(13.51) *There is a probability space* $(\Omega, \mathcal{S}, \Pr)$*, satisfying the required closure and consistency properties, such that* Ω *is the sample space defined above,* $\mathcal{E}_\sigma \in \mathcal{S}$ *for each finite sequence* σ*, and* $\Pr\left[\mathcal{E}_\sigma\right] = w(x_1)w(x_2)\cdots w(x_s)$*.*

Once we have this fact, the closure of $\mathcal{S}$ under complement and countable union, and the consistency of Pr with respect to these operations, allow us to compute probabilities of essentially any "reasonable" subset of Ω.

In our infinite sample space Ω, with events and probabilities defined as above, we encounter a phenomenon that does not naturally arise with finite sample spaces. Suppose the set X used to generate Ω is equal to $\{0, 1\}$, and $w(0) = w(1) = 1/2$. Let $\mathcal{E}$ denote the set consisting of all sequences that contain at least one entry equal to 1. (Note that $\mathcal{E}$ omits the "all-0" sequence.) We observe that $\mathcal{E}$ is an event in $\mathcal{S}$, since we can define σ_i to be the sequence of $i-1$ 0s followed by a 1, and observe that $\mathcal{E} = \cup_{i=1}^{\infty}\mathcal{E}_{\sigma_i}$. Moreover, all the events $\mathcal{E}_{\sigma_i}$ are pairwise disjoint, and so

$$\Pr\left[\mathcal{E}\right] = \sum_{i=1}^{\infty} \Pr\left[\mathcal{E}_{\sigma_i}\right] = \sum_{i=1}^{\infty} 2^{-i} = 1.$$

Here, then, is the phenomenon: It's possible for an event to have probability 1 even when it's not equal to the whole sample space Ω. Similarly, $\Pr\left[\overline{\mathcal{E}}\right] = 1 - \Pr\left[\mathcal{E}\right] = 0$, and so we see that it's possible for an event to have probability 0 even when it's not the empty set. There is nothing wrong with any of these results; in a sense, it's a necessary step if we want probabilities defined over infinite sets to make sense. It's simply that in such cases, we should be careful to distinguish between the notion that an event has probability 0 and the intuitive idea that the event "can't happen."

Solved Exercises

Solved Exercise 1

Suppose we have a collection of small, low-powered devices scattered around a building. The devices can exchange data over short distances by wireless communication, and we suppose for simplicity that each device has enough range to communicate with d other devices. Thus we can model the wireless connections among these devices as an undirected graph $G = (V, E)$ in which each node is incident to exactly d edges.

Now we'd like to give some of the nodes a stronger *uplink transmitter* that they can use to send data back to a base station. Giving such a transmitter to every node would ensure that they can all send data like this, but we can achieve this while handing out fewer transmitters. Suppose that we find a subset S of the nodes with the property that every node in $V - S$ is adjacent to a node in S. We call such a set S a *dominating set*, since it "dominates" all other nodes in the graph. If we give uplink transmitters only to the nodes in a dominating set S, we can still extract data from all nodes: Any node $u \notin S$ can choose a neighbor $v \in S$, send its data to v, and have v relay the data back to the base station.

The issue is now to find a dominating set S of minimum possible size, since this will minimize the number of uplink transmitters we need. This is an NP-hard problem; in fact, proving this is the crux of Exercise 29 in Chapter 8. (It's also worth noting here the difference between dominating sets and vertex covers: in a dominating set, it is fine to have an edge (u, v) with neither u nor v in the set S as long as both u and v have neighbors in S. So, for example, a graph consisting of three nodes all connected by edges has a dominating set of size 1, but no vertex cover of size 1.)

Despite the NP-hardness, it's important in applications like this to find as small a dominating set as one can, even if it is not optimal. We will see here that a simple randomized strategy can be quite effective. Recall that in our graph G, each node is incident to exactly d edges. So clearly any dominating set will need to have size at least $\frac{n}{d+1}$, since each node we place in a dominating set can take care only of itself and its d neighbors. We want to show that a random selection of nodes will, in fact, get us quite close to this simple lower bound.

Specifically, show that for some constant c, a set of $\frac{cn \log n}{d+1}$ nodes chosen uniformly at random from G will be a dominating set with high probability. (In other words, this completely random set is likely to form a dominating set that is only $O(\log n)$ times larger than our simple lower bound of $\frac{n}{d+1}$.)

Solution Let $k = \frac{cn \log n}{d}$, where we will choose the constant c later, once we have a better idea of what's going on. Let $\mathcal{E}$ be the event that a random choice of k nodes is a dominating set for G. To make the analysis simpler, we will consider a model in which the nodes are selected one at a time, and the same node may be selected twice (if it happens to be picked twice by our sequence of random choices).

Now we want to show that if c (and hence k) is large enough, then $\Pr[\mathcal{E}]$ is close to 1. But $\mathcal{E}$ is a very complicated-looking event, so we begin by breaking it down into much simpler events whose probabilities we can analyze more easily.

To start with, we say that a node w *dominates* a node v if w is a neighbor of v, or $w = v$. We say that a set S dominates a node v if some element of S dominates v. (These definitions let us say that a dominating set is simply a set of nodes that dominates every node in the graph.) Let $\mathcal{D}[v, t]$ denote the event that the t^{th} random node we choose dominates node v. The probability of this event can be determined quite easily: of the n nodes in the graph, we must choose v or one of its d neighbors, and so

$$\Pr\left[\mathcal{D}[v,t]\right] = \frac{d+1}{n}.$$

Let $\mathcal{D}_v$ denote the event that the random set consisting of all k selected nodes dominates v. Thus

$$\mathcal{D}_v = \bigcup_{t=1}^{k} \mathcal{D}[v,t].$$

For independent events, we've seen in the text that it's easier to work with intersections—where we can simply multiply out the probabilities—than with unions. So rather than thinking about $\mathcal{D}_v$, we'll consider the complementary "failure event" $\overline{\mathcal{D}_v}$, that no node in the random set dominates v. In order for no node to dominate v, each of our choices has to fail to do so, and hence we have

$$\overline{\mathcal{D}_v} = \bigcap_{t=1}^{k} \overline{\mathcal{D}[v,t]}.$$

Since the events $\overline{\mathcal{D}[v,t]}$ are independent, we can compute the probability on the right-hand side by multiplying all the individual probabilities; thus

$$\Pr\left[\overline{\mathcal{D}_v}\right] = \prod_{t=1}^{k} \Pr\left[\overline{\mathcal{D}[v,t]}\right] = \left(1 - \frac{d+1}{n}\right)^k.$$

Now, $k = \frac{cn \log n}{d+1}$, so we can write this last expression as

$$\left(1 - \frac{d+1}{n}\right)^k = \left[\left(1 - \frac{d+1}{n}\right)^{n/(d+1)}\right]^{c \log n} \le \left(\frac{1}{e}\right)^{c \log n},$$

where the inequality follows from (13.1) that we stated earlier in the chapter.

We have not yet specified the base of the logarithm we use to define k, but it's starting to look like base e is a good choice. Using this, we can further simplify the last expression to

$$\Pr\left[\overline{\mathcal{D}_v}\right] \le \left(\frac{1}{e}\right)^{c \ln n} = \frac{1}{n^c}.$$

We are now very close to done. We have shown that for each node v, the probability that our random set fails to dominate it is at most n^{-c}, which we can drive down to a very small quantity by making c moderately large. Now recall the original event $\mathcal{E}$, that our random set is a dominating set. This fails to occur if and only if one of the events $\mathcal{D}_v$ fails to occur, so $\overline{\mathcal{E}} = \cup_v \overline{\mathcal{D}_v}$. Thus, by the Union Bound (13.2), we have

$$\Pr\left[\overline{\mathcal{E}}\right] \le \sum_{v \in V} \Pr\left[\overline{\mathcal{D}_v}\right] \le n \cdot \frac{1}{n^c} = \frac{1}{n^{c-1}}.$$

Simply choosing $c = 2$ makes this probability $\frac{1}{n}$, which is much less than 1. Thus, with high probability, the event $\mathcal{E}$ holds and our random choice of nodes is indeed a dominating set.

It's interesting to note that the probability of success, as a function of k, exhibits behavior very similar to what we saw in the contention-resolution example in Section 13.1. Setting $k = \Theta(n/d)$ is enough to guarantee that each individual node is dominated with constant probability. This, however, is not enough to get anything useful out of the Union Bound. Then, raising k by another logarithmic factor is enough to drive up the probability of dominating each node to something very close to 1, at which point the Union Bound can come into play.

Solved Exercise 2

Suppose we are given a set of n variables $x_1, x_2, \ldots, x_n$, each of which can take one of the values in the set $\{0, 1\}$. We are also given a set of k equations; the r^{th} equation has the form

$$(x_i + x_j) \bmod 2 = b_r$$

for some choice of two distinct variables x_i, x_j, and for some value b_r that is either 0 or 1. Thus each equation specifies whether the sum of two variables is even or odd.

Consider the problem of finding an assignment of values to variables that maximizes the number of equations that are satisfied (i.e., in which equality actually holds). This problem is NP-hard, though you don't have to prove this.

For example, suppose we are given the equations

$$(x_1 + x_2) \bmod 2 = 0$$

$$(x_1 + x_3) \bmod 2 = 0$$

$$(x_2 + x_4) \bmod 2 = 1$$

$$(x_3 + x_4) \bmod 2 = 0$$

over the four variables $x_1, \ldots, x_4$. Then it's possible to show that no assignment of values to variables will satisfy all equations simultaneously, but setting all variables equal to 0 satisfies three of the four equations.