Math 410 (Prof. Bayly) SPORTS RANKINGS‡ Summer 2005

Sports fascinates people. Besides the pleasure we take in exerting our bodies and spending time with our friends in good-natured competition, we love to watch others compete. Even if we don't know the athletes personally, we develop favorites whose fortunes we follow, and it means a lot to us whether our team wins or loses. Interest in sports is a global phenomenon, and there is a great deal of money in the sports industry, especially in professional team sports such as soccer, basketball, football, and baseball.

A great deal of the interest in a sport centers on the questions "Who is the best?" and "Who is the better of two teams or individuals?" In Olympic-style sports like track-and-field or swimming, there are objective measures of how fast each athlete can swim 100 meters or how far they can throw a standard javelin. In sports like tennis or soccer, however, players or teams compete one-to-one in pairs, and there is no direct information about who is "better" between two players or teams who have not yet played each other.

But that does not stop the speculators! The college and professional football leagues are ranked by the sports magazines every week of the season, the entrants to major tennis tournaments are "seeded", and bookmakers* and gamblers continually strive to make better predictions than each other about the outcomes of future competitions.

Since ranking and assessing are quantitative exercises, it's natural to think that mathematics can help us. In practice the information that we get is limited by the fact that human beings are complicated, and what actually determines who wins a game depends on a lot of things that we can't observe. Nevertheless if we imagine that athletes are simple quantifiable entities, we can devise ranking systems that make sense for our imaginary athletes. With appropriate caution we can input the data for real athletes, and draw tentative speculations about their future performances.

Model. Let's imagine a typical college sport in which a team visits another team's campus and they play a game. Each team scores points, and the team with the higher number of points is the winner. This basic structure is common to baseball, basketball, football, and soccer; we're obviously leaving out a lot of important aspects of specific sports. We assume that the better team wins, and that the score difference indicates how much better they are. This assumption clearly leaves out a lot of factors, some of which are unpredictable (like whether the quarterback stayed up late studying for a Cosmology midterm) and some are predictable (like home court advantage) but difficult to quantify, especially in our first model.

Now we make a big assumption: that each team has a certain intrinsic quality level q which

‡ This material is adapted from the discussion in Gil Strang's text "Introduction to Linear Algebra and its Applications".

* It has been said that "a bookmaker is a pickpocket who lets you use your own hands."

reflects how good it is, in the sense that we expect the (score of team A) minus (score of team B) to equal $q_A - q_B$ when they play each other. Our goal, of possible, is to use the scores of the games that have been played so far to calculate the q values for all the teams, and therefore be able to make predictions about the outcomes of future games.

We shall use a digraph to represent the total history of games played so far in a league. The nodes will represent teams (usually denoted $A, B, C, ...$), and the edges (denoted $a, b, c, ...$) will represent games. The arrow will indicate which team is home (the head of the arrow) and which is the visitor (the tail of the arrow); we visualize the arrow indicating the trip made by the visiting team. Associated with each edge will be the algebraic difference (home score) - (visitor score) for that game.

Example. Four teams A, B, C, D play two games a, b.

In game a team B visits team A; B wins by 4 to 3.

In game b team C visits team D; D wins by 2 to 1.

The equations for the team qualities are

$$q_A - q_B = 3 - 4 = -1 \quad , \quad q_D - q_C = 2 - 1 = 1, \tag{1a}$$

or $G\vec{q} = \vec{d}$, where

$$G = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \quad , \quad \vec{q} == \begin{pmatrix} q_A \\ q_B \\ q_C \\ q_D \end{pmatrix} \quad , \quad \vec{d} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \tag{1b}$$

Note that G is the edge-node matrix of the directed graph, and $\vec{d}$ is the vector of algebraic score differences, always calculated as (home) - (visitor).

The system (1) has an infinite number of solutions with two free variables q_B, q_D:

$$\vec{q} = \begin{pmatrix} q_B - 1 \\ q_B \\ q_D - 1 \\ q_D \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ -1 \\ 0 \end{pmatrix} + q_B \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + q_D \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}; \tag{2}$$

note that the two null vectors correspond to the two separate pieces of the digraph. Team A is ranked 1 unit lower than team B, and team C is one unit lower than D. But because q_B, q_D are independent, we cannot compare A or B with C or D. It's not at all surprising in context: the group A,B could be major-league professionals and C,D could be a couple of neighborhood pickup teams, and without any inter-group games we have no basis for comparisons between members of different groups.

To rank all the teams, we need a game between one of A,B and one of C,D. Suppose in game c team C visits team B and wins by 5 to 2, which adds the equation $q_B - q_C = 2 - 5 = -3$ to the system. The solution still has an infinite number of solutions (with free variable q_D):

$$\vec{q} = \begin{pmatrix} q_D - 5 \\ q_D - 4 \\ q_D - 1 \\ q_D \end{pmatrix} = \begin{pmatrix} -5 \\ -4 \\ -1 \\ 0 \end{pmatrix} + q_D \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \tag{2}$$

The relative rankings, however, are unaffected by q_D; only the particular solution is important.

Note that the solution predicts that if A were to play D, then D would beat A by 5 points. The likely point spread between teams that have not yet played each other is popular ground for wagering, and this kind of information can be quite useful. (We are NOT responsible for you losing money on bets made with this system, though if you happen to hit the jackpot we would not decline a portion of your winnings as a token of your gratitude.)

Back on the field, the season continues. In game d, A visits C and wins by 9 to 6, adding the equation $q_A - q_C = 3$ to the system. Unfortunately this equation is inconsistent with the previous equations $q_B - q_C = -3$ and $q_A - q_B = -1$. It's not that the actual game outcomes are inconsistent; it frequently happens that B beats A, C beats B, and later A beats C. The error is the assumption that the score differences exactly correspond with underlying quality differences. Mathematically, the problem is that the system $G\vec{q} = \vec{d}$, where now

$$G = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 \end{pmatrix} \quad , \quad \vec{d} = \begin{pmatrix} -1 \\ 1 \\ -3 \\ 3 \end{pmatrix}, \tag{3}$$

has no solutions.

This would have been a problem in chapter 1, but now that we have completed chapter 4 we know not to give up when a system of equations has no solution. The next best thing (from one point of view) is simply to seek the least squares approximate solution, i.e. the solution of the normal equations $G^T G\vec{q} = G^T\vec{d}$. Explicitly

$$G^T G = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ 0 & 0 & 01 & 1 \end{pmatrix} \quad , \quad G^T\vec{d} = \begin{pmatrix} 2 \\ -2 \\ -1 \\ 1 \end{pmatrix} \tag{4a}$$

and the normal equations have the infinite family of solutions

$$\vec{q} = \begin{pmatrix} q_D - 5/6 \\ q_D - 5/3 \\ q_D - 1 \\ q_d \end{pmatrix} = \begin{pmatrix} -5/6 \\ -5/3 \\ -1 \\ 0 \end{pmatrix} + q_d \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \tag{4b}$$

So we conclude that D is the highest ranked team, followed by A, then C, and B lowest; the numerical values indicating our best estimates of likely point spreads. As before, the free variable q_D does not affect the rankings.

To summarize, the method I suggest for ranking athletes or teams uses information about who has beaten whom so far, and by how much. The designation into home and visitors is convenient, but not essential - it just gives an unambiguous way to put directions on the edges of the digraph representing the games, and the corresponding ± 1's in the edge-node incidence matrix G. (We can only expect a globally-defined ranking if the digraph is connected.) The linear system for the "quality" measures is $G\vec{q} = \vec{d}$, where $\vec{d}$ is the vector of (home)-(visitor) score differences. This system may have no solutions, in which case we try to find a least squares approximate solution by solving the normal equations $G^T G = G^T \vec{d}$.

This method may seem extremely simplistic, and it is; we're not claiming any particular merit for it besides its simplicity! But that's the whole idea of mathematical modeling - to invent simplified versions of complicated systems, and analyze the simplified systems. In the following critique we discuss this further, but one variation that is easy to incorporate into this scheme is home court advantage. If you believe that playing at home lets a team score h more points than when they play away, we just add h to every entry in $\vec{d}$. If you aren't sure what h should be, you can treat it as an additional variable to be solved for or estimated as part of the solution process.

Critique. The biggest problem with this strategy for ranking teams is that the model is just not realistic. How likely one athlete is to beat another athlete cannot be reduced to one number representing overall "quality", and teams have even more dimensions of variation than individuals do. A very modest way to address this shortcoming would be to represent a team by two "quality" variables, one representing offense and the other defense. The actual number of points scored by one team represents the difference between its offense and its opponent's defense. A more sophisticated approach would be to consider more aspects of a team's potential, and perhaps keep track of more game statistics, such as the number of turnovers or shots taken by offense or blocked by defense.

A further complication is that there is a lot more than points to the story. For one thing, the actual numerical margin by which a team wins or loses means nothing as far as the sort is concerned - only the victory or loss is recorded. It's well known that people (including athletes) relax when the situation seems safe, and usually don't play as hard when they have a comfortable lead. Coaches often intentionally give their second-string players more play time when it seems that victory is assured; in fact this is a good strategy for having a more successful season or multiyear program. Conversely, a team who can really pour it on when they are behind, or ahead by a small number of points, might also be a much better team (in terms of its win-loss record) than its point scores indicate.

One way to accommodate this effect (as Strang writes) is to modify the entries in $\vec{d}$ (as we did with home court advantage) to give a bonus for the simple achievement of victory. Indeed, in my children's community soccer league tournaments, the number of "points" awarded a team is equal to the number of goals scored, plus one if the team won its game, and another point is given if the opposing team scored no goals. The latter modification recognizes that a shutout reflects substantial dominance by the winning team (consider a 3-0 versus a 4-1 victory).

The bad news is that if you're serious about using mathematics to predict the outcomes of sporting events, there are a great many factors to be considered, some of which do not lend themselves to quantification. The good news is that the field is completely open! Select a sport and follow it for a season, get some data and see how well the simple method works, and perhaps modify for home court advantage or victory bonus. Try a few crazy ideas of your own. If you find something that works better than my method, let me know - or at least, please cut me a slice of your winnings!