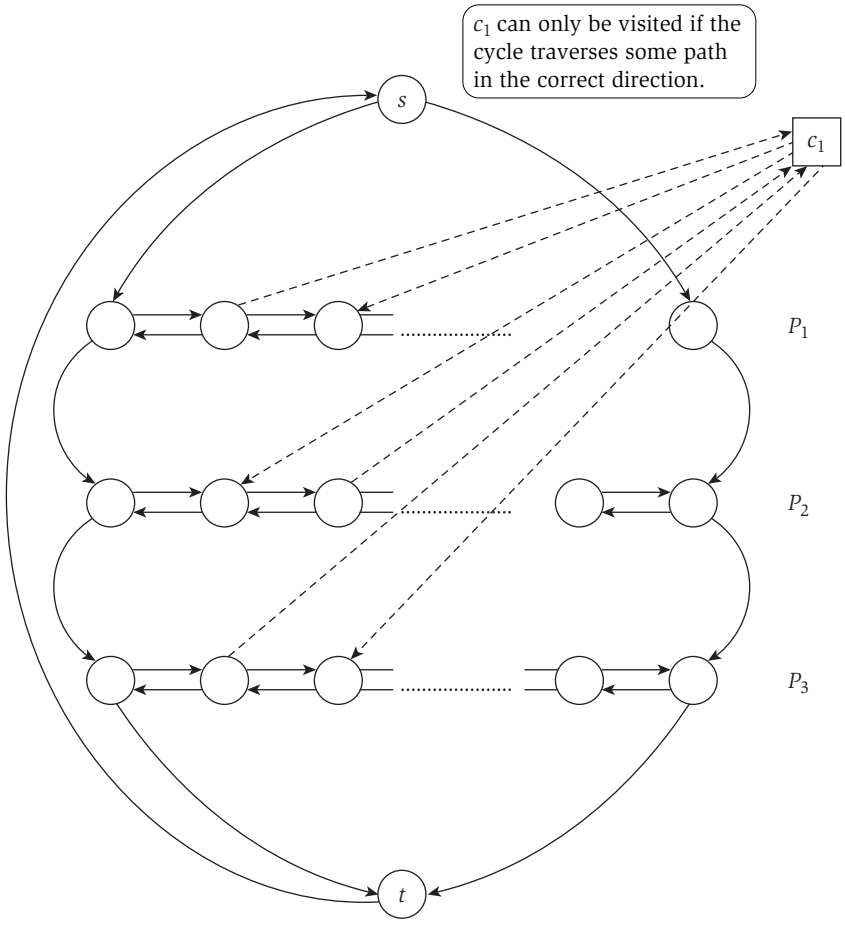

Figure 8.8 The reduction from 3-SAT to Hamiltonian Cycle: part 2.

the nodes immediately before and after c_j in the cycle $\mathcal{C}$ are joined by an edge e in G; thus, if we remove c_j from the cycle and insert this edge e for each j, then we obtain a Hamiltonian cycle $\mathcal{C}'$ on the subgraph $G - \{c_1, \ldots, c_k\}$. This is our original subgraph, before we added the clause nodes; as we noted above, any Hamiltonian cycle in this subgraph must traverse each P_i fully in one direction or the other. We thus use $\mathcal{C}'$ to define the following truth assignment for the 3-SAT instance. If $\mathcal{C}'$ traverses P_i left to right, then we set $x_i = 1$; otherwise we set $x_i = 0$. Since the larger cycle $\mathcal{C}$ was able to visit each clause node c_j, at least one of the paths was traversed in the "correct" direction relative to the node c_j, and so the assignment we have defined satisfies all the clauses.

Having established that the 3-SAT instance is satisfiable if and only if G has a Hamiltonian cycle, our proof is complete. ■

Proving Traveling Salesman is NP-Complete

Armed with our basic hardness result for Hamiltonian Cycle, we can move on to show the hardness of Traveling Salesman.

(8.18) Traveling Salesman *is NP-complete.*

Proof. It is easy to see that Traveling Salesman is in $\mathcal{NP}$: The certificate is a permutation of the cities, and a certifier checks that the length of the corresponding tour is at most the given bound.

We now show that Hamiltonian Cycle $\leq_P$ Traveling Salesman. Given a directed graph $G = (V, E)$, we define the following instance of Traveling Salesman. We have a city v'_i for each node v_i of the graph G. We define $d(v'_i, v'_j)$ to be 1 if there is an edge (v_i, v_j) in G, and we define it to be 2 otherwise.

Now we claim that G has a Hamiltonian cycle if and only if there is tour of length at most n in our Traveling Salesman instance. For if G has a Hamiltonian cycle, then this ordering of the corresponding cities defines a tour of length n. Conversely, suppose there is a tour of length at most n. The expression for the length of this tour is a sum of n terms, each of which is at least 1; thus it must be the case that all the terms are equal to 1. Hence each pair of nodes in G that correspond to consecutive cities on the tour must be connected by an edge; it follows that the ordering of these corresponding nodes must form a Hamiltonian cycle. ■

Note that allowing *asymmetric* distances in the Traveling Salesman Problem ($d(v'_i, v'_j) \neq d(v'_j, v'_i)$) played a crucial role; since the graph in the Hamiltonian Cycle instance is directed, our reduction yielded a Traveling Salesman instance with asymmetric distances.

In fact, the analogue of the Hamiltonian Cycle Problem for undirected graphs is also NP-complete; although we will not prove this here, it follows via a not-too-difficult reduction from directed Hamiltonian Cycle. Using this undirected Hamiltonian Cycle Problem, an exact analogue of (8.18) can be used to prove that the Traveling Salesman Problem with symmetric distances is also NP-complete.

Of course, the most famous special case of the Traveling Salesman Problem is the one in which the distances are defined by a set of n points in the plane. It is possible to reduce Hamiltonian Cycle to this special case as well, though this is much trickier.

Extensions: The Hamiltonian Path Problem

It is also sometimes useful to think about a variant of Hamiltonian Cycle in which it is not necessary to return to one's starting point. Thus, given a directed graph $G = (V, E)$, we say that a path P in G is a *Hamiltonian path* if it contains each vertex exactly once. (The path is allowed to start at any node and end at any node, provided it respects this constraint.) Thus such a path consists of distinct nodes $v_{i_1}, v_{i_2}, \ldots, v_{i_n}$ in order, such that they collectively constitute the entire vertex set V; by way of contrast with a Hamiltonian cycle, it is not necessary for there to be an edge from v_{i_n} back to v_{i_1}. Now, the *Hamiltonian Path Problem* asks:

Given a directed graph G, does it contain a Hamiltonian path?

Using the hardness of Hamiltonian Cycle, we show the following.

(8.19) Hamiltonian Path *is NP-complete.*

Proof. First of all, Hamiltonian Path is in $\mathcal{NP}$: A certificate could be a path in G, and a certifier could then check that it is indeed a path and that it contains each node exactly once.

One way to show that Hamiltonian Path is NP-complete is to use a reduction from 3-SAT that is almost identical to the one we used for Hamiltonian Cycle: We construct the same graph that appears in Figure 8.7, *except* that we do not include an edge from t to s. If there is any Hamiltonian path in this modified graph, it must begin at s (since s has no incoming edges) and end at t (since t has no outgoing edges). With this one change, we can adapt the argument used in the Hamiltonian Cycle reduction more or less word for word to argue that there is a satisfying assignment for the instance of 3-SAT if and only if there is a Hamiltonian path.

An alternate way to show that Hamiltonian Path is NP-complete is to prove that Hamiltonian Cycle $\leq_P$ Hamiltonian Path. Given an instance of Hamiltonian Cycle, specified by a directed graph G, we construct a graph G' as follows. We choose an arbitrary node v in G and replace it with two new nodes v' and v''. All edges out of v in G are now out of v'; and all edges into v in G are now into v''. More precisely, each edge (v, w) in G is replaced by an edge (v', w); and each edge (u, v) in G is replaced by an edge (u, v''). This completes the construction of G'.

We claim that G' contains a Hamiltonian path if and only if G contains a Hamiltonian cycle. Indeed, suppose C is a Hamiltonian cycle in G, and consider traversing it beginning and ending at node v. It is easy to see that the same ordering of nodes forms a Hamiltonian path in G' that begins at v' and ends at v''. Conversely, suppose P is a Hamiltonian path in G'. Clearly P must begin

at v' (since v' has no incoming edges) and end at v'' (since v'' has no outgoing edges). If we replace v' and v'' with v, then this ordering of nodes forms a Hamiltonian cycle in G. ■

8.6 Partitioning Problems

In the next two sections, we consider two fundamental *partitioning* problems, in which we are searching over ways of dividing a collection of objects into subsets. Here we show the NP-completeness of a problem that we call *3-Dimensional Matching*. In the next section we consider *Graph Coloring*, a problem that involves partitioning the nodes of a graph.

The 3-Dimensional Matching Problem

We begin by discussing the 3-Dimensional Matching Problem, which can be motivated as a harder version of the Bipartite Matching Problem that we considered earlier. We can view the Bipartite Matching Problem in the following way: We are given two sets X and Y, each of size n, and a set P of pairs drawn from $X \times Y$. The question is: Does there exist a set of n pairs in P so that each element in $X \cup Y$ is contained in exactly one of these pairs? The relation to Bipartite Matching is clear: the set P of pairs is simply the edges of the bipartite graph.

Now Bipartite Matching is a problem we know how to solve in polynomial time. But things get much more complicated when we move from ordered pairs to ordered triples. Consider the following 3-Dimensional Matching Problem:

> *Given disjoint sets X, Y, and Z, each of size n, and given a set $T \subseteq X \times Y \times Z$ of ordered triples, does there exist a set of n triples in T so that each element of $X \cup Y \cup Z$ is contained in exactly one of these triples?*

Such a set of triples is called a *perfect three-dimensional matching.*

An interesting thing about 3-Dimensional Matching, beyond its relation to Bipartite Matching, is that it simultaneously forms a special case of both Set Cover and Set Packing: we are seeking to *cover* the ground set $X \cup Y \cup Z$ with a collection of *disjoint* sets. More concretely, *3-Dimensional Matching* is a special case of *Set Cover* since we seek to cover the ground set $U = X \cup Y \cup Z$ using at most n sets from a given collection (the triples). Similarly, 3-Dimensional Matching is a special case of Set Packing, since we are seeking n disjoint subsets of the ground set $U = X \cup Y \cup Z$.

Proving 3-Dimensional Matching Is NP-Complete

The arguments above can be turned quite easily into proofs that 3-Dimensional Matching $\leq_P$ Set Cover and that 3-Dimensional Matching $\leq_P$ Set Packing.

But this doesn't help us establish the NP-completeness of 3-Dimensional Matching, since these reductions simply show that 3-Dimensional Matching can be reduced to some very hard problems. What we need to show is the other direction: that a known NP-complete problem can be reduced to 3-Dimensional Matching.

(8.20) 3-Dimensional Matching *is NP-complete.*

Proof. Not surprisingly, it is easy to prove that 3-Dimensional Matching is in $\mathcal{NP}$. Given a collection of triples $T \subset X \times Y \times Z$, a certificate that there is a solution could be a collection of triples $T' \subseteq T$. In polynomial time, one could verify that each element in $X \cup Y \cup Z$ belongs to exactly one of the triples in T'.

For the reduction, we again return all the way to 3-SAT. This is perhaps a little more curious than in the case of Hamiltonian Cycle, since 3-Dimensional Matching is so closely related to both Set Packing and Set Cover; but in fact the partitioning requirement is very hard to encode using either of these problems.

Thus, consider an arbitrary instance of 3-SAT, with n variables $x_1, \ldots, x_n$ and k clauses $C_1, \ldots, C_k$. We will show how to solve it, given the ability to detect perfect three-dimensional matchings.

The overall strategy in this reduction will be similar (at a very high level) to the approach we followed in the reduction from 3-SAT to Hamiltonian Cycle. We will first design gadgets that encode the independent choices involved in the truth assignment to each variable; we will then add gadgets that encode the constraints imposed by the clauses. In performing this construction, we will initially describe all the elements in the 3-Dimensional Matching instance simply as "elements," without trying to specify for each one whether it comes from X, Y, or Z. At the end, we will observe that they naturally decompose into these three sets.

Here is the basic gadget associated with variable x_i. We define elements $A_i = \{a_{i1}, a_{i2}, \ldots, a_{i,2k}\}$ that constitute the *core* of the gadget; we define elements $B_i = \{b_{i1}, \ldots, b_{i,2k}\}$ at the *tips* of the gadget. For each $j = 1, 2, \ldots, 2k$, we define a triple $t_{ij} = (a_{ij}, a_{i,j+1}, b_{ij})$, where we interpret addition modulo $2k$. Three of these gadgets are pictured in Figure 8.9. In gadget i, we will call a triple t_{ij} *even* if j is even, and *odd* if j is odd. In an analogous way, we will refer to a tip b_{ij} as being either *even* or *odd*.

These will be the only triples that contain the elements in A_i, so we can already say something about how they must be covered in any perfect matching: we must either use all the even triples in gadget i, or all the odd triples in gadget i. This will be our basic way of encoding the idea that x_i can

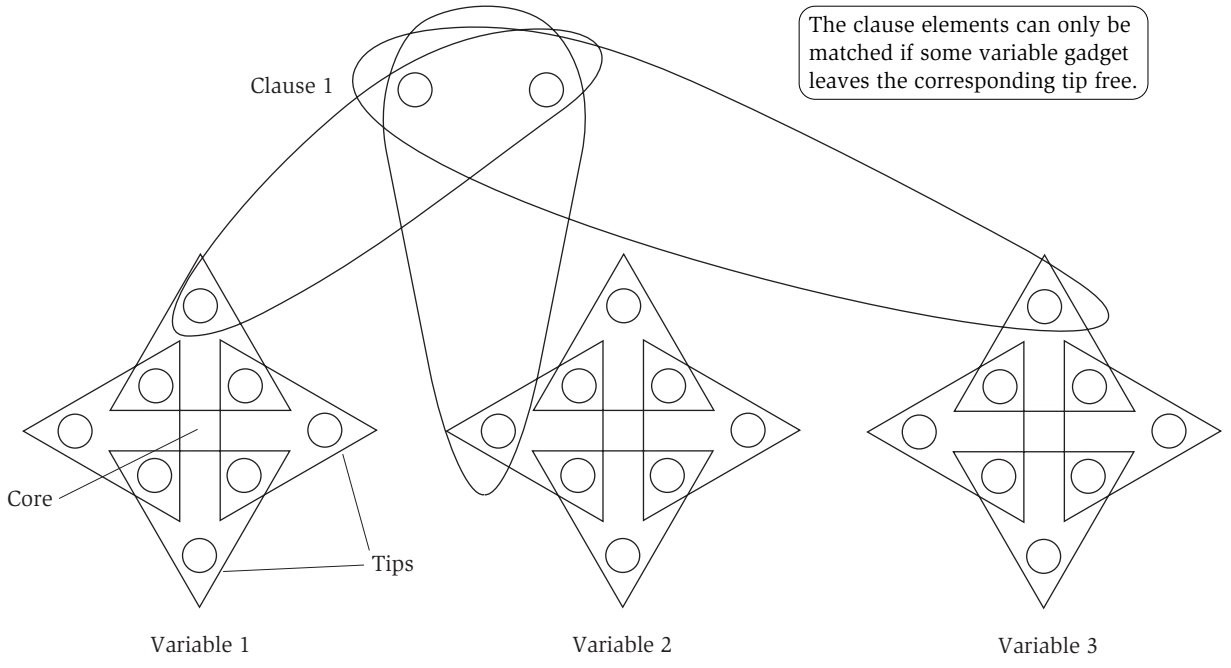

Figure 8.9 The reduction from 3-SAT to 3-Dimensional Matching.

be set to either 0 or 1; if we select all the even triples, this will represent setting $x_i = 0$, and if we select all the odd triples, this will represent setting $x_i = 1$.

Here is another way to view the odd/even decision, in terms of the tips of the gadget. If we decide to use the even triples, we cover the even tips of the gadget and leave the odd tips free. If we decide to use the odd triples, we cover the odd tips of the gadget and leave the even tips free. Thus our decision of how to set x_i can be viewed as follows: Leaving the odd tips free corresponds to 0, while leaving the even tips free corresponds to 1. This will actually be the more useful way to think about things in the remainder of the construction.

So far we can make this even/odd choice independently for each of the n variable gadgets. We now add elements to model the clauses and to constrain the assignments we can choose. As in the proof of (8.17), let's consider the example of a clause

$$C_1 = x_1 \vee \overline{x_2} \vee x_3.$$

In the language of three-dimensional matchings, it tells us, "The matching on the cores of the gadgets should leave the even tips of the first gadget free; or it should leave the odd tips of the second gadget free; or it should leave the even tips of the third gadget free." So we add a *clause gadget* that does precisely

this. It consists of a set of two *core* elements $P_1 = \{p_1, p_1'\}$, and three triples that contain them. One has the form (p_1, p_1', b_{1j}) for an even tip b_{1j}; another includes p_1, p_1', and an odd tip $b_{2,j'}$; and a third includes p_1, p_1', and an even tip $b_{3,j''}$. These are the only three triples that cover P_1, so we know that one of them must be used; this enforces the clause constraint exactly.

In general, for clause C_j, we create a gadget with two core elements $P_j = \{p_j, p_j'\}$, and we define three triples containing P_j as follows. Suppose clause C_j contains a term t. If $t = x_i$, we define a triple $(p_j, p_j', b_{i,2j})$; if $t = \overline{x_i}$, we define a triple $(p_j, p_j', b_{i,2j-1})$. Note that only clause gadget j makes use of tips b_{im} with $m = 2j$ or $m = 2j - 1$; thus, the clause gadgets will never "compete" with each other for free tips.

We are almost done with the construction, but there's still one problem. Suppose the set of clauses has a satisfying assignment. Then we make the corresponding choices of odd/even for each variable gadget; this leaves at least one free tip for each clause gadget, and so all the core elements of the clause gadgets get covered as well. The problem is that *we haven't covered all the tips.* We started with $n \cdot 2k = 2nk$ tips; the triples $\{t_{ij}\}$ covered nk of them; and the clause gadgets covered an additional k of them. This leaves $(n - 1)k$ tips left to be covered.

We handle this problem with a very simple trick: we add $(n - 1)k$ "cleanup gadgets" to the construction. Cleanup gadget i consists of two core elements $Q_i = \{q_i, q_i'\}$, and there is a triple (q_i, q_i, b) for *every* tip b in every variable gadget. This is the final piece of the construction.

Thus, if the set of clauses has a satisfying assignment, then we make the corresponding choices of odd/even for each variable gadget; as before, this leaves at least one free tip for each clause gadget. Using the cleanup gadgets to cover the remaining tips, we see that all core elements in the variable, clause, and cleanup gadgets have been covered, and all tips have been covered as well.

Conversely, suppose there is a perfect three-dimensional matching in the instance we have constructed. Then, as we argued above, in each variable gadget the matching chooses either all the even $\{t_{ij}\}$ or all the odd $\{t_{ij}\}$. In the former case, we set $x_i = 0$ in the 3-SAT instance; and in the latter case, we set $x_i = 1$. Now consider clause C_j; has it been satisfied? Because the two core elements in P_j have been covered, at least one of the three variable gadgets corresponding to a term in C_j made the "correct" odd/even decision, and this induces a variable assignment that satisfies C_j.

This concludes the proof, except for one last thing to worry about: Have we really constructed an instance of 3-Dimensional Matching? We have a collection of elements, and triples containing certain of them, but can the elements really be partitioned into appropriate sets X, Y, and Z of equal size?

Fortunately, the answer is yes. We can define X to be set of all a_{ij} with j even, the set of all p_j, and the set of all q_i. We can define Y to be set of all a_{ij} with j odd, the set of all p'_j, and the set of all q'_i. Finally, we can define Z to be the set of all tips b_{ij}. It is now easy to check that each triple consists of one element from each of X, Y, and Z. ■

8.7 Graph Coloring

When you color a map (say, the states in a U.S. map or the countries on a globe), the goal is to give neighboring regions different colors so that you can see their common borders clearly while minimizing visual distraction by using only a few colors. In the middle of the 19th century, Francis Guthrie noticed that you could color a map of the counties of England this way with only four colors, and he wondered whether the same was true for every map. He asked his brother, who relayed the question to one of his professors, and thus a famous mathematical problem was born: the *Four-Color Conjecture*.

The Graph Coloring Problem

Graph coloring refers to the same process on an undirected graph G, with the nodes playing the role of the regions to be colored, and the edges representing pairs that are neighbors. We seek to assign a *color* to each node of G so that if (u, v) is an edge, then u and v are assigned different colors; and the goal is to do this while using a small set of colors. More formally, a *k-coloring* of G is a function $f : V \to \{1, 2, \ldots, k\}$ so that for every edge (u, v), we have $f(u) \neq f(v)$. (So the available colors here are named $1, 2, \ldots, k$, and the function f represents our choice of a color for each node.) If G has a k-coloring, then we will say that it is a *k-colorable graph*.

In contrast with the case of maps in the plane, it's clear that there's not some fixed constant k so that every graph has a k-coloring: For example, if we take a set of n nodes and join each pair of them by an edge, the resulting graph needs n colors. However, the algorithmic version of the problem is very interesting:

Given a graph G and a bound k, does G have a k-coloring?

We will refer to this as the *Graph Coloring Problem*, or as *k-Coloring* when we wish to emphasize a particular choice of k.

Graph Coloring turns out to be a problem with a wide range of applications. While it's not clear there's ever been much genuine demand from cartographers, the problem arises naturally whenever one is trying to allocate resources in the presence of conflicts.

- Suppose, for example, that we have a collection of n processes on a system that can run multiple jobs concurrently, but certain pairs of jobs cannot be scheduled at the same time because they both need a particular resource. Over the next k time steps of the system, we'd like to schedule each process to run in at least one of them. Is this possible? If we construct a graph G on the set of processes, joining two by an edge if they have a conflict, then a k-coloring of G represents a conflict-free schedule of the processes: all nodes colored j can be scheduled in step j, and there will never be contention for any of the resources.
- Another well-known application arises in the design of compilers. Suppose we are compiling a program and are trying to assign each variable to one of k registers. If two variables are in use at a common point in time, then they cannot be assigned to the same register. (Otherwise one would end up overwriting the other.) Thus we can build a graph G on the set of variables, joining two by an edge if they are both in use at the same time. Now a k-coloring of G corresponds to a safe way of allocating variables to registers: All nodes colored j can be assigned to register j, since no two of them are in use at the same time.
- A third application arises in wavelength assignment for wireless communication devices: We'd like to assign one of k transmitting wavelengths to each of n devices; but if two devices are sufficiently close to each other, then they need to be assigned different wavelengths to prevent interference. To deal with this, we build a graph G on the set of devices, joining two nodes if they're close enough to interfere with each other; a k-coloring of this graph is now an assignment of wavelengths so that any nodes assigned the same wavelength are far enough apart that interference won't be a problem. (Interestingly, this is an application of graph coloring where the "colors" being assigned to nodes are positions on the electromagnetic spectrum—in other words, under a slightly liberal interpretation, they're actually colors.)

The Computational Complexity of Graph Coloring

What is the complexity of k-Coloring? First of all, the case $k = 2$ is a problem we've already seen in Chapter 3. Recall, there, that we considered the problem of determining whether a graph G is bipartite, and we showed that this is equivalent to the following question: Can one color the nodes of G red and blue so that every edge has one red end and one blue end?

But this latter question is precisely the Graph Coloring Problem in the case when there are $k = 2$ colors (i.e., red and blue) available. Thus we have argued that

(8.21) *A graph G is 2-colorable if and only if it is bipartite.*

This means we can use the algorithm from Section 3.4 to decide whether an input graph G is 2-colorable in $O(m+n)$ time, where n is the number of nodes of G and m is the number of edges.

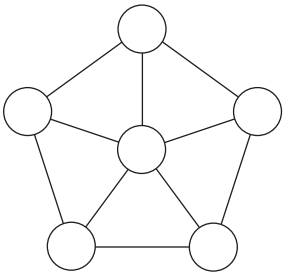

Figure 8.10 A graph that is not 3-colorable.

As soon as we move up to $k = 3$ colors, things become much harder. No simple efficient algorithm for the 3-Coloring Problem suggests itself, as it did for 2-Coloring, and it is also a very difficult problem to reason about. For example, one might initially suspect that any graph that is not 3-colorable will contain a "proof" in the form of four nodes that are all mutually adjacent (and hence would need four different colors)—but this is not true. The graph in Figure 8.10, for instance, is not 3-colorable for a somewhat more subtle (though still explainable) reason, and it is possible to draw much more complicated graphs that are not 3-colorable for reasons that seem very hard to state succinctly.

In fact, the case of three colors is already a very hard problem, as we show now.

Proving 3-Coloring Is NP-Complete

(8.22) 3-Coloring *is NP-complete.*

Proof. It is easy to see why the problem is in $\mathcal{NP}$. Given G and k, one certificate that the answer is yes is simply a k-coloring: One can verify in polynomial time that at most k colors are used, and that no pair of nodes joined by an edge receive the same color.

Like the other problems in this section, 3-Coloring is a problem that is hard to relate at a superficial level to other NP-complete problems we've seen. So once again, we're going to reach all the way back to 3-SAT. Given an arbitrary instance of 3-SAT, with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_k$, we will solve it using a black box for 3-Coloring.

The beginning of the reduction is quite intuitive. Perhaps the main power of 3-Coloring for encoding Boolean expressions lies in the fact that we can associate graph nodes with particular terms, and by joining them with edges we ensure that they get different colors; this can be used to set one true and the other false. So with this in mind, we define nodes v_i and $\overline{v_i}$ corresponding to each variable x_i and its negation $\overline{x_i}$. We also define three "special nodes" T, F, and B, which we refer to as *True*, *False*, and *Base*.

To begin, we join each pair of nodes v_i, $\overline{v_i}$ to each other by an edge, and we join both these nodes to *Base*. (This forms a triangle on v_i, $\overline{v_i}$, and *Base*, for each i.) We also join *True*, *False*, and *Base* into a triangle. The simple graph

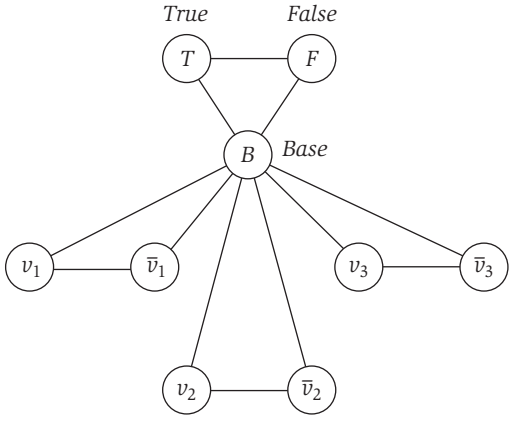

Figure 8.11 The beginning of the reduction for 3-Coloring.

G we have defined thus far is pictured in Figure 8.11, and it already has some useful properties.

- In any 3-coloring of G, the nodes v_i and $\overline{v_i}$ must get different colors, and both must be different from *Base*.
- In any 3-coloring of G, the nodes *True*, *False*, and *Base* must get all three colors in some permutation. Thus we can refer to the three colors as the *True* color, the *False* color, and the *Base* color, based on which of these three nodes gets which color. In particular, this means that for each i, one of v_i or $\overline{v_i}$ gets the *True* color, and the other gets the *False* color. For the remainder of the construction, we will consider the variable x_i to be set to 1 in the given instance of 3-SAT if and only if the node v_i gets assigned the *True* color.

So in summary, we now have a graph G in which any 3-coloring implicitly determines a truth assignment for the variables in the 3-SAT instance. We now need to grow G so that only satisfying assignments can be extended to 3-colorings of the full graph. How should we do this?

As in other 3-SAT reductions, let's consider a clause like $x_1 \vee \overline{x_2} \vee x_3$. In the language of 3-colorings of G, it says, "At least one of the nodes v_1, $\overline{v_2}$, or v_3 should get the *True* color." So what we need is a little subgraph that we can plug into G, so that any 3-coloring that extends into this subgraph must have the property of assigning the *True* color to at least one of v_1, $\overline{v_2}$, or v_3. It takes some experimentation to find such a subgraph, but one that works is depicted in Figure 8.12.

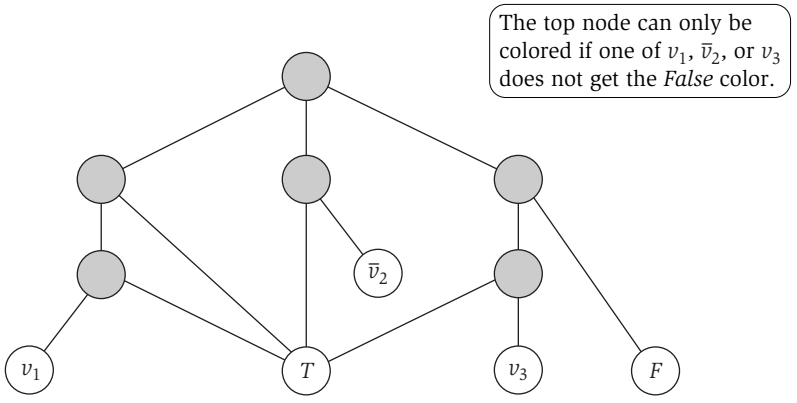

Figure 8.12 Attaching a subgraph to represent the clause $x_1 \vee \bar{x}_2 \vee x_3$.

This six-node subgraph "attaches" to the rest of G at five existing nodes: *True*, *False*, and those corresponding to the three terms in the clause that we're trying to represent (in this case, v_1, $\overline{v_2}$, and v_3.) Now suppose that in some 3-coloring of G all three of v_1, $\overline{v_2}$, and v_3 are assigned the *False* color. Then the lowest two shaded nodes in the subgraph must receive the *Base* color, the three shaded nodes above them must receive, respectively, the *False*, *Base*, and *True* colors, and hence there's no color that can be assigned to the topmost shaded node. In other words, a 3-coloring in which none of v_1, $\overline{v_2}$, or v_3 is assigned the *True* color cannot be extended to a 3-coloring of this subgraph.[2]

Finally, and conversely, some hand-checking of cases shows that as long as one of v_1, $\overline{v_2}$, or v_3 is assigned the *True* color, the full subgraph can be 3-colored.

So from this, we can complete the construction: We start with the graph G defined above, and for each clause in the 3-SAT instance, we attach a six-node subgraph as shown in Figure 8.12. Let us call the resulting graph G'.

[2] This argument actually gives considerable insight into how one comes up with this subgraph in the first place. The goal is to have a node like the topmost one that cannot receive any color. So we start by "plugging in" three nodes corresponding to the terms, all colored *False*, at the bottom. For each one, we then work upward, pairing it off with a node of a known color to force the node above to have the third color. Proceeding in this way, we can arrive at a node that is forced to have any color we want. So we force each of the three different colors, starting from each of the three different terms, and then we plug all three of these differently colored nodes into our topmost node, arriving at the impossibility.

We now claim that the given 3-SAT instance is satisfiable if and only if G' has a 3-coloring. First, suppose that there is a satisfying assignment for the 3-SAT instance. We define a coloring of G' by first coloring *Base*, *True*, and *False* arbitrarily with the three colors, then, for each i, assigning v_i the *True* color if $x_i = 1$ and the *False* color if $x_i = 0$. We then assign $\overline{v_i}$ the only available color. Finally, as argued above, it is now possible to extend this 3-coloring into each six-node clause subgraph, resulting in a 3-coloring of all of G'.

Conversely, suppose G' has a 3-coloring. In this coloring, each node v_i is assigned either the *True* color or the *False* color; we set the variable x_i correspondingly. Now we claim that in each clause of the 3-SAT instance, at least one of the terms in the clause has the truth value 1. For if not, then all three of the corresponding nodes has the *False* color in the 3-coloring of G' and, as we have seen above, there is no 3-coloring of the corresponding clause subgraph consistent with this—a contradiction. ■

When $k > 3$, it is very easy to reduce the 3-Coloring Problem to k-Coloring. Essentially, all we do is to take an instance of 3-Coloring, represented by a graph G, add $k - 3$ new nodes, and join these new nodes to each other and to every node in G. The resulting graph is k-colorable if and only if the original graph G is 3-colorable. Thus k-Coloring for any $k > 3$ is NP-complete as well.

Coda: The Resolution of the Four-Color Conjecture

To conclude this section, we should finish off the story of the Four-Color Conjecture for maps in the plane as well. After more than a hundred years, the conjecture was finally proved by Appel and Haken in 1976. The structure of the proof was a simple induction on the number of regions, but the induction step involved nearly two thousand fairly complicated cases, and the verification of these cases had to be carried out by a computer. This was not a satisfying outcome for most mathematicians: Hoping for a proof that would yield some insight into why the result was true, they instead got a case analysis of enormous complexity whose proof could not be checked by hand. The problem of finding a reasonably short, human-readable proof still remains open.

8.8 Numerical Problems

We now consider some computationally hard problems that involve arithmetic operations on numbers. We will see that the intractability here comes from the way in which some of the problems we have seen earlier in the chapter can be encoded in the representations of very large integers.

The Subset Sum Problem

Our basic problem in this genre will be *Subset Sum*, a special case of the Knapsack Problem that we saw before in Section 6.4 when we covered dynamic programming. We can formulate a decision version of this problem as follows.

> *Given natural numbers $w_1, \ldots, w_n$, and a target number W, is there a subset of $\{w_1, \ldots, w_n\}$ that adds up to precisely W?*

We have already seen an algorithm to solve this problem; why are we now including it on our list of computationally hard problems? This goes back to an issue that we raised the first time we considered Subset Sum in Section 6.4. The algorithm we developed there has running time $O(nW)$, which is reasonable when W is small, but becomes hopelessly impractical as W (and the numbers w_i) grow large. Consider, for example, an instance with 100 numbers, each of which is 100 bits long. Then the input is only $100 \times 100 = 10{,}000$ digits, but W is now roughly 2^{100}.

To phrase this more generally, since integers will typically be given in bit representation, or base-10 representation, the quantity W is really *exponential* in the size of the input; our algorithm was not a polynomial-time algorithm. (We referred to it as *pseudo-polynomial*, to indicate that it ran in time polynomial in the magnitude of the input numbers, but not polynomial in the size of their representation.)

This is an issue that comes up in many settings; for example, we encountered it in the context of network flow algorithms, where the capacities had integer values. Other settings may be familiar to you as well. For example, the security of a cryptosystem such as RSA is motivated by the sense that factoring a 1,000-bit number is difficult. But if we considered a running time of 2^{1000} steps feasible, factoring such a number would not be difficult at all.

It is worth pausing here for a moment and asking: Is this notion of polynomial time for numerical operations too severe a restriction? For example, given two natural numbers w_1 and w_2 represented in base-d notation for some $d > 1$, how long does it take to add, subtract, or multiply them? This is an issue we touched on in Section 5.5, where we noted that the standard ways that kids in elementary school learn to perform these operations have (low-degree) polynomial running times. Addition and subtraction (with carries) take $O(\log w_1 + \log w_2)$ time, while the standard multiplication algorithm runs in $O(\log w_1 \cdot \log w_2)$ time. (Recall that in Section 5.5 we discussed the design of an asymptotically faster multiplication algorithm that elementary schoolchildren are unlikely to invent on their own.)

So a basic question is: Can Subset Sum be solved by a (genuinely) polynomial-time algorithm? In other words, could there be an algorithm with running time polynomial in n and $\log W$? Or polynomial in n alone?

Proving Subset Sum Is NP-Complete

The following result suggests that this is not likely to be the case.

(8.23) Subset Sum *is NP-complete.*

Proof. We first show that Subset Sum is in $\mathcal{NP}$. Given natural numbers $w_1, \ldots, w_n$, and a target W, a certificate that there is a solution would be the subset $w_{i_1}, \ldots, w_{i_k}$ that is purported to add up to W. In polynomial time, we can compute the sum of these numbers and verify that it is equal to W.

We now reduce a known NP-complete problem to Subset Sum. Since we are seeking a set that adds up to *exactly* a given quantity (as opposed to being bounded above or below by this quantity), we look for a combinatorial problem that is based on meeting an *exact* bound. The 3-Dimensional Matching Problem is a natural choice; we show that 3-Dimensional Matching $\leq_P$ Subset Sum. The trick will be to encode the manipulation of sets via the addition of integers.

So consider an instance of 3-Dimensional Matching specified by sets X, Y, Z, each of size n, and a set of m triples $T \subseteq X \times Y \times Z$. A common way to represent sets is via *bit-vectors*: Each entry in the vector corresponds to a different element, and it holds a 1 if and only if the set contains that element. We adopt this type of approach for representing each triple $t = (x_i, y_j, z_k) \in T$: we construct a number w_t with $3n$ digits that has a 1 in positions i, $n + j$, and $2n + k$, and a 0 in all other positions. In other words, for some base $d > 1$, $w_t = d^{i-1} + d^{n+j-1} + d^{2n+k-1}$.

Note how taking the union of triples *almost* corresponds to integer addition: The 1s fill in the places where there is an element in any of the sets. But we say *almost* because addition includes *carries*: too many 1s in the same column will "roll over" and produce a nonzero entry in the next column. This has no analogue in the context of the union operation.

In the present situation, we handle this problem by a simple trick. We have only m numbers in all, and each has digits equal to 0 or 1; so if we assume that our numbers are written in base $d = m + 1$, then there will be no carries at all.

Thus we construct the following instance of Subset Sum. For each triple $t = (x_i, y_j, z_k) \in T$, we construct a number w_t in base $m + 1$ as defined above. We define W to be the number in base $m + 1$ with $3n$ digits, each of which is equal to 1, that is, $W = \sum_{i=0}^{3n-1} (m + 1)^i$.

We claim that the set T of triples contains a perfect three-dimensional matching if and only if there is a subset of the numbers $\{w_t\}$ that adds up to W. For suppose there is a perfect three-dimensional matching consisting of

triples $t_1, \ldots, t_n$. Then in the sum $w_{t_1} + \cdots + w_{t_n}$, there is a single 1 in each of the $3n$ digit positions, and so the result is equal to W.

Conversely, suppose there exists a set of numbers $w_{t_1}, \ldots, w_{t_k}$ that adds up to W. Then since each w_{t_i} has three 1s in its representation, and there are no carries, we know that $k = n$. It follows that for each of the $3n$ digit positions, exactly one of the w_{t_i} has a 1 in that position. Thus, $t_1, \ldots, t_k$ constitute a perfect three-dimensional matching. ■

Extensions: The Hardness of Certain Scheduling Problems

The hardness of Subset Sum can be used to establish the hardness of a range of scheduling problems—including some that do not obviously involve the addition of numbers. Here is a nice example, a natural (but much harder) generalization of a scheduling problem we solved in Section 4.2 using a greedy algorithm.

Suppose we are given a set of n jobs that must be run on a single machine. Each job i has a *release time* r_i when it is first available for processing; a *deadline* d_i by which it must be completed; and a *processing duration* t_i. We will assume that all of these parameters are natural numbers. In order to be completed, job i must be allocated a contiguous slot of t_i time units somewhere in the interval $[r_i, d_i]$. The machine can run only one job at a time. The question is: Can we schedule all jobs so that each completes by its deadline? We will call this an instance of *Scheduling with Release Times and Deadlines*.

(8.24) Scheduling with Release Times and Deadlines *is NP-complete.*

Proof. Given an instance of the problem, a certificate that it is solvable would be a specification of the starting time for each job. We could then check that each job runs for a distinct interval of time, between its release time and deadline. Thus the problem is in $\mathcal{NP}$.

We now show that Subset Sum is reducible to this scheduling problem. Thus, consider an instance of Subset Sum with numbers $w_1, \ldots, w_n$ and a target W. In constructing an equivalent scheduling instance, one is struck initially by the fact that we have so many parameters to manage: release times, deadlines, and durations. The key is to sacrifice most of this flexibility, producing a "skeletal" instance of the problem that still encodes the Subset Sum Problem.

Let $S = \sum_{i=1}^{n} w_i$. We define jobs $1, 2, \ldots, n$; job i has a release time of 0, a deadline of $S + 1$, and a duration of w_i. For this set of jobs, we have the freedom to arrange them in any order, and they will all finish on time.

We now further constrain the instance so that the only way to solve it will be to group together a subset of the jobs whose durations add up precisely to W. To do this, we define an $(n+1)^{\text{st}}$ job; it has a release time of W, a deadline of $W+1$, and a duration of 1.

Now consider any feasible solution to this instance of the scheduling problem. The $(n+1)^{\text{st}}$ job must be run in the interval $[W, W+1]$. This leaves S available time units between the common release time and the common deadline; and there are S time units worth of jobs to run. Thus the machine must not have any idle time, when no jobs are running. In particular, if jobs $i_1, \ldots, i_k$ are the ones that run before time W, then the corresponding numbers $w_{i_1}, \ldots, w_{i_k}$ in the Subset Sum instance add up to exactly W.

Conversely, if there are numbers $w_{i_1}, \ldots, w_{i_k}$ that add up to exactly W, then we can schedule these before job $n+1$ and the remainder after job $n+1$; this is a feasible solution to the scheduling instance. ■

Caveat: Subset Sum with Polynomially Bounded Numbers

There is a very common source of pitfalls involving the Subset Sum Problem, and while it is closely connected to the issues we have been discussing already, we feel it is worth discussing explicitly. The pitfall is the following.

> *Consider the special case of Subset Sum, with n input numbers, in which W is bounded by a polynomial function of n. Assuming $\mathcal{P} \neq \mathcal{NP}$, this special case is* not *NP-complete.*

It is not NP-complete for the simple reason that it can be solved in time $O(nW)$, by our dynamic programming algorithm from Section 6.4; when W is bounded by a polynomial function of n, this is a polynomial-time algorithm.

All this is very clear; so you may ask: Why dwell on it? The reason is that there is a genre of problem that is often wrongly claimed to be NP-complete (even in published papers) via reduction from this special case of Subset Sum. Here is a basic example of such a problem, which we will call *Component Grouping*.

> *Given a graph G that is not connected, and a number k, does there exist a subset of its connected components whose union has size exactly k?*

Incorrect Claim. Component Grouping is NP-complete.

Incorrect Proof. Component Grouping is in $\mathcal{NP}$, and we'll skip the proof of this. We now attempt to show that Subset Sum $\leq_P$ Component Grouping. Given an instance of Subset Sum with numbers $w_1, \ldots, w_n$ and target W, we construct an instance of Component Grouping as follows. For each i, we construct a path P_i of length w_i. The graph G will be the union of the paths

$P_1, \ldots, P_n$, each of which is a separate connected component. We set $k = W$. It is clear that G has a set of connected components whose union has size k if and only if some subset of the numbers $w_1, \ldots, w_n$ adds up to W. ■

The error here is subtle; in particular, the claim in the last sentence is correct. The problem is that the construction described above does not establish that Subset Sum $\leq_P$ Component Grouping, because it requires more than polynomial time. In constructing the input to our black box that solves Component Grouping, we had to build the encoding of a graph of size $w_1 + \cdots + w_n$, and this takes time exponential in the size of the input to the Subset Sum instance. In effect, Subset Sum works with the numbers $w_1, \ldots, w_n$ in a very compact representation, but Component Grouping does not accept "compact" encodings of graphs.

The problem is more fundamental than the incorrectness of this proof; in fact, Component Grouping is a problem that can be solved in polynomial time. If $n_1, n_2, \ldots, n_c$ denote the sizes of the connected components of G, we simply use our dynamic programming algorithm for Subset Sum to decide whether some subset of these numbers $\{n_i\}$ adds up to k. The running time required for this is $O(ck)$; and since c and k are both bounded by n, this is $O(n^2)$ time.

Thus we have discovered a new polynomial-time algorithm by reducing in the other direction, to a polynomial-time solvable special case of Subset Sum.

8.9 Co-NP and the Asymmetry of NP

As a further perspective on this general class of problems, let's return to the definitions underlying the class $\mathcal{NP}$. We've seen that the notion of an efficient certifier doesn't suggest a concrete algorithm for actually solving the problem that's better than brute-force search.

Now here's another observation: The definition of efficient certification, and hence of $\mathcal{NP}$, is fundamentally *asymmetric*. An input string s is a "yes" instance if and only if there exists a short t so that $B(s, t) = \texttt{yes}$. Negating this statement, we see that an input string s is a "no" instance if and only if *for all* short t, it's the case that $B(s, t) = \texttt{no}$.

This relates closely to our intuition about $\mathcal{NP}$: When we have a "yes" instance, we can provide a short proof of this fact. But when we have a "no" instance, no correspondingly short proof is guaranteed by the definition; the answer is no simply because there is no string that will serve as a proof. In concrete terms, recall our question from Section 8.3: Given an unsatisfiable set of clauses, what evidence could we show to quickly convince you that there is no satisfying assignment?

For every problem X, there is a natural *complementary* problem $\overline{X}$: For all input strings s, we say $s \in \overline{X}$ if and only if $s \notin X$. Note that if $X \in \mathcal{P}$, then $\overline{X} \in \mathcal{P}$, since from an algorithm A that solves X, we can simply produce an algorithm $\overline{A}$ that runs A and then flips its answer.

But it is far from clear that if $X \in \mathcal{NP}$, it should follow that $\overline{X} \in \mathcal{NP}$. The problem $\overline{X}$, rather, has a different property: for all s, we have $s \in \overline{X}$ if and only if for all t of length at most $p(|s|)$, $B(s, t) = \texttt{no}$. This is a fundamentally different definition, and it can't be worked around by simply "inverting" the output of the efficient certifier B to produce $\overline{B}$. The problem is that the "exists t" in the definition of $\mathcal{NP}$ has become a "for all t," and this is a serious change.

There is a class of problems parallel to $\mathcal{NP}$ that is designed to model this issue; it is called, naturally enough, co-$\mathcal{NP}$. A problem X belongs to co-$\mathcal{NP}$ if and only if the complementary problem $\overline{X}$ belongs to $\mathcal{NP}$. We do not know for sure that $\mathcal{NP}$ and co-$\mathcal{NP}$ are different; we can only ask

(8.25) *Does* $\mathcal{NP} =$ co-$\mathcal{NP}$*?*

Again, the widespread belief is that $\mathcal{NP} \neq$ co-$\mathcal{NP}$: Just because the "yes" instances of a problem have short proofs, it is not clear why we should believe that the "no" instances have short proofs as well.

Proving $\mathcal{NP} \neq$ co-$\mathcal{NP}$ would be an even bigger step than proving $\mathcal{P} \neq \mathcal{NP}$, for the following reason:

(8.26) *If* $\mathcal{NP} \neq$ co-$\mathcal{NP}$*, then* $\mathcal{P} \neq \mathcal{NP}$.

Proof. We'll actually prove the contrapositive statement: $\mathcal{P} = \mathcal{NP}$ implies $\mathcal{NP} =$ co-$\mathcal{NP}$. Essentially, the point is that $\mathcal{P}$ is closed under complementation; so if $\mathcal{P} = \mathcal{NP}$, then $\mathcal{NP}$ would be closed under complementation as well. More formally, starting from the assumption $\mathcal{P} = \mathcal{NP}$, we have

$$X \in \mathcal{NP} \Longrightarrow X \in \mathcal{P} \Longrightarrow \overline{X} \in \mathcal{P} \Longrightarrow \overline{X} \in \mathcal{NP} \Longrightarrow X \in \text{co-}\mathcal{NP}$$

and

$$X \in \text{co-}\mathcal{NP} \Longrightarrow \overline{X} \in \mathcal{NP} \Longrightarrow \overline{X} \in \mathcal{P} \Longrightarrow X \in \mathcal{P} \Longrightarrow X \in \mathcal{NP}.$$

Hence it would follow that $\mathcal{NP} \subseteq$ co-$\mathcal{NP}$ and co-$\mathcal{NP} \subseteq \mathcal{NP}$, whence $\mathcal{NP} =$ co-$\mathcal{NP}$. ■

Good Characterizations: The Class NP ∩ co-NP

If a problem X belongs to both $\mathcal{NP}$ and co-$\mathcal{NP}$, then it has the following nice property: When the answer is yes, there is a short proof; and when the answer is no, there is also a short proof. Thus problems that belong to this intersection

$\mathcal{NP} \cap$ co-$\mathcal{NP}$ are said to have a *good characterization*, since there is always a nice certificate for the solution.

This notion corresponds directly to some of the results we have seen earlier. For example, consider the problem of determining whether a flow network contains a flow of value at least ν, for some quantity ν. To prove that the answer is yes, we could simply exhibit a flow that achieves this value; this is consistent with the problem belonging to $\mathcal{NP}$. But we can also prove the answer is no: We can exhibit a cut whose capacity is strictly less than ν. This duality between "yes" and "no" instances is the crux of the Max-Flow Min-Cut Theorem.

Similarly, Hall's Theorem for matchings from Section 7.5 proved that the Bipartite Perfect Matching Problem is in $\mathcal{NP} \cap$ co-$\mathcal{NP}$: We can exhibit either a perfect matching, or a set of vertices $A \subseteq X$ such that the total number of neighbors of A is strictly less than $|A|$.

Now, if a problem X is in $\mathcal{P}$, then it belongs to both $\mathcal{NP}$ and co-$\mathcal{NP}$; thus, $\mathcal{P} \subseteq \mathcal{NP} \cap$ co-$\mathcal{NP}$. Interestingly, both our proof of the Max-Flow Min-Cut Theorem and our proof of Hall's Theorem came hand in hand with proofs of the stronger results that Maximum Flow and Bipartite Matching are problems in $\mathcal{P}$. Nevertheless, the good characterizations themselves are so clean that formulating them separately still gives us a lot of conceptual leverage in reasoning about these problems.

Naturally, one would like to know whether there's a problem that has a good characterization but no polynomial-time algorithm. But this too is an open question:

(8.27) *Does* $\mathcal{P} = \mathcal{NP} \cap$ co-$\mathcal{NP}$*?*

Unlike questions (8.11) and (8.25), general opinion seems somewhat mixed on this one. In part, this is because there are many cases in which a problem was found to have a nontrivial good characterization; and then (sometimes many years later) it was also discovered to have a polynomial-time algorithm.

8.10 A Partial Taxonomy of Hard Problems

We've now reached the end of the chapter, and we've encountered a fairly rich array of NP-complete problems. In a way, it's useful to know a good number of different NP-complete problems: When you encounter a new problem X and want to try proving it's NP-complete, you want to show $Y \leq_P X$ for some known NP-complete problem Y—so the more options you have for Y, the better.

At the same time, the more options you have for Y, the more bewildering it can be to try choosing the right one to use in a particular reduction. Of course, the whole point of NP-completeness is that one of these problems will work in your reduction if and only if any of them will (since they're all equivalent with respect to polynomial-time reductions); but the reduction to a given problem X can be much, much easier starting from some problems than from others.

With this in mind, we spend this concluding section on a review of the NP-complete problems we've come across in the chapter, grouped into six basic genres. Together with this grouping, we offer some suggestions as to how to choose a starting problem for use in a reduction.

Packing Problems

Packing problems tend to have the following structure: You're given a collection of objects, and you want to choose at least k of them; making your life difficult is a set of conflicts among the objects, preventing you from choosing certain groups simultaneously.

We've seen two basic packing problems in this chapter.

- *Independent Set*: Given a graph G and a number k, does G contain an independent set of size at least k?
- *Set Packing*: Given a set U of n elements, a collection $S_1, \ldots, S_m$ of subsets of U, and a number k, does there exist a collection of at least k of these sets with the property that no two of them intersect?

Covering Problems

Covering problems form a natural contrast to packing problems, and one typically recognizes them as having the following structure: you're given a collection of objects, and you want to choose a subset that collectively achieves a certain goal; the challenge is to achieve this goal while choosing only k of the objects.

We've seen two basic covering problems in this chapter.

- *Vertex Cover*: Given a graph G and a number k, does G contain a vertex cover of size at most k?
- *Set Cover*: Given a set U of n elements, a collection $S_1, \ldots, S_m$ of subsets of U, and a number k, does there exist a collection of at most k of these sets whose union is equal to all of U?

Partitioning Problems

Partitioning problems involve a search over all ways to divide up a collection of objects into subsets so that each object appears in exactly one of the subsets.

One of our two basic partitioning problems, 3-Dimensional Matching, arises naturally whenever you have a collection of sets and you want to solve a covering problem and a packing problem simultaneously: Choose some of the sets in such a way that they are disjoint, yet completely cover the ground set.

- *3-Dimensional Matching*: Given disjoint sets X, Y, and Z, each of size n, and given a set $T \subseteq X \times Y \times Z$ of ordered triples, does there exist a set of n triples in T so that each element of $X \cup Y \cup Z$ is contained in exactly one of these triples?

Our other basic partitioning problem, Graph Coloring, is at work whenever you're seeking to partition objects in the presence of conflicts, and conflicting objects aren't allowed to go into the same set.

- *Graph Coloring*: Given a graph G and a bound k, does G have a k-coloring?

Sequencing Problems

Our first three types of problems have involved searching over subsets of a collection of objects. Another type of computationally hard problem involves searching over the set of all *permutations* of a collection of objects.

Two of our basic sequencing problems draw their difficulty from the fact that you are required to order n objects, but there are restrictions preventing you from placing certain objects after certain others.

- *Hamiltonian Cycle*: Given a directed graph G, does it contain a Hamiltonian cycle?
- *Hamiltonian Path*: Given a directed graph G, does it contain a Hamiltonian path?

Our third basic sequencing problem is very similar; it softens these restrictions by simply imposing a cost for placing one object after another.

- *Traveling Salesman*: Given a set of distances on n cities, and a bound D, is there a tour of length at most D?

Numerical Problems

The hardness of the numerical problems considered in this chapter flowed principally from Subset Sum, the special case of the Knapsack Problem that we considered in Section 8.8.

- *Subset Sum*: Given natural numbers $w_1, \ldots, w_n$, and a target number W, is there a subset of $\{w_1, \ldots, w_n\}$ that adds up to precisely W?

It is natural to try reducing from *Subset Sum* whenever one has a problem with weighted objects and the goal is to select objects conditioned on a constraint on

the total weight of the objects selected. This, for example, is what happened in the proof of (8.24), showing that Scheduling with Release Times and Deadlines is NP-complete.

At the same time, one must heed the warning that Subset Sum only becomes hard with truly large integers; when the magnitudes of the input numbers are bounded by a polynomial function of n, the problem is solvable in polynomial time by dynamic programming.

Constraint Satisfaction Problems

Finally, we considered basic constraint satisfaction problems, including Circuit Satisfiability, SAT, and 3-SAT. Among these, the most useful for the purpose of designing reductions is 3-SAT.

- *3-SAT*: Given a set of clauses $C_1, \ldots, C_k$, each of length 3, over a set of variables $X = \{x_1, \ldots, x_n\}$, does there exist a satisfying truth assignment?

Because of its expressive flexibility, 3-SAT is often a useful starting point for reductions where none of the previous five categories seem to fit naturally onto the problem being considered. In designing 3-SAT reductions, it helps to recall the advice given in the proof of (8.8), that there are two distinct ways to view an instance of 3-SAT: (a) as a search over assignments to the variables, subject to the constraint that all clauses must be satisfied, and (b) as a search over ways to choose a single term (to be satisfied) from each clause, subject to the constraint that one mustn't choose conflicting terms from different clauses. Each of these perspectives on 3-SAT is useful, and each forms the key idea behind a large number of reductions.

Solved Exercises

Solved Exercise 1

You're consulting for a small high-tech company that maintains a high-security computer system for some sensitive work that it's doing. To make sure this system is not being used for any illicit purposes, they've set up some logging software that records the IP addresses that all their users are accessing over time. We'll assume that each user accesses at most one IP address in any given minute; the software writes a log file that records, for each user u and each minute m, a value $I(u, m)$ that is equal to the IP address (if any) accessed by user u during minute m. It sets $I(u, m)$ to the null symbol $\perp$ if u did not access any IP address during minute m.

The company management just learned that yesterday the system was used to launch a complex attack on some remote sites. The attack was carried out by accessing t distinct IP addresses over t consecutive minutes: In minute

1, the attack accessed address i_1; in minute 2, it accessed address i_2; and so on, up to address i_t in minute t.

Who could have been responsible for carrying out this attack? The company checks the logs and finds to its surprise that there's no single user u who accessed each of the IP addresses involved at the appropriate time; in other words, there's no u so that $I(u, m) = i_m$ for each minute m from 1 to t.

So the question becomes: What if there were a small *coalition* of k users that collectively might have carried out the attack? We will say a subset S of users is a *suspicious coalition* if, for each minute m from 1 to t, there is at least one user $u \in S$ for which $I(u, m) = i_m$. (In other words, each IP address was accessed at the appropriate time by at least one user in the coalition.)

The *Suspicious Coalition Problem* asks: Given the collection of all values $I(u, m)$, and a number k, is there a suspicious coalition of size at most k?

Solution First of all, Suspicious Coalition is clearly in $\mathcal{NP}$: If we were to be shown a set S of users, we could check that S has size at most k, and that for each minute m from 1 to t, at least one of the users in S accessed the IP address i_m.

Now we want to find a known NP-complete problem and reduce it to Suspicious Coalition. Although Suspicious Coalition has lots of features (users, minutes, IP addresses), it's very clearly a covering problem (following the taxonomy described in the chapter): We need to explain all t suspicious accesses, and we're allowed a limited number of users (k) with which to do this. Once we've decided it's a covering problem, it's natural to try reducing Vertex Cover or Set Cover to it. And in order to do this, it's useful to push most of its complicated features into the background, leaving just the bare-bones features that will be used to encode Vertex Cover or Set Cover.

Let's focus on reducing Vertex Cover to it. In Vertex Cover, we need to cover every edge, and we're only allowed k nodes. In Suspicious Coalition, we need to "cover" all the accesses, and we're only allowed k users. This parallelism strongly suggests that, given an instance of Vertex Cover consisting of a graph $G = (V, E)$ and a bound k, we should construct an instance of Suspicious Coalition in which the users represent the nodes of G and the suspicious accesses represent the edges.

So suppose the graph G for the Vertex Cover instance has m edges $e_1, \ldots, e_m$, and $e_j = (v_j, w_j)$. We construct an instance of Suspicious Coalition as follows. For each node of G we construct a user, and for each edge $e_t = (v_t, w_t)$ we construct a minute t. (So there will be m minutes total.) In minute t, the users associated with the two ends of e_t access an IP address i_t, and all other users access nothing. Finally, the attack consists of accesses to addresses $i_1, i_2, \ldots, i_m$ in minutes $1, 2, \ldots, m$, respectively.

The following claim will establish that Vertex Cover $\leq_P$ Suspicious Coalition and hence will conclude the proof that Suspicious Coalition is NP-complete. Given how closely our construction of the instance shadows the original Vertex Cover instance, the proof is completely straightforward.

(8.28) *In the instance constructed, there is a suspicious coalition of size at most k if and only if the graph G contains a vertex cover of size at most k.*

Proof. First, suppose that G contains a vertex cover C of size at most k. Then consider the corresponding set S of users in the instance of Suspicious Coalition. For each t from 1 to m, at least one element of C is an end of the edge e_t, and the corresponding user in S accessed the IP address i_t. Hence the set S is a suspicious coalition.

Conversely, suppose that there is a suspicious coalition S of size at most k, and consider the corresponding set of nodes C in G. For each t from 1 to m, at least one user in S accessed the IP address i_t, and the corresponding node in C is an end of the edge e_t. Hence the set C is a vertex cover. ■

Solved Exercise 2

You've been asked to organize a freshman-level seminar that will meet once a week during the next semester. The plan is to have the first portion of the semester consist of a sequence of ℓ guest lectures by outside speakers, and have the second portion of the semester devoted to a sequence of p hands-on projects that the students will do.

There are n options for speakers overall, and in week number i (for $i = 1, 2, \ldots, \ell$) a subset L_i of these speakers is available to give a lecture.

On the other hand, each project requires that the students have seen certain background material in order for them to be able to complete the project successfully. In particular, for each project j (for $j = 1, 2, \ldots, p$), there is a subset P_j of relevant speakers so that the students need to have seen a lecture by *at least one of* the speakers in the set P_j in order to be able to complete the project.

So this is the problem: Given these sets, can you select exactly one speaker for each of the first ℓ weeks of the seminar, so that you only choose speakers who are available in their designated week, and so that for each project j, the students will have seen at least one of the speakers in the relevant set P_j? We'll call this the *Lecture Planning Problem*.

To make this clear, let's consider the following sample instance. Suppose that $\ell = 2$, $p = 3$, and there are $n = 4$ speakers that we denote A, B, C, D. The availability of the speakers is given by the sets $L_1 = \{A, B, C\}$ and $L_2 = \{A, D\}$. The relevant speakers for each project are given by the sets $P_1 = \{B, C\}$,

$P_2 = \{A, B, D\}$, and $P_3 = \{C, D\}$. Then the answer to this instance of the problem is yes, since we can choose speaker B in the first week and speaker D in the second week; this way, for each of the three projects, students will have seen at least one of the relevant speakers.

Prove that Lecture Planning is NP-complete.

Solution The problem is in $\mathcal{NP}$ since, given a sequence of speakers, we can check (a) all speakers are available in the weeks when they're scheduled, and (b) that for each project, at least one of the relevant speakers has been scheduled.

Now we need to find a known NP-complete problem that we can reduce to Lecture Planning. This is less clear-cut than in the previous exercise, because the statement of the Lecture Planning Problem doesn't immediately map into the taxonomy from the chapter.

There is a useful intuitive view of Lecture Planning, however, that is characteristic of a wide range of constraint satisfaction problems. This intuition is captured, in a strikingly picturesque way, by a description that appeared in the *New Yorker* of the lawyer David Boies's cross-examination style:

> *During a cross-examination, David takes a friendly walk down the hall with you while he's quietly closing doors. They get to the end of the hall and David turns on you and there's no place to go. He's closed all the doors.*[3]

What does constraint satisfaction have to do with cross-examination? In Lecture Planning, as in many similar problems, there are two conceptual phases. There's a first phase in which you walk through a set of choices, selecting some and thereby closing the door on others; this is followed by a second phase in which you find out whether your choices have left you with a valid solution or not.

In the case of Lecture Planning, the first phase consists of choosing a speaker for each week, and the second phase consists of verifying that you've picked a relevant speaker for each project. But there are many NP-complete problems that fit this description at a high level, and so viewing Lecture Planning this way helps us search for a plausible reduction. We will in fact describe two reductions, first from 3-SAT and then from Vertex Cover. Of course, either one of these by itself is enough to prove NP-completeness, but both make for useful examples.

3-SAT is the canonical example of a problem with the two-phase structure described above: We first walk through the variables, setting each one to true or false; we then look over each clause and see whether our choices

[3] Ken Auletta quoting Jeffrey Blattner, *The New Yorker*, 16 August 1999.

have satisfied it. This parallel to Lecture Planning already suggests a natural reduction showing that 3-SAT $\leq_P$ Lecture Planning: We set things up so that the choice of lecturers sets the variables, and then the feasibility of the projects represents the satisfaction of the clauses.

More concretely, suppose we are given an instance of 3-SAT consisting of clauses $C_1, \ldots, C_k$ over the variables $x_1, x_2, \ldots, x_n$. We construct an instance of Lecture Planning as follows. For each variable x_i, we create two lecturers z_i and z_i' that will correspond to x_i and its negation. We begin with n weeks of lectures; in week i, the only two lecturers available are z_i and z_i'. Then there is a sequence of k projects; for project j, the set of relevant lecturers P_j consists of the three lecturers corresponding to the terms in clause C_j. Now, if there is a satisfying assignment v for the 3-SAT instance, then in week i we choose the lecturer among z_i, z_i' that corresponds to the value assigned to x_i by v; in this case, we will select at least one speaker from each relevant set P_j. Conversely, if we find a way to choose speakers so that there is at least one from each relevant set, then we can set the variables x_i as follows: x_i is set to 1 if z_i is chosen, and it is set to 0 if z_i' is chosen. In this way, at least one of the three variables in each clause C_j is set in a way that satisfies it, and so this is a satisfying assignment. This concludes the reduction and its proof of correctness.

Our intuitive view of Lecture Planning leads naturally to a reduction from Vertex Cover as well. (What we describe here could be easily modified to work from Set Cover or 3-Dimensional Matching too.) The point is that we can view Vertex Cover as having a similar two-phase structure: We first choose a set of k nodes from the input graph, and we then verify for each edge that these choices have covered all the edges.

Given an input to Vertex Cover, consisting of a graph $G = (V, E)$ and a number k, we create a lecturer z_v for each node v. We set $\ell = k$, and define $L_1 = L_2 = \cdots = L_k = \{z_v : v \in V\}$. In other words, for the first k weeks, all lecturers are available. After this, we create a project j for each edge $e_j = (v, w)$, with set $P_j = \{z_v, z_w\}$.

Now, if there is a vertex cover S of at most k nodes, then consider the set of lecturers $Z_S = \{z_v : v \in S\}$. For each project P_j, at least one of the relevant speakers belongs to Z_S, since S covers all edges in G. Moreover, we can schedule all the lecturers in Z_S during the first k weeks. Thus it follows that there is a feasible solution to the instance of Lecture Planning.

Conversely, suppose there is a feasible solution to the instance of Lecture Planning, and let T be the set of all lecturers who speak in the first k weeks. Let X be the set of nodes in G that correspond to lecturers in T. For each project P_j, at least one of the two relevant speakers appears in T, and hence at least

one end of each edge e_j is in the set X. Thus X is a vertex cover with at most k nodes.

This concludes the proof that Vertex Cover $\leq_P$ Lecture Planning.

Exercises

1. For each of the two questions below, decide whether the answer is (i) "Yes," (ii) "No," or (iii) "Unknown, because it would resolve the question of whether $\mathcal{P} = \mathcal{NP}$." Give a brief explanation of your answer.

(a) Let's define the decision version of the Interval Scheduling Problem from Chapter 4 as follows: Given a collection of intervals on a time-line, and a bound k, does the collection contain a subset of nonoverlapping intervals of size at least k?

Question: Is it the case that Interval Scheduling $\leq_P$ Vertex Cover?

(b) Question: Is it the case that Independent Set $\leq_P$ Interval Scheduling?

2. A store trying to analyze the behavior of its customers will often maintain a two-dimensional array A, where the rows correspond to its customers and the columns correspond to the products it sells. The entry $A[i, j]$ specifies the quantity of product j that has been purchased by customer i.

Here's a tiny example of such an array A.

	liquid detergent	**beer**	**diapers**	**cat litter**
Raj	0	6	0	3
Alanis	2	3	0	0
Chelsea	0	0	0	7

One thing that a store might want to do with this data is the following. Let us say that a subset S of the customers is *diverse* if no two of the of the customers in S have ever bought the same product (i.e., for each product, at most one of the customers in S has ever bought it). A diverse set of customers can be useful, for example, as a target pool for market research.

We can now define the Diverse Subset Problem as follows: Given an $m \times n$ array A as defined above, and a number $k \leq m$, is there a subset of at least k of customers that is *diverse*?

Show that Diverse Subset is NP-complete.

3. Suppose you're helping to organize a summer sports camp, and the following problem comes up. The camp is supposed to have at least

one counselor who's skilled at each of the n sports covered by the camp (baseball, volleyball, and so on). They have received job applications from m potential counselors. For each of the n sports, there is some subset of the m applicants qualified in that sport. The question is: For a given number $k < m$, is it possible to hire at most k of the counselors and have at least one counselor qualified in each of the n sports? We'll call this the *Efficient Recruiting Problem*.

Show that Efficient Recruiting is NP-complete.

4. Suppose you're consulting for a group that manages a high-performance real-time system in which asynchronous processes make use of shared resources. Thus the system has a set of n *processes* and a set of m *resources*. At any given point in time, each process specifies a set of resources that it requests to use. Each resource might be requested by many processes at once; but it can only be used by a single process at a time. Your job is to allocate resources to processes that request them. If a process is allocated all the resources it requests, then it is *active*; otherwise it is *blocked*. You want to perform the allocation so that as many processes as possible are active. Thus we phrase the *Resource Reservation Problem* as follows: Given a set of processes and resources, the set of requested resources for each process, and a number k, is it possible to allocate resources to processes so that at least k processes will be active?

Consider the following list of problems, and for each problem either give a polynomial-time algorithm or prove that the problem is NP-complete.

(a) The general Resource Reservation Problem defined above.

(b) The special case of the problem when $k = 2$.

(c) The special case of the problem when there are two types of resources—say, people and equipment—and each process requires at most one resource of each type (In other words, each process requires one specific person and one specific piece of equipment.)

(d) The special case of the problem when each resource is requested by at most two processes.

5. Consider a set $A = \{a_1, \ldots, a_n\}$ and a collection $B_1, B_2, \ldots, B_m$ of subsets of A (i.e., $B_i \subseteq A$ for each i).

We say that a set $H \subseteq A$ is a *hitting set* for the collection $B_1, B_2, \ldots, B_m$ if H contains at least one element from each B_i—that is, if $H \cap B_i$ is not empty for each i (so H "hits" all the sets B_i).

We now define the *Hitting Set Problem* as follows. We are given a set $A = \{a_1, \ldots, a_n\}$, a collection $B_1, B_2, \ldots, B_m$ of subsets of A, and a number

k. We are asked: Is there a hitting set $H \subseteq A$ for $B_1, B_2, \ldots, B_m$ so that the size of H is at most k?

Prove that Hitting Set is NP-complete.

6. Consider an instance of the Satisfiability Problem, specified by clauses $C_1, \ldots, C_k$ over a set of Boolean variables $x_1, \ldots, x_n$. We say that the instance is *monotone* if each term in each clause consists of a nonnegated variable; that is, each term is equal to x_i, for some i, rather than $\overline{x_i}$. Monotone instances of Satisfiability are very easy to solve: They are always satisfiable, by setting each variable equal to 1.

For example, suppose we have the three clauses

$$(x_1 \vee x_2), (x_1 \vee x_3), (x_2 \vee x_3).$$

This is monotone, and indeed the assignment that sets all three variables to 1 satisfies all the clauses. But we can observe that this is not the only satisfying assignment; we could also have set x_1 and x_2 to 1, and x_3 to 0. Indeed, for any monotone instance, it is natural to ask how few variables we need to set to 1 in order to satisfy it.

Given a monotone instance of Satisfiability, together with a number k, the problem of *Monotone Satisfiability with Few True Variables* asks: Is there a satisfying assignment for the instance in which at most k variables are set to 1? Prove this problem is NP-complete.

7. Since the 3-Dimensional Matching Problem is NP-complete, it is natural to expect that the corresponding 4-Dimensional Matching Problem is at least as hard. Let us define *4-Dimensional Matching* as follows. Given sets W, X, Y, and Z, each of size n, and a collection C of ordered 4-tuples of the form (w_i, x_j, y_k, z_ℓ), do there exist n 4-tuples from C so that no two have an element in common?

Prove that 4-Dimensional Matching is NP-complete.

8. Your friends' preschool-age daughter Madison has recently learned to spell some simple words. To help encourage this, her parents got her a colorful set of refrigerator magnets featuring the letters of the alphabet (some number of copies of the letter A, some number of copies of the letter B, and so on), and the last time you saw her the two of you spent a while arranging the magnets to spell out words that she knows.

Somehow with you and Madison, things always end up getting more elaborate than originally planned, and soon the two of you were trying to spell out words so as to use up all the magnets in the full set—that is, picking words that she knows how to spell, so that once they were all spelled out, each magnet was participating in the spelling of exactly one

of the words. (Multiple copies of words are okay here; so for example, if the set of refrigerator magnets includes two copies each of C, A, and T, it would be okay to spell out CAT twice.)

This turned out to be pretty difficult, and it was only later that you realized a plausible reason for this. Suppose we consider a general version of the problem of *Using Up All the Refrigerator Magnets*, where we replace the English alphabet by an arbitrary collection of symbols, and we model Madison's vocabulary as an arbitrary set of strings over this collection of symbols. The goal is the same as in the previous paragraph.

Prove that the problem of Using Up All the Refrigerator Magnets is NP-complete.

9. Consider the following problem. You are managing a communication network, modeled by a directed graph $G = (V, E)$. There are c *users* who are interested in making use of this network. User i (for each $i = 1, 2, \ldots, c$) issues a *request* to reserve a specific path P_i in G on which to transmit data.

You are interested in accepting as many of these path requests as possible, subject to the following restriction: if you accept both P_i and P_j, then P_i and P_j cannot share any nodes.

Thus, the *Path Selection Problem* asks: Given a directed graph $G = (V, E)$, a set of requests $P_1, P_2, \ldots, P_c$—each of which must be a path in G—and a number k, is it possible to select at least k of the paths so that no two of the selected paths share any nodes?

Prove that Path Selection is NP-complete.

10. Your friends at WebExodus have recently been doing some consulting work for companies that maintain large, publicly accessible Web sites—contractual issues prevent them from saying which ones—and they've come across the following *Strategic Advertising Problem*.

A company comes to them with the map of a Web site, which we'll model as a directed graph $G = (V, E)$. The company also provides a set of t *trails* typically followed by users of the site; we'll model these trails as directed paths $P_1, P_2, \ldots, P_t$ in the graph G (i.e., each P_i is a path in G).

The company wants WebExodus to answer the following question for them: Given G, the paths $\{P_i\}$, and a number k, is it possible to place advertisements on at most k of the nodes in G, so that each path P_i includes at least one node containing an advertisement? We'll call this the Strategic Advertising Problem, with input G, $\{P_i : i = 1, \ldots, t\}$, and k.

Your friends figure that a good algorithm for this will make them all rich; unfortunately, things are never quite this simple.

(a) Prove that Strategic Advertising is NP-complete.

(b) Your friends at WebExodus forge ahead and write a pretty fast algorithm $\mathcal{S}$ that produces yes/no answers to arbitrary instances of the Strategic Advertising Problem. You may assume that the algorithm $\mathcal{S}$ is always correct.

Using the algorithm $\mathcal{S}$ as a black box, design an algorithm that takes input G, $\{P_i\}$, and k as in part (a), and does one of the following two things:

- Outputs a set of at most k nodes in G so that each path P_i includes at least one of these nodes, *or*
- Outputs (correctly) that no such set of at most k nodes exists.

Your algorithm should use at most a polynomial number of steps, together with at most a polynomial number of calls to the algorithm $\mathcal{S}$.

11. As some people remember, and many have been told, the idea of hypertext predates the World Wide Web by decades. Even hypertext fiction is a relatively old idea: Rather than being constrained by the linearity of the printed page, you can plot a story that consists of a collection of interlocked virtual "places" joined by virtual "passages."[4] So a piece of hypertext fiction is really riding on an underlying directed graph; to be concrete (though narrowing the full range of what the domain can do), we'll model this as follows.

Let's view the structure of a piece of hypertext fiction as a directed graph $G = (V, E)$. Each node $u \in V$ contains some text; when the reader is currently at u, he or she can choose to follow any edge out of u; and if the reader chooses $e = (u, v)$, he or she arrives next at the node v. There is a start node $s \in V$ where the reader begins, and an end node $t \in V$; when the reader first reaches t, the story ends. Thus any path from s to t is a valid *plot* of the story. Note that, unlike one's experience using a Web browser, there is not necessarily a way to go back; once you've gone from u to v, you might not be able to ever return to u.

In this way, the hypertext structure defines a huge number of different plots on the same underlying content; and the relationships among all these possibilities can grow very intricate. Here's a type of problem one encounters when reasoning about a structure like this. Consider a piece of hypertext fiction built on a graph $G = (V, E)$ in which there are certain crucial *thematic elements*: love, death, war, an intense desire to major in computer science, and so forth. Each thematic element i is represented by a set $T_i \subseteq V$ consisting of the nodes in G at which this theme

[4] See, e.g., *http://www.eastgate.com*.

appears. Now, given a particular set of thematic elements, we may ask: Is there a valid plot of the story in which each of these elements is encountered? More concretely, given a directed graph G, with start node s and end node t, and thematic elements represented by sets $T_1, T_2, \ldots, T_k$, the *Plot Fulfillment Problem* asks: Is there a path from s to t that contains at least one node from each of the sets T_i?

Prove that Plot Fulfillment is NP-complete.

12. Some friends of yours maintain a popular news and discussion site on the Web, and the traffic has reached a level where they want to begin differentiating their visitors into paying and nonpaying customers. A standard way to do this is to make all the content on the site available to customers who pay a monthly subscription fee; meanwhile, visitors who don't subscribe can still view a subset of the pages (all the while being bombarded with ads asking them to become subscribers).

Here are two simple ways to control access for nonsubscribers: You could (1) designate a fixed subset of pages as viewable by nonsubscribers, or (2) allow any page in principle to be viewable, but specify a maximum number of pages that can be viewed by a nonsubscriber in a single session. (We'll assume the site is able to track the path followed by a visitor through the site.)

Your friends are experimenting with a way of restricting access that is different from and more subtle than either of these two options. They want nonsubscribers to be able to sample different sections of the Web site, so they designate certain subsets of the pages as constituting particular *zones*—for example, there can be a zone for pages on politics, a zone for pages on music, and so forth. It's possible for a page to belong to more than one zone. Now, as a nonsubscribing user passes through the site, the access policy allows him or her to visit one page from each zone, but an attempt by the user to access a second page from the same zone later in the browsing session will be disallowed. (Instead, the user will be directed to an ad suggesting that he or she become a subscriber.)

More formally, we can model the site as a directed graph $G = (V, E)$, in which the nodes represent Web pages and the edges represent directed hyperlinks. There is a distinguished *entry node* $s \in V$, and there are *zones* $Z_1, \ldots, Z_k \subseteq V$. A path P taken by a nonsubscriber is restricted to include at most one node from each zone Z_i.

One issue with this more complicated access policy is that it gets difficult to answer even basic questions about reachability, including: Is it possible for a nonsubscriber to visit a given node t? More precisely, we define the *Evasive Path Problem* as follows: Given G, $Z_1, \ldots, Z_k$, $s \in V$, and

a *destination node* $t \in V$, is there an s-t path in G that includes at most one node from each zone Z_i? Prove that Evasive Path is NP-complete.

13. A *combinatorial auction* is a particular mechanism developed by economists for selling a collection of items to a collection of potential buyers. (The Federal Communications Commission has studied this type of auction for assigning stations on the radio spectrum to broadcasting companies.)

Here's a simple type of combinatorial auction. There are n items for sale, labeled $I_1, \ldots, I_n$. Each item is indivisible and can only be sold to one person. Now, m different people place *bids*: The i^{th} bid specifies a subset S_i of the items, and an *offering price* x_i that the bidder is willing to pay for the items in the set S_i, as a single unit. (We'll represent this bid as the pair (S_i, x_i).)

An auctioneer now looks at the set of all m bids; she chooses to *accept* some of these bids and to *reject* the others. Each person whose bid i is accepted gets to take all the items in the corresponding set S_i. Thus the rule is that no two accepted bids can specify sets that contain a common item, since this would involve giving the same item to two different people.

The auctioneer collects the sum of the offering prices of all accepted bids. (Note that this is a "one-shot" auction; there is no opportunity to place further bids.) The auctioneer's goal is to collect as much money as possible.

Thus, the problem of *Winner Determination for Combinatorial Auctions* asks: Given items $I_1, \ldots, I_n$, bids $(S_1, x_1), \ldots, (S_m, x_m)$, and a bound B, is there a collection of bids that the auctioneer can accept so as to collect an amount of money that is at least B?

Example. Suppose an auctioneer decides to use this method to sell some excess computer equipment. There are four items labeled "`PC`," "`monitor`," "`printer`", and "`scanner`"; and three people place bids. Define

$$S_1 = \{\texttt{PC}, \texttt{monitor}\}, S_2 = \{\texttt{PC}, \texttt{printer}\}, S_3 = \{\texttt{monitor}, \texttt{printer}, \texttt{scanner}\}$$

and

$$x_1 = x_2 = x_3 = 1.$$

The bids are $(S_1, x_1), (S_2, x_2), (S_3, x_3)$, and the bound B is equal to 2.

Then the answer to this instance is no: The auctioneer can accept at most one of the bids (since any two bids have a desired item in common), and this results in a total monetary value of only 1.

Prove that the problem of Winner Determination in Combinatorial Auctions is NP-complete.

14. We've seen the Interval Scheduling Problem in Chapters 1 and 4. Here we consider a computationally much harder version of it that we'll call *Multiple Interval Scheduling*. As before, you have a processor that is available to run jobs over some period of time (e.g., 9 A.M. to 5 P.M).

People submit jobs to run on the processor; the processor can only work on one job at any single point in time. Jobs in this model, however, are more complicated than we've seen in the past: each job requires a *set* of intervals of time during which it needs to use the processor. Thus, for example, a single job could require the processor from 10 A.M. to 11 A.M., and again from 2 P.M. to 3 P.M.. If you accept this job, it ties up your processor during those two hours, but you could still accept jobs that need any other time periods (including the hours from 11 A.M. to 2 A.M.).

Now you're given a set of n jobs, each specified by a set of time intervals, and you want to answer the following question: For a given number k, is it possible to accept at least k of the jobs so that no two of the accepted jobs have any overlap in time?

Show that Multiple Interval Scheduling is NP-complete.

15. You're sitting at your desk one day when a FedEx package arrives for you. Inside is a cell phone that begins to ring, and you're not entirely surprised to discover that it's your friend Neo, whom you haven't heard from in quite a while. Conversations with Neo all seem to go the same way: He starts out with some big melodramatic justification for why he's calling, but in the end it always comes down to him trying to get you to volunteer your time to help with some problem he needs to solve.

This time, for reasons he can't go into (something having to do with protecting an underground city from killer robot probes), he and a few associates need to monitor radio signals at various points on the electromagnetic spectrum. Specifically, there are n different frequencies that need monitoring, and to do this they have available a collection of *sensors*.

There are two components to the monitoring problem.

- A set L of m geographic locations at which sensors can be placed; and
- A set S of b *interference sources*, each of which blocks certain frequencies at certain locations. Specifically, each interference source i is specified by a pair (F_i, L_i), where F_i is a subset of the frequencies and L_i is a subset of the locations; it signifies that (due to radio inter-

ference) a sensor placed at any location in the set L_i will not be able to receive signals on any frequency in the set F_i.

We say that a subset $L' \subseteq L$ of locations is *sufficient* if, for each of the n frequencies j, there is some location in L' where frequency j is not blocked by any interference source. Thus, by placing a sensor at each location in a sufficient set, you can successfully monitor each of the n frequencies.

They have k sensors, and hence they want to know whether there is a sufficient set of locations of size at most k. We'll call this an instance of the *Nearby Electromagnetic Observation Problem*: Given frequencies, locations, interference sources, and a parameter k, is there a sufficient set of size at most k?

Example. Suppose we have four frequencies $\{f_1, f_2, f_3, f_4\}$ and four locations $\{\ell_1, \ell_2, \ell_3, \ell_4\}$. There are three interference sources, with

$$(F_1, L_1) = (\{f_1, f_2\}, \{\ell_1, \ell_2, \ell_3\})$$
$$(F_2, L_2) = (\{f_3, f_4\}, \{\ell_3, \ell_4\})$$
$$(F_3, L_3) = (\{f_2, f_3\}, \{\ell_1\})$$

Then there is a sufficient set of size 2: We can choose locations ℓ_2 and ℓ_4 (since f_1 and f_2 are not blocked at ℓ_4, and f_3 and f_4 are not blocked at ℓ_2).

Prove that Nearby Electromagnetic Observation is NP-complete.

16. Consider the problem of reasoning about the identity of a set from the size of its intersections with other sets. You are given a finite set U of size n, and a collection $A_1, \ldots, A_m$ of subsets of U. You are also given numbers $c_1, \ldots, c_m$. The question is: Does there exist a set $X \subset U$ so that for each $i = 1, 2, \ldots, m$, the cardinality of $X \cap A_i$ is equal to c_i? We will call this an instance of the *Intersection Inference Problem*, with input U, $\{A_i\}$, and $\{c_i\}$.

Prove that Intersection Inference is NP-complete.

17. You are given a directed graph $G = (V, E)$ with weights w_e on its edges $e \in E$. The weights can be negative or positive. The *Zero-Weight-Cycle Problem* is to decide if there is a simple cycle in G so that the sum of the edge weights on this cycle is exactly 0. Prove that this problem is NP-complete.

18. You've been asked to help some organizational theorists analyze data on group decision-making. In particular, they've been looking at a dataset that consists of decisions made by a particular governmental policy committee, and they're trying to decide whether it's possible to identify a small set of influential members of the committee.

Here's how the committee works. It has a set $M = \{m_1, \ldots, m_n\}$ of n members, and over the past year it's voted on t different issues. On each issue, each member can vote either "Yes," "No," or "Abstain"; the overall

effect is that the committee presents an affirmative decision on the issue if the number of "Yes" votes is strictly greater than the number of "No" votes (the "Abstain" votes don't count for either side), and it delivers a negative decision otherwise.

Now we have a big table consisting of the vote cast by each committee member on each issue, and we'd like to consider the following definition. We say that a subset of the members $M' \subseteq M$ is *decisive* if, had we looked just at the votes cast by the members in M', the committee's decision on *every* issue would have been the same. (In other words, the overall outcome of the voting among the members in M' is the same on every issue as the overall outcome of the voting by the entire committee.) Such a subset can be viewed as a kind of "inner circle" that reflects the behavior of the committee as a whole.

Here's the question: Given the votes cast by each member on each issue, and given a parameter k, we want to know whether there is a decisive subset consisting of at most k members. We'll call this an instance of the *Decisive Subset Problem*.

Example. Suppose we have four committee members and three issues. We're looking for a decisive set of size at most $k = 2$, and the voting went as follows.

Issue #	m_1	m_2	m_3	m_4
Issue 1	Yes	Yes	Abstain	No
Issue 2	Abstain	No	No	Abstain
Issue 3	Yes	Abstain	Yes	Yes

Then the answer to this instance is "Yes," since members m_1 and m_3 constitute a decisive subset.

Prove that Decisive Subset is NP-complete.

19. Suppose you're acting as a consultant for the port authority of a small Pacific Rim nation. They're currently doing a multi-billion-dollar business per year, and their revenue is constrained almost entirely by the rate at which they can unload ships that arrive in the port.

Handling hazardous materials adds additional complexity to what is, for them, an already complicated task. Suppose a convoy of ships arrives in the morning and delivers a total of n cannisters, each containing a different kind of hazardous material. Standing on the dock is a set of m trucks, each of which can hold up to k containers.

Here are two related problems, which arise from different types of constraints that might be placed on the handling of hazardous materials. For each of the two problems, give one of the following two answers:

- A polynomial-time algorithm to solve it; or
- A proof that it is NP-complete.

(a) For each cannister, there is a specified subset of the trucks in which it may be safely carried. Is there a way to load all n cannisters into the m trucks so that no truck is overloaded, and each container goes in a truck that is allowed to carry it?

(b) In this different version of the problem, any cannister can be placed in any truck; however, there are certain pairs of cannisters that cannot be placed together in the same truck. (The chemicals they contain may react explosively if brought into contact.) Is there a way to load all n cannisters into the m trucks so that no truck is overloaded, and no two cannisters are placed in the same truck when they are not supposed to be?

20. There are many different ways to formalize the intuitive problem of *clustering*, where the goal is to divide up a collection of objects into groups that are "similar" to one another.

First, a natural way to express the input to a clustering problem is via a set of objects $p_1, p_2, \ldots, p_n$, with a numerical distance $d(p_i, p_j)$ defined on each pair. (We require only that $d(p_i, p_i) = 0$; that $d(p_i, p_j) > 0$ for distinct p_i and p_j; and that distances are symmetric: $d(p_i, p_j) = d(p_j, p_i)$.)

In Section 4.7, earlier in the book, we considered one reasonable formulation of the clustering problem: Divide the objects into k sets so as to *maximize* the minimum distance between any pair of objects in distinct clusters. This turns out to be solvable by a nice application of the Minimum Spanning Tree Problem.

A different but seemingly related way to formalize the clustering problem would be as follows: Divide the objects into k sets so as to *minimize* the maximum distance between any pair of objects in the same cluster. Note the change. Where the formulation in the previous paragraph sought clusters so that no two were "close together," this new formulation seeks clusters so that none of them is too "wide"—that is, no cluster contains two points at a large distance from each other.

Given the similarities, it's perhaps surprising that this new formulation is computationally hard to solve optimally. To be able to think about this in terms of NP-completeness, let's write it first as a yes/no decision problem. Given n objects $p_1, p_2, \ldots, p_n$ with distances on them as above,

and a bound B, we define the *Low-Diameter Clustering Problem* as follows: Can the objects be partitioned into k sets, so that no two points in the same set are at a distance greater than B from each other?

Prove that Low-Diameter Clustering is NP-complete.

21. After a few too many days immersed in the popular entrepreneurial self-help book *Mine Your Own Business*, you've come to the realization that you need to upgrade your office computing system. This, however, leads to some tricky problems.

In configuring your new system, there are k *components* that must be selected: the operating system, the text editing software, the e-mail program, and so forth; each is a separate component. For the j^{th} component of the system, you have a set A_j of options; and a *configuration* of the system consists of a selection of one element from each of the sets of options $A_1, A_2, \ldots, A_k$.

Now the trouble arises because certain pairs of options from different sets may not be compatible. We say that option $x_i \in A_i$ and option $x_j \in A_j$ form an *incompatible pair* if a single system cannot contain them both. (For example, Linux (as an option for the operating system) and Microsoft Word (as an option for the text-editing software) form an incompatible pair.) We say that a configuration of the system is *fully compatible* if it consists of elements $x_1 \in A_1, x_2 \in A_2, \ldots x_k \in A_k$ such that none of the pairs (x_i, x_j) is an incompatible pair.

We can now define the *Fully Compatible Configuration (FCC) Problem*. An instance of FCC consists of disjoint sets of options $A_1, A_2, \ldots, A_k$, and a set P of *incompatible pairs* (x, y), where x and y are elements of different sets of options. The problem is to decide whether there exists a fully compatible configuration: a selection of an element from each option set so that no pair of selected elements belongs to the set P.

Example. Suppose $k = 3$, and the sets A_1, A_2, A_3 denote options for the operating system, the text-editing software, and the e-mail program, respectively. We have

$$A_1 = \{\texttt{Linux}, \texttt{Windows NT}\},$$
$$A_2 = \{\texttt{emacs}, \texttt{Word}\},$$
$$A_3 = \{\texttt{Outlook}, \texttt{Eudora}, \texttt{rmail}\},$$

with the set of incompatible pairs equal to

$$P = \{(\texttt{Linux}, \texttt{Word}), (\texttt{Linux}, \texttt{Outlook}), (\texttt{Word}, \texttt{rmail})\}.$$

Then the answer to the decision problem in this instance of FCC is yes—for example, the choices $\texttt{Linux} \in A_1$, $\texttt{emacs} \in A_2$, $\texttt{rmail} \in A_3$ is a fully compatible configuration according to the definitions above.

Prove that Fully Compatible Configuration is NP-complete.

22. Suppose that someone gives you a black-box algorithm $\mathcal{A}$ that takes an undirected graph $G = (V, E)$, and a number k, and behaves as follows.

- If G is not connected, it simply returns "G is not connected."
- If G is connected and has an independent set of size at least k, it returns "yes."
- If G is connected and does not have an independent set of size at least k, it returns "no."

Suppose that the algorithm $\mathcal{A}$ runs in time polynomial in the size of G and k.

Show how, using calls to $\mathcal{A}$, you could then solve the Independent Set Problem in polynomial time: Given an arbitrary undirected graph G, and a number k, does G contain an independent set of size at least k?

23. Given a set of finite binary strings $S = \{s_1, \ldots, s_k\}$, we say that a string u is a *concatenation* over S if it is equal to $s_{i_1} s_{i_2} \cdots s_{i_t}$ for some indices $i_1, \ldots, i_t \in \{1, \ldots, k\}$.

A friend of yours is considering the following problem: Given two sets of finite binary strings, $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, does there exist any string u so that u is both a concatenation over A and a concatenation over B?

Your friend announces, "At least the problem is in $\mathcal{NP}$, since I would just have to exhibit such a string u in order to prove the answer is yes." You point out (politely, of course) that this is a completely inadequate explanation; how do we know that the shortest such string u doesn't have length exponential in the size of the input, in which case it would not be a polynomial-size certificate?

However, it turns out that this claim can be turned into a proof of membership in $\mathcal{NP}$. Specifically, prove the following statement.

> *If there is a string u that is a concatenation over both A and B, then there is such a string whose length is bounded by a polynomial in the sum of the lengths of the strings in $A \cup B$.*

24. Let $G = (V, E)$ be a bipartite graph; suppose its nodes are partitioned into sets X and Y so that each edge has one end in X and the other in Y. We define an (a, b)-*skeleton* of G to be a set of edges $E' \subseteq E$ so that *at most*

a nodes in X are incident to an edge in E', and *at least* b nodes in Y are incident to an edge in E'.

Show that, given a bipartite graph G and numbers a and b, it is NP-complete to decide whether G has an (a, b)-skeleton.

25. For functions $g_1, \ldots, g_\ell$, we define the function $\max(g_1, \ldots, g_\ell)$ via

$$[\max(g_1, \ldots, g_\ell)](x) = \max(g_1(x), \ldots, g_\ell(x)).$$

Consider the following problem. You are given n piecewise linear, continuous functions $f_1, \ldots, f_n$ defined over the interval $[0, t]$ for some integer t. You are also given an integer B. You want to decide: Do there exist k of the functions $f_{i_1}, \ldots, f_{i_k}$ so that

$$\int_0^t [\max(f_{i_1}, \ldots, f_{i_k})](x)\, dx \geq B?$$

Prove that this problem is NP-complete.

26. You and a friend have been trekking through various far-off parts of the world and have accumulated a big pile of souvenirs. At the time you weren't really thinking about which of these you were planning to keep and which your friend was going to keep, but now the time has come to divide everything up.

Here's a way you could go about doing this. Suppose there are n objects, labeled $1, 2, \ldots, n$, and object i has an agreed-upon *value* x_i. (We could think of this, for example, as a monetary resale value; the case in which you and your friend don't agree on the value is something we won't pursue here.) One reasonable way to divide things would be to look for a *partition* of the objects into two sets, so that the total value of the objects in each set is the same.

This suggests solving the following *Number Partitioning Problem*. You are given positive integers $x_1, \ldots, x_n$; you want to decide whether the numbers can be partitioned into two sets S_1 and S_2 with the same sum:

$$\sum_{x_i \in S_1} x_i = \sum_{x_j \in S_2} x_j.$$

Show that Number Partitioning is NP-complete.

27. Consider the following problem. You are given positive integers $x_1, \ldots, x_n$, and numbers k and B. You want to know whether it is possible to *partition*

the numbers $\{x_i\}$ into k sets $S_1, \ldots, S_k$ so that the squared sums of the sets add up to at most B:

$$\sum_{i=1}^{k} \left(\sum_{x_j \in S_i} x_j \right)^2 \leq B.$$

Show that this problem is NP-complete.

28. The following is a version of the Independent Set Problem. You are given a graph $G = (V, E)$ and an integer k. For this problem, we will call a set $I \subset V$ *strongly independent* if, for any two nodes $v, u \in I$, the edge (v, u) does not belong to E, and there is also no path of two edges from u to v, that is, there is no node w such that both $(u, w) \in E$ and $(w, v) \in E$. The Strongly Independent Set Problem is to decide whether G has a strongly independent set of size at least k.

Prove that the Strongly Independent Set Problem is NP-complete.

29. You're configuring a large network of workstations, which we'll model as an undirected graph G; the nodes of G represent individual workstations and the edges represent direct communication links. The workstations all need access to a common *core database*, which contains data necessary for basic operating system functions.

You could replicate this database on each workstation; this would make lookups very fast from any workstation, but you'd have to manage a huge number of copies. Alternately, you could keep a single copy of the database on one workstation and have the remaining workstations issue requests for data over the network G; but this could result in large delays for a workstation that's many hops away from the site of the database.

So you decide to look for the following compromise: You want to maintain a small number of copies, but place them so that any workstation either has a copy of the database or is connected by a direct link to a workstation that has a copy of the database. In graph terminology, such a set of locations is called a *dominating set*.

Thus we phrase the *Dominating Set Problem* as follows. Given the network G, and a number k, is there a way to place k copies of the database at k different nodes so that every node either has a copy of the database or is connected by a direct link to a node that has a copy of the database?

Show that Dominating Set is NP-complete.

30. One thing that's not always apparent when thinking about traditional "continuous math" problems is the way discrete, combinatorial issues

of the kind we're studying here can creep into what look like standard calculus questions.

Consider, for example, the traditional problem of minimizing a one-variable function like $f(x) = 3 + x - 3x^2$ over an interval like $x \in [0, 1]$. The derivative has a zero at $x = 1/6$, but this in fact is a maximum of the function, not a minimum; to get the minimum, one has to heed the standard warning to check the values on the boundary of the interval as well. (The minimum is in fact achieved on the boundary, at $x = 1$.)

Checking the boundary isn't such a problem when you have a function in one variable; but suppose we're now dealing with the problem of minimizing a function in n variables $x_1, x_2, \ldots, x_n$ over the unit cube, where each of $x_1, x_2, \ldots, x_n \in [0, 1]$. The minimum may be achieved on the interior of the cube, but it may be achieved on the boundary; and this latter prospect is rather daunting, since the boundary consists of 2^n "corners" (where each x_i is equal to either 0 or 1) as well as various pieces of other dimensions. Calculus books tend to get suspiciously vague around here, when trying to describe how to handle multivariable minimization problems in the face of this complexity.

It turns out there's a reason for this: Minimizing an n-variable function over the unit cube in n dimensions is as hard as an NP-complete problem. To make this concrete, let's consider the special case of polynomials with integer coefficients over n variables $x_1, x_2, \ldots, x_n$. To review some terminology, we say a *monomial* is a product of a real-number coefficient c and each variable x_i raised to some nonnegative integer power a_i; we can write this as $cx_1^{a_1}x_2^{a_2} \cdots x_n^{a_n}$. (For example, $2x_1^2x_2x_3^4$ is a monomial.) A *polynomial* is then a sum of a finite set of monomials. (For example, $2x_1^2x_2x_3^4 + x_1x_3 - 6x_2^2x_3^2$ is a polynomial.)

We define the *Multivariable Polynomial Minimization Problem* as follows: Given a polynomial in n variables with integer coefficients, and given an integer bound B, is there a choice of real numbers $x_1, x_2, \ldots, x_n \in [0, 1]$ that causes the polynomial to achieve a value that is $\leq B$?

Choose a problem Y from this chapter that is known to be NP-complete and show that

$$Y \leq_P \text{Multivariable Polynomial Minimization.}$$

31. Given an undirected graph $G = (V, E)$, a *feedback set* is a set $X \subseteq V$ with the property that $G - X$ has no cycles. The *Undirected Feedback Set Problem* asks: Given G and k, does G contain a feedback set of size at most k? Prove that Undirected Feedback Set is NP-complete.

32. The mapping of genomes involves a large array of difficult computational problems. At the most basic level, each of an organism's chromosomes can be viewed as an extremely long string (generally containing millions of symbols) over the four-letter alphabet $\{A, C, G, T\}$. One family of approaches to genome mapping is to generate a large number of short, overlapping snippets from a chromosome, and then to infer the full long string representing the chromosome from this set of overlapping substrings.

While we won't go into these string assembly problems in full detail, here's a simplified problem that suggests some of the computational difficulty one encounters in this area. Suppose we have a set $S = \{s_1, s_2, \ldots, s_n\}$ of short DNA strings over a q-letter alphabet; and each string s_i has length 2ℓ, for some number $\ell \geq 1$. We also have a library of additional strings $T = \{t_1, t_2, \ldots, t_m\}$ over the same alphabet; each of these also has length 2ℓ. In trying to assess whether the string s_b might come directly after the string s_a in the chromosome, we will look to see whether the library T contains a string t_k so that the first ℓ symbols in t_k are equal to the last ℓ symbols in s_a, and the last ℓ symbols in t_k are equal to the first ℓ symbols in s_b. If this is possible, we will say that t_k *corroborates* the pair (s_a, s_b). (In other words, t_k could be a snippet of DNA that straddled the region in which s_b directly followed s_a.)

Now we'd like to concatenate all the strings in S in some order, one after the other with no overlaps, so that each consecutive pair is corroborated by some string in the library T. That is, we'd like to order the strings in S as $s_{i_1}, s_{i_2}, \ldots, s_{i_n}$, where $i_1, i_2, \ldots, i_n$ is a permutation of $\{1, 2, \ldots, n\}$, so that for each $j = 1, 2, \ldots, n-1$, there is a string t_k that corroborates the pair $(s_{i_j}, s_{i_{j+1}})$. (The same string t_k can be used for more than one consecutive pair in the concatenation.) If this is possible, we will say that the set S has a *perfect assembly*.

Given sets S and T, the *Perfect Assembly Problem* asks: Does S have a perfect assembly with respect to T? Prove that Perfect Assembly is NP-complete.

Example. Suppose the alphabet is $\{A, C, G, T\}$, the set $S = \{AG, TC, TA\}$, and the set $T = \{AC, CA, GC, GT\}$ (so each string has length $2\ell = 2$). Then the answer to this instance of Perfect Assembly is yes: We can concatenate the three strings in S in the order $TCAGTA$ (so $s_{i_1} = s_2$, $s_{i_2} = s_1$, and $s_{i_3} = s_3$). In this order, the pair (s_{i_1}, s_{i_2}) is corroborated by the string CA in the library T, and the pair (s_{i_2}, s_{i_3}) is corroborated by the string GT in the library T.

33. In a barter economy, people trade goods and services directly, without money as an intermediate step in the process. Trades happen when each

party views the set of goods they're getting as more valuable than the set of goods they're giving in return. Historically, societies tend to move from barter-based to money-based economies; thus various online systems that have been experimenting with barter can be viewed as intentional attempts to regress to this earlier form of economic interaction. In doing this, they've rediscovered some of the inherent difficulties with barter relative to money-based systems. One such difficulty is the complexity of identifying opportunities for trading, even when these opportunities exist.

To model this complexity, we need a notion that each person assigns a *value* to each object in the world, indicating how much this object would be worth to them. Thus we assume there is a set of n people $p_1, \ldots, p_n$, and a set of m distinct objects $a_1, \ldots, a_m$. Each object is owned by one of the people. Now each person p_i has a *valuation function* v_i, defined so that $v_i(a_j)$ is a nonnegative number that specifies how much object a_j is worth to p_i—the larger the number, the more valuable the object is to the person. Note that everyone assigns a valuation to each object, including the ones they don't currently possess, and different people can assign very different valuations to the same object.

A two-person trade is possible in a system like this when there are people p_i and p_j, and subsets of objects A_i and A_j possessed by p_i and p_j, respectively, so that each person would prefer the objects in the subset they don't currently have. More precisely,

- p_i's total valuation for the objects in A_j exceeds his or her total valuation for the objects in A_i, and
- p_j's total valuation for the objects in A_i exceeds his or her total valuation for the objects in A_j.

(Note that A_i doesn't have to be *all* the objects possessed by p_i (and likewise for A_j); A_i and A_j can be arbitrary subsets of their possessions that meet these criteria.)

Suppose you are given an instance of a barter economy, specified by the above data on people's valuations for objects. (To prevent problems with representing real numbers, we'll assume that each person's valuation for each object is a natural number.) Prove that the problem of determining whether a two-person trade is possible is NP-complete.

34. In the 1970s, researchers including Mark Granovetter and Thomas Schelling in the mathematical social sciences began trying to develop models of certain kinds of collective human behaviors: Why do particular fads catch on while others die out? Why do particular technological innovations achieve widespread adoption, while others remain focused

on a small group of users? What are the dynamics by which rioting and looting behavior sometimes (but only rarely) emerges from a crowd of angry people? They proposed that these are all examples of *cascade processes*, in which an individual's behavior is highly influenced by the behaviors of his or her friends, and so if a few individuals instigate the process, it can spread to more and more people and eventually have a very wide impact. We can think of this process as being like the spread of an illness, or a rumor, jumping from person to person.

The most basic version of their models is the following. There is some underlying *behavior* (e.g., playing ice hockey, owning a cell phone, taking part in a riot), and at any point in time each person is either an *adopter* of the behavior or a *nonadopter*. We represent the population by a directed graph $G = (V, E)$ in which the nodes correspond to people and there is an edge (v, w) if person v has influence over the behavior of person w: If person v adopts the behavior, then this helps induce person w to adopt it as well. Each person w also has a given *threshold* $\theta(w) \in [0, 1]$, and this has the following meaning: At any time when at least a $\theta(w)$ fraction of the nodes with edges to w are adopters of the behavior, the node w will become an adopter as well.

Note that nodes with lower thresholds are more easily convinced to adopt the behavior, while nodes with higher thresholds are more conservative. A node w with threshold $\theta(w) = 0$ will adopt the behavior immediately, with no inducement from friends. Finally, we need a convention about nodes with no incoming edges: We will say that they become adopters if $\theta(w) = 0$, and cannot become adopters if they have any larger threshold.

Given an instance of this model, we can simulate the spread of the behavior as follows.

```
Initially, set all nodes w with θ(w) = 0 to be adopters
   (All other nodes start out as nonadopters)
Until there is no change in the set of adopters:
   For each nonadopter w simultaneously:
      If at least a θ(w) fraction of nodes with edges to w are
         adopters then
           w becomes an adopter
      Endif
   Endfor
End
Output the final set of adopters
```

Note that this process terminates, since there are only n individuals total, and at least one new person becomes an adopter in each iteration.

Now, in the last few years, researchers in marketing and data mining have looked at how a model like this could be used to investigate "word-of-mouth" effects in the success of new products (the so-called *viral marketing* phenomenon). The idea here is that the behavior we're concerned with is the use of a new product; we may be able to convince a few key people in the population to try out this product, and hope to trigger as large a cascade as possible.

Concretely, suppose we choose a set of nodes $S \subseteq V$ and we reset the threshold of each node in S to 0. (By convincing them to try the product, we've ensured that they're adopters.) We then run the process described above, and see how large the final set of adopters is. Let's denote the size of this final set of adopters by $f(S)$ (note that we write it as a function of S, since it naturally depends on our choice of S). We could think of $f(S)$ as the *influence* of the set S, since it captures how widely the behavior spreads when "seeded" at S.

The goal, if we're marketing a product, is to find a small set S whose influence $f(S)$ is as large as possible. We thus define the *Influence Maximization Problem* as follows: Given a directed graph $G = (V, E)$, with a threshold value at each node, and parameters k and b, is there a set S of at most k nodes for which $f(S) \geq b$?

Prove that Influence Maximization is NP-complete.

Example. Suppose our graph $G = (V, E)$ has five nodes $\{a, b, c, d, e\}$, four edges $(a, b), (b, c), (e, d), (d, c)$, and all node thresholds equal to $2/3$. Then the answer to the Influence Maximization instance defined by G, with $k = 2$ and $b = 5$, is yes: We can choose $S = \{a, e\}$, and this will cause the other three nodes to become adopters as well. (This is the only choice of S that will work here. For example, if we choose $S = \{a, d\}$, then b and c will become adopters, but e won't; if we choose $S = \{a, b\}$, then none of c, d, or e will become adopters.)

35. Three of your friends work for a large computer-game company, and they've been working hard for several months now to get their proposal for a new game, *Droid Trader!*, approved by higher management. In the process, they've had to endure all sorts of discouraging comments, ranging from "You're really going to have to work with Marketing on the name" to "Why don't you emphasize the parts where people get to kick each other in the head?"

At this point, though, it's all but certain that the game is really heading into production, and your friends come to you with one final

issue that's been worrying them: What if the overall premise of the game is too simple, so that players get really good at it and become bored too quickly?

It takes you a while, listening to their detailed description of the game, to figure out what's going on; but once you strip away the space battles, kick-boxing interludes, and Stars-Wars-inspired pseudo-mysticism, the basic idea is as follows. A player in the game controls a spaceship and is trying to make money buying and selling droids on different planets. There are n different types of droids and k different planets. Each planet p has the following properties: there are $s(j, p) \geq 0$ droids of type j available for sale, at a fixed price of $x(j, p) \geq 0$ each, for $j = 1, 2, \ldots, n$; and there is a demand for $d(j, p) \geq 0$ droids of type j, at a fixed price of $y(j, p) \geq 0$ each. (We will assume that a planet does not simultaneously have both a positive supply and a positive demand for a single type of droid; so for each j, at least one of $s(j, p)$ or $d(j, p)$ is equal to 0.)

The player begins on planet s with z units of money and must end at planet t; there is a directed acyclic graph G on the set of planets, such that s-t paths in G correspond to valid routes by the player. (G is chosen to be acyclic to prevent arbitrarily long games.) For a given s-t path P in G, the player can engage in transactions as follows. Whenever the player arrives at a planet p on the path P, she can buy up to $s(j, p)$ droids of type j for $x(j, p)$ units of money each (provided she has sufficient money on hand) and/or sell up to $d(j, p)$ droids of type j for $y(j, p)$ units of money each (for $j = 1, 2, \ldots, n$). The player's *final score* is the total amount of money she has on hand when she arrives at planet t. (There are also bonus points based on space battles and kick-boxing, which we'll ignore for the purposes of formulating this question.)

So basically, the underlying problem is to achieve a high score. In other words, given an instance of this game, with a directed acyclic graph G on a set of planets, all the other parameters described above, and also a target bound B, is there a path P in G and a sequence of transactions on P so that the player ends with a final score that is at least B? We'll call this an instance of the *High-Score-on-Droid-Trader! Problem*, or HSoDT! for short.

Prove that HSoDT! is NP-complete, thereby guaranteeing (assuming $P \neq NP$) that there isn't a simple strategy for racking up high scores on your friends' game.

36. Sometimes you can know people for years and never really understand them. Take your friends Raj and Alanis, for example. Neither of them is a morning person, but now they're getting up at 6 AM every day to visit

local farmers' markets, gathering fresh fruits and vegetables for the new health-food restaurant they've opened, Chez Alanisse.

In the course of trying to save money on ingredients, they've come across the following thorny problem. There is a large set of n possible raw ingredients they could buy, $I_1, I_2, \ldots, I_n$ (e.g., bundles of dandelion greens, jugs of rice vinegar, and so forth). Ingredient I_j must be purchased in units of size $s(j)$ grams (any purchase must be for a whole number of units), and it costs $c(j)$ dollars per unit. Also, it remains safe to use for $t(j)$ days from the date of purchase.

Now, over the next k days, they want to make a set of k different daily specials, one each day. (The order in which they schedule the specials is up to them.) The i^{th} daily special uses a subset $S_i \subseteq \{I_1, I_2, \ldots, I_n\}$ of the raw ingredients. Specifically, it requires $a(i, j)$ grams of ingredient I_j. And there's a final constraint: The restaurant's rabidly loyal customer base only remains rabidly loyal if they're being served the freshest meals available; so for each daily special, the ingredients S_i are partitioned into two subsets: those that must be purchased on the very day when the daily special is being offered, and those that can be used any day while they're still safe. (For example, the mesclun-basil salad special needs to be made with basil that has been purchased that day; but the arugula-basil pesto with Cornell dairy goat cheese special can use basil that is several days old, as long as it is still safe.)

This is where the opportunity to save money on ingredients comes up. Often, when they buy a unit of a certain ingredient I_j, they don't need the whole thing for the special they're making that day. Thus, if they can follow up quickly with another special that uses I_j but doesn't require it to be fresh that day, then they can save money by not having to purchase I_j again. Of course, scheduling the basil recipes close together may make it harder to schedule the goat cheese recipes close together, and so forth—this is where the complexity comes in.

So we define the *Daily Special Scheduling Problem* as follows: Given data on ingredients and recipes as above, and a budget x, is there a way to schedule the k daily specials so that the total money spent on ingredients over the course of all k days is at most x?

Prove that *Daily Special Scheduling* is NP-complete.

37. There are those who insist that the initial working title for Episode XXVII of the Star Wars series was "$\mathcal{P} = \mathcal{NP}$"—but this is surely apocryphal. In any case, if you're so inclined, it's easy to find NP-complete problems lurking just below the surface of the original *Star Wars* movies.

Consider the problem faced by Luke, Leia, and friends as they tried to make their way from the Death Star back to the hidden Rebel base. We can view the galaxy as an undirected graph $G = (V, E)$, where each node is a star system and an edge (u, v) indicates that one can travel directly from u to v. The Death Star is represented by a node s, the hidden Rebel base by a node t. Certain edges in this graph represent longer distances than others; thus each edge e has an integer *length* $\ell_e \geq 0$. Also, certain edges represent routes that are more heavily patrolled by evil Imperial spacecraft; so each edge e also has an integer *risk* $r_e \geq 0$, indicating the expected amount of damage incurred from special-effects-intensive space battles if one traverses this edge.

It would be safest to travel through the outer rim of the galaxy, from one quiet upstate star system to another; but then one's ship would run out of fuel long before getting to its destination. Alternately, it would be quickest to plunge through the cosmopolitan core of the galaxy; but then there would be far too many Imperial spacecraft to deal with. In general, for any path P from s to t, we can define its *total length* to be the sum of the lengths of all its edges; and we can define its *total risk* to be the sum of the risks of all its edges.

So Luke, Leia, and company are looking at a complex type of shortest-path problem in this graph: they need to get from s to t along a path whose total length and total risk are *both* reasonably small. In concrete terms, we can phrase the *Galactic Shortest-Path Problem* as follows: Given a setup as above, and integer bounds L and R, is there a path from s to t whose total length is at most L, *and* whose total risk is at most R?

Prove that Galactic Shortest Path is NP-complete.

38. Consider the following version of the Steiner Tree Problem, which we'll refer to as *Graphical Steiner Tree*. You are given an undirected graph $G = (V, E)$, a set $X \subseteq V$ of vertices, and a number k. You want to decide whether there is a set $F \subseteq E$ of at most k edges so that in the graph (V, F), X belongs to a single connected component.

Show that Graphical Steiner Tree is NP-complete.

39. The *Directed Disjoint Paths Problem* is defined as follows. We are given a directed graph G and k pairs of nodes $(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)$. The problem is to decide whether there exist node-disjoint paths $P_1, P_2, \ldots, P_k$ so that P_i goes from s_i to t_i.

Show that Directed Disjoint Paths is NP-complete.

40. Consider the following problem that arises in the design of broadcasting schemes for networks. We are given a directed graph $G = (V, E)$, with a

designated node $r \in V$ and a designated set of "target nodes" $T \subseteq V - \{r\}$. Each node v has a *switching time* s_v, which is a positive integer.

At time 0, the node r generates a message that it would like every node in T to receive. To accomplish this, we want to find a scheme whereby r tells some of its neighbors (in sequence), who in turn tell some of their neighbors, and so on, until every node in T has received the message. More formally, a *broadcast scheme* is defined as follows. Node r may send a copy of the message to one of its neighbors at time 0; this neighbor will receive the message at time 1. In general, at time $t \geq 0$, any node v that has already received the message may send a copy of the message to one of its neighbors, provided it has not sent a copy of the message in any of the time steps $t - s_v + 1, t - s_v + 2, \ldots, t - 1$. (This reflects the role of the *switching time*; v needs a pause of $s_v - 1$ steps between successive sendings of the message. Note that if $s_v = 1$, then no restriction is imposed by this.)

The *completion time* of the broadcast scheme is the minimum time t by which all nodes in T have received the message. The *Broadcast Time Problem* is the following: Given the input described above, and a bound b, is there a broadcast scheme with completion time at most b?

Prove that Broadcast Time is NP-complete.

Example. Suppose we have a directed graph $G = (V, E)$, with $V = \{r, a, b, c\}$; edges (r, a), (a, b), (r, c); the set $T = \{b, c\}$; and switching time $s_v = 2$ for each $v \in V$. Then a broadcast scheme with minimum completion time would be as follows: r sends the message to a at time 0; a sends the message to b at time 1; r sends the message to c at time 2; and the scheme completes at time 3 when c receives the message. (Note that a can send the message as soon as it receives it at time 1, since this is its first sending of the message; but r cannot send the message at time 1 since $s_r = 2$ and it sent the message at time 0.)

41. Given a directed graph G, a *cycle cover* is a set of node-disjoint cycles so that each node of G belongs to a cycle. The *Cycle Cover Problem* asks whether a given directed graph has a cycle cover.

(a) Show that the Cycle Cover Problem can be solved in polynomial time. (*Hint:* Use Bipartite Matching.)

(b) Suppose we require each cycle to have at most three edges. Show that determining whether a graph G has such a cycle cover is NP-complete.

42. Suppose you're consulting for a company in northern New Jersey that designs communication networks, and they come to you with the follow-

ing problem. They're studying a specific n-node communication network, modeled as a directed graph $G = (V, E)$. For reasons of fault tolerance, they want to divide up G into as many virtual "domains" as possible: A *domain* in G is a set X of nodes, of size at least 2, so that for each pair of nodes $u, v \in X$ there are directed paths from u to v and v to u that are contained entirely in X.

Show that the following *Domain Decomposition Problem* is NP-complete. Given a directed graph $G = (V, E)$ and a number k, can V be *partitioned* into at least k sets, each of which is a domain?

Notes and Further Reading

In the notes to Chapter 2, we described some of the early work on formalizing computational efficiency using polynomial time; NP-completeness evolved out of this work and grew into its central role in computer science following the papers of Cook (1971), Levin (1973), and Karp (1972). Edmonds (1965) is credited with drawing particular attention to the class of problems in $\mathcal{NP} \cap$ co-$\mathcal{NP}$—those with "good characterizations." His paper also contains the explicit conjecture that the Traveling Salesman Problem cannot be solved in polynomial time, thereby prefiguring the $P \neq NP$ question. Sipser (1992) is a useful guide to all of this historical context.

The book by Garey and Johnson (1979) provides extensive material on NP-completeness and concludes with a very useful catalog of known NP-complete problems. While this catalog, necessarily, only covers what was known at the time of the book's publication, it is still a very useful reference when one encounters a new problem that looks like it might be NP-complete. In the meantime, the space of known NP-complete problems has continued to expand dramatically; as Christos Papadimitriou said in a lecture, "Roughly 6,000 papers every year contain an NP-completeness result. That means another NP-complete problem has been discovered since lunch." (His lecture was at 2:00 in the afternoon.)

One can interpret NP-completeness as saying that each individual NP-complete problem contains the entire complexity of NP hidden inside it. A concrete reflection of this is the fact that several of the NP-complete problems we discuss here are the subject of entire books: the Traveling Salesman is the subject of Lawler et al. (1985); Graph Coloring is the subject of Jensen and Toft (1995); and the Knapsack Problem is the subject of Martello and Toth (1990). NP-completeness results for scheduling problems are discussed in the survey by Lawler et al. (1993).

Notes on the Exercises A number of the exercises illustrate further problems that emerged as paradigmatic examples early in the development of NP-completeness; these include Exercises 5, 26, 29, 31, 38, 39, 40, and 41. Exercise 33 is based on discussions with Daniel Golovin, and Exercise 34 is based on our work with David Kempe. Exercise 37 is an example of the class of *Bicriteria Shortest-Path problems*; its motivating application here was suggested by Maverick Woo.