

**The baseball elimination problem**

We shall now explore a much more sophisticated application of the maximum flow problem. While this application is traditionally phrased in terms of baseball, it really has nothing to do with that sport in particular, especially as it is now played in the major leagues.

The setting is as follows. There are  $n$  teams: team 1, team 2,  $\dots$ , team  $n$ . When two teams play each other in a game, one team wins, and the other loses. (There are no ties, and for simplicity, no rain-outs; each game really takes place, and at the end, one team wins and one team loses, without exception.) There is an agreed-upon schedule that specifies which team should play which other team when; in the entire season, each team is scheduled to play, in total, the same number of games. Each team wants to come in first place; that is, it wants to end the schedule having won the most number of games. If the season ends with two or more teams tied for first place, then all of these teams are considered to have come in first place.

You follow one of these teams closely and are rooting for them to come in first place; we shall assume that we have indexed the teams so that your team is team  $n$ . Unfortunately, it is mid-season and your team is not doing very well at all. You would like to know if it is still possible for them to end the season in first place, or whether, to your great dismay, they have already been eliminated. We shall show that this problem can be formulated as a maximum flow problem.

The input for this *baseball elimination problem* consists of the following. A summary of the remaining games of the season: for each pair of teams  $i$  and  $j$ ,  $i, j = 1, \dots, n$ , the number of games remaining to be played between them, which is denoted  $g(i, j)$ . The standings thus far, which can be summarized by specifying, for each team  $i = 1, \dots, n$ , the number of games that team  $i$  has won thus far, which shall be denoted  $w(i)$ .

We next wish to make precise what it means for our team  $n$  to have been eliminated. We indicate instead what it would mean if team  $n$  were not eliminated. How many games would team  $n$  win if it goes undefeated for the rest of the season (which is clearly the most rosy scenario)? It currently has  $w(n)$  wins, and there are  $\sum_{j=1}^{n-1} g(j, n)$  games remaining on its schedule, and so a perfect completion to its season would give team  $n$  with  $w(n) + \sum_{j=1}^{n-1} g(j, n)$  wins in total. Let this number of wins be denoted  $W$ . Team  $n$  is not eliminated if there is some way that the rest of the games between the other teams can turn out so that each other team has at most  $W$  wins at the end of the season. There are  $g(i, j)$  games remaining between teams  $i$  and  $j$ . Suppose that  $x(i, j)$  denotes the number of games between  $i$  and  $j$ , of those remaining, that team  $i$  wins. Note that  $x(i, j) + x(j, i) = g(i, j) = g(j, i)$  for each  $i, j = 1, \dots, n$ . (Be sure that you understand why this is true; there is little point in reading on before you do! As a further check, what is  $x(i, i)$  and  $g(i, i)$  for each  $i = 1, \dots, n$ ?) Given this notation, we can express the number of wins that each team  $i$  has at the end of the season:  $w(i) + \sum_{j=1}^n x(i, j)$ ; call this value  $W(i)$ ,  $i = 1, \dots, n - 1$ .

Team  $n$  can still come in first place if there exists some way to complete the season such that  $W \geq W(i)$ ,  $i = 1, \dots, n - 1$ . In our notation, that means that team  $n$  is *not eliminated* if there is some assignment of integer values  $x(i, j) \geq 0$ ,  $i, j = 1, \dots, n$  such that

$$x(i, j) + x(j, i) = g(i, j) = g(j, i), \quad \text{for each } i, j = 1, \dots, n - 1, \tag{1}$$

and

$$W = w(n) + \sum_{j=1}^{n-1} g(j, n) \geq w(i) + \sum_{j=1}^{n-1} x(i, j), \quad \text{for each } i = 1, \dots, n - 1. \tag{2}$$

We wish to decide if team  $n$  is not eliminated.

This problem is not as easy as you might think (or as sportscasters might have led you to believe). Consider the following example. There are 4 teams, and in the entire season, each team plays each of the

other 3 teams for 3 games each. That is, each team plays a 9-game season. Suppose that, so far, team 1 has played each of teams 3 and 4 twice, and team 2 has played each of teams 3 and 4 twice. In all cases, the lower indexed team has won. We wish to know if team 4 has is not been eliminated. So in this case, the input consists of the following data: the table of games remaining

Team $i$	Wins $w(i)$
1	4
2	4
3	0
4	0

Games Remaining $g(i, j)$				
$i$ vs. $j$	1	2	3	4
1	–	3	1	1
2	3	–	1	1
3	1	1	–	3
4	1	1	3	–

Team 4 can still end the season with 5 wins, so  $W = 5$ . None of the teams has more than 4 wins thus far, and so the outlook for team 4 does not appear so dim. Or does it? Teams 1 and 2 still have 3 games between them. No matter what happens, one of these two teams must win at least 2 of the 3; but each of these teams already has 4 wins, and so one of these two teams **must** have at least 6 wins at the end of the season. Team 4 can do no better than 5 wins; hence team 4 has been eliminated already.

Another way to think of the question of whether team  $n$  is not eliminated is as follows: can the games between all of the teams  $1, 2, \dots, n - 1$  be played out with each team's win total not exceeding  $W$ , or equivalently, with team  $i$  winning at most  $W - w(i)$  of its remaining games, for each  $i = 1, \dots, n - 1$ ? The key to formulating this question as a maximum flow problem is that instead of oil being shipped through the network, the commodity here is "remaining games between teams  $i$  and  $j$ ,  $i, j = 1, \dots, n - 1$ ". What is more, when a game starts out at the source, its outcome is undecided, but by flowing to the sink (and more specifically, by taking a particular path), a winner has been declared for it.

To construct the input, we introduce a source  $s$ , a sink  $t$ , one node for each pair of teams  $\{i, j\}$ , where  $i, j = 1, \dots, n - 1$  (called the *pair nodes*), and one node for each team  $i$ , where  $i = 1, \dots, n - 1$  (called the *team nodes*). Let the set of pair nodes be denoted  $N_P$  and the set of team nodes be denoted  $N_T$ . There is an arc from the source to each pair node; the capacity of the arc from  $s$  to  $\{i, j\}$  is  $g(i, j)$ . There is an arc leaving each pair node  $\{i, j\}$  to each of the team nodes for  $i$  and  $j$ ; The capacity of each of these is  $G$ , where  $G$  is the total number of games remaining among teams  $\{1, 2, \dots, n - 1\}$ ; that is,

$$G = \sum_{i=1}^{n-1} \sum_{j=1}^{i-1} g(i, j) = \sum_{\{i, j\} \in N_P} g(i, j).$$

(An explanation of notation is probably useful here: if you have  $\sum_{i=k}^l y(i)$  when  $k > l$ , then this sum is 0 no matter what the values  $y(\cdot)$  are; if  $k = l$ , then the sum is  $y(k) = y(l)$ .) For each team node  $i$ ,  $i = 1, \dots, n - 1$ , there is an arc of capacity  $W - w(i)$  leaving this node to the sink  $t$ . Of course, if  $W < w(i)$ , then team  $n$  is clearly eliminated, and hence we don't need this maximum flow formulation at all; hence, we shall assume that  $W \geq w(i)$ . This completes the construction of the input. An example of this construction applied to the specific input given above is shown in figure 1.

We claim that team  $n$  is not eliminated precisely when the maximum flow value for this input is equal to  $G$ , the total number of games remaining to be played between teams in  $\{1, 2, \dots, n - 1\}$ . Why should this claim be true? To show it, we must show two things: (A) when team  $n$  is not eliminated, then this input does have such a flow; (B) when the input has maximum flow value  $G$ , then team  $n$  is not eliminated.

Before doing this more formally, let us first give some intuition behind the construction. Look at one unit of flow that goes from the source to the sink. We think of each unit of flow as corresponding to one of the remaining games between teams in  $\{1, 2, \dots, n - 1\}$ . First it passes through one of the pair nodes  $\{i, j\}$ . This classifies the game as being between teams  $i$  and  $j$ . This game will be won by either team  $i$  or team  $j$ . The unit of flow then passes through one of the two team nodes  $i$  and  $j$ . This classifies this game as being won by that team. It then flows from that team node to the sink. The capacity of the arc from the source to the pair node  $\{i, j\}$  is  $g(i, j)$  and this ensures that no more than  $g(i, j)$  games are classified as being played between teams  $i$  and  $j$ . The capacity from team node  $i$  to the sink is  $W - w(i)$ . This ensures that for any feasible flow, at most that many of the remaining games are won by team  $i$ . Next observe that the cut defined by  $s$  as the source set, and all of the remaining nodes in the sink set is a cut of capacity

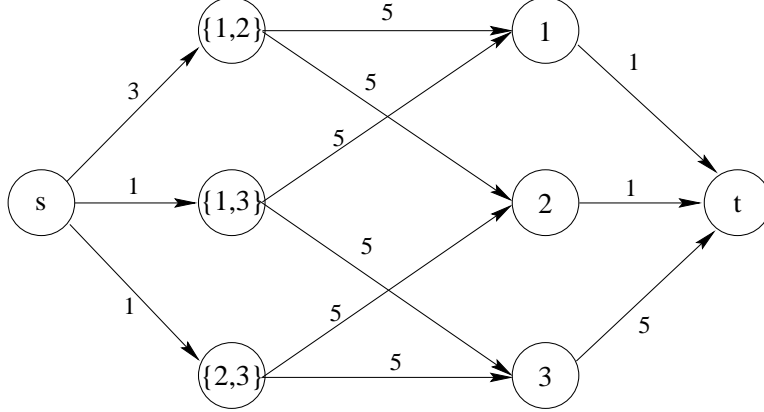


Figure 1: An example of a baseball elimination formulation

$G$ . Hence, if there is a feasible flow  $f$  of value  $G$ , then each arc from  $s$  to  $\{i, j\}$  must have  $g(i, j)$  units of flow in  $f$ . So this flow is calling for all remaining games to be played. And for each of these games, one of the teams playing is a winner. But the capacity restrictions of the arcs going into the sink ensure that each team  $i$  wins at most  $W - w(i)$  of these games. But this means that team  $n$  is not eliminated.

Now we give this argument more formally. We first argue that (A) is true. By the definition of being not eliminated, there exist values  $x(i, j) \geq 0$ ,  $i, j = 1, \dots, n$  satisfying constraints (1) and (2). We assign flow value  $g(i, j)$  to the arc from  $s$  to  $\{i, j\}$ , for each pair node  $\{i, j\}$ ; this flow value clearly satisfies the capacity constraint on this arc. For each pair node  $\{i, j\}$ , we assign flow value  $x(i, j)$  to the arc from  $\{i, j\}$  to  $i$ , and flow value  $x(j, i)$  to the arc from  $\{i, j\}$  to  $j$ . Since  $g(i, j) = x(i, j) + x(j, i)$  by (1), the node flow conservation constraint is satisfied at each pair node  $\{i, j\}$ . The total flow assigned to enter team node  $i$  is  $\sum_{j=1}^{n-1} x(i, j)$ ; we let this sum be the flow value on the arc from team node  $i$  to the sink, and hence the node flow conservation constraint is satisfied at each team node  $i$ . To see that the capacity constraint of the arc from team node  $i$  to the sink is satisfied, observe that constraint (2) implies that

$$W - w(i) \geq \sum_{j=1}^{n-1} x(i, j), \quad \text{for each } i = 1, \dots, n-1,$$

which is exactly what we wished to show. (Be sure you understand this!) We have just given a flow of value  $G$ , and since there is a cut of capacity  $G$ , we know that  $G$  is the maximum flow value.

Now we will argue that (B) is true. Suppose that there is a flow  $f$  of value  $G$  for our maximum flow input. Since all of the capacities are integers, we can assume that the flow value assigned to each arc is an integer. As we argued above, the cut in which the source side consists only of the source itself has capacity  $G$ , and so the flow value  $f(s, \{i, j\})$  is equal to  $g(i, j)$ , for each pair node  $\{i, j\}$ . Set  $\bar{x}(i, j)$  equal to the flow value  $f(\{i, j\}, i)$  and  $\bar{x}(j, i) = f(\{i, j\}, j)$ . Since  $g(i, j)$  units of the flow  $f$  go into each pair node  $\{i, j\}$ , and there are only two arcs coming out this node,

$$g(i, j) = f(s, \{i, j\}) = f(\{i, j\}, i) + f(\{i, j\}, j) = \bar{x}(i, j) + \bar{x}(j, i).$$

In other words,  $\bar{x}$  satisfies (1). Next observe that  $\sum_{j=1}^{n-1} \bar{x}(i, j)$  units of flow  $f$  go into team node  $i$ . There is only one arc leaving team node  $i$ , going to the sink, and so  $f(i, t) = \sum_{j=1}^{n-1} \bar{x}(i, j)$ , for each  $i = 1, \dots, n-1$ . But  $f$  satisfies all capacity constraints, and so  $f(i, t) \leq W - w(i)$ ,  $i = 1, \dots, n-1$ . By combining the last two relations, we get that  $\sum_{j=1}^{n-1} \bar{x}(i, j) \leq W - w(i)$ ,  $i = 1, \dots, n-1$ . This final inequality is equivalent to writing that  $\bar{x}$  satisfies (2). By the integrality property, the flow  $f$  has been assumed to take only integer values. Hence  $\bar{x}(i, j)$  are all non-negative integer values. We have exhibited the required values to prove that team  $n$  is not eliminated. This completes the argument that our formulation correctly determines whether team  $n$  has been eliminated.

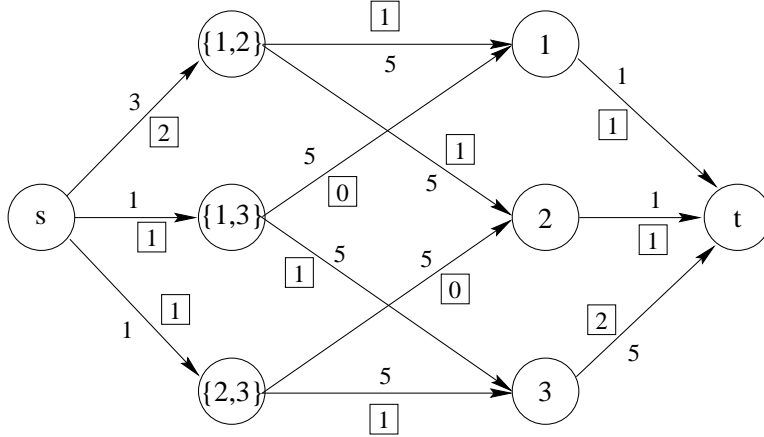


Figure 2: An optimal flow for the input given above

Suppose that you have computed the maximum flow for the input in figure 1. In fact, the maximum flow is given in Figure 2, and its value is 4, which is less than the total number of games remaining among teams 1,2, and 3, which is equal to 5. Hence, team 4 has been eliminated. But we knew this already; we have already given a quite simple argument that shows that team 4 cannot end the season in first place. But suppose that you use this procedure on some other data, and once again, this procedure indicates that your team has been eliminated. You want to convince a friend that this is the case. How can you do this? One option is to teach the friend everything you have learned in this class about the maximum flow problem, but that is probably more than the friend is willing to listen to. There is another alternative: for the input constructed by our formulation, the minimum cut can always be interpreted to give a short proof of the fact that your team has been eliminated (assuming that the max flow value is less than  $G$ , of course).

Suppose that team  $n$  is not eliminated, and look at a subset  $R$  of the other teams, i.e.,  $R \subseteq \{1, 2, \dots, n-1\}$ . There are  $\sum_{\{i,j\} \subseteq R} g(i, j)$  games remaining to be played between teams (both) from the set  $R$ . For each of these games, some team in  $R$  wins the game. So the teams in  $R$ , in total, must win at least  $\sum_{\{i,j\} \subseteq R} g(i, j)$  more games. However, since team  $n$  is not eliminated, each team  $i \in R$  can win at most  $W - w(i)$  of its remaining games. In total, the teams in  $R$  can win at most  $\sum_{i \in R} (W - w(i))$  more games. Hence, it must be the case, that

$$\sum_{\{i,j\} \subseteq R} g(i, j) \leq \sum_{i \in R} (W - w(i)).$$

However, suppose you checked the data that you are given, and this is not the case; that is, you have found a subset of teams  $R$  such that

$$\sum_{\{i,j\} \subseteq R} g(i, j) > \sum_{i \in R} (W - w(i)). \quad (3)$$

What conclusions can you draw from this? You must conclude that team  $n$  has, in fact, been eliminated. In fact, we have already used this argument. We just took  $R = \{1, 2\}$  to convince ourselves that team 4 has been eliminated; there are 3 games remaining between these two teams, and each of these two teams can only win 1 more game without eliminating team 4.

But this is an argument you can explain to your friend without teaching everything we have learned about the maximum flow problem! The question remains, how can we find the set  $R$ , and even worse, might team  $n$  be eliminated (in some cases) without there being such a simple proof that it has been eliminated. The answer is that there always is such a simple proof, and that we can find the set  $R$  from the minimum cut.

Suppose that the maximum flow value is less than  $G$ ; hence, there is a cut  $(S, T)$  of capacity less than  $G$ . The source side of this cut,  $S$ , must contain the source  $s$ , along with a subset  $P$  of the pair nodes, and a subset  $R$  of the team nodes; that is,  $S = \{s\} \cup P \cup R$ . Let  $\{i, j\}$  be some node in  $P$ . We know then that team node  $i$  must be in  $R$ , since otherwise, the arc  $(\{i, j\}, i)$  of capacity  $G$  would cross the cut, which contradicts

the fact that the capacity of the entire cut is less than  $G$ . Similarly,  $\{i, j\} \in P$  also implies that team node  $j$  is in the set  $R$ . This implies that

$$\sum_{\{i,j\} \in P} g(i, j) \leq \sum_{\{i,j\} \subseteq R} g(i, j). \quad (4)$$

We can also conclude that the capacity of this cut is the sum of the capacity of some arcs leaving the source, plus the capacity of some of the arcs entering the sink. More precisely, the capacity of  $(S, T)$  is

$$\sum_{\{i,j\} \in N_P - P} g(i, j) + \sum_{i \in R} (W - w(i)) < G,$$

where the inequality simply reflects that we have assumed that the min cut capacity is less than  $G$ . This inequality implies that

$$\sum_{i \in R} (W - w(i)) < G - \sum_{\{i,j\} \in N_P - P} g(i, j) = \sum_{\{i,j\} \in N_P} g(i, j) - \sum_{\{i,j\} \in N_P - P} g(i, j) = \sum_{\{i,j\} \in P} g(i, j) \leq \sum_{\{i,j\} \subseteq R} g(i, j),$$

where the last inequality follows from (4). But this is just the inequality (3): the set  $R$  can be used to explain to your friend that team  $n$  has been eliminated. (Double check that the minimum cut for the example in Figure 1 does generate  $R = \{1, 2\}$ !)