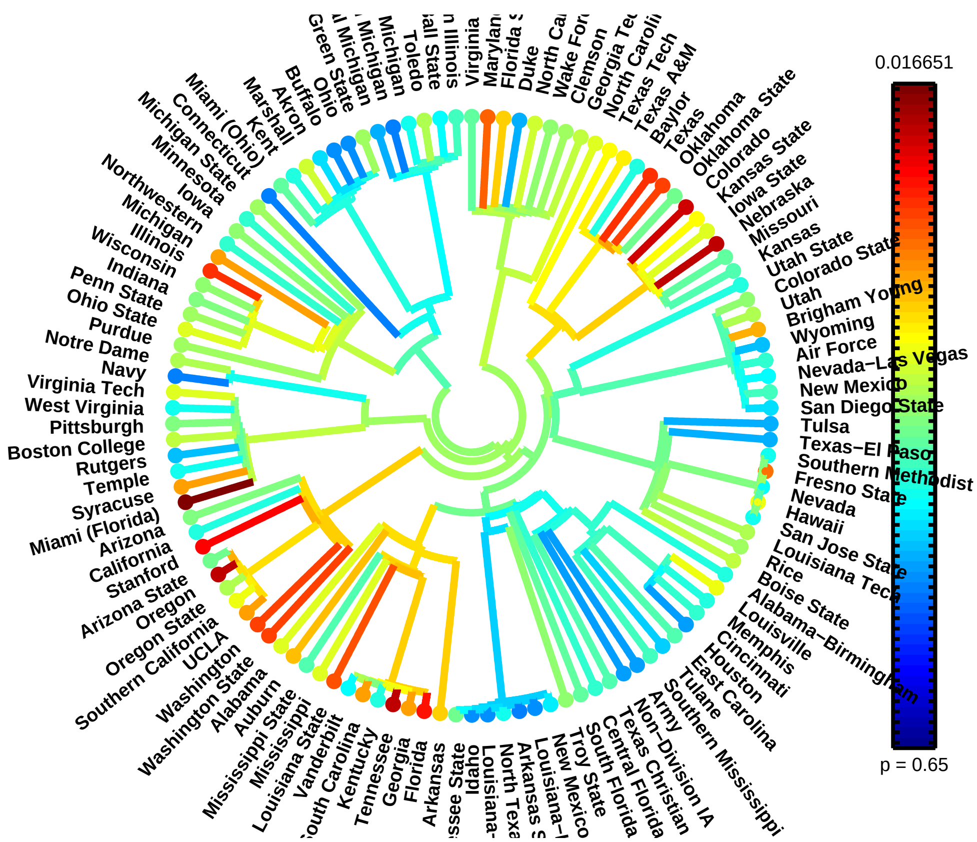
Random Walker Ranking for NCAA Division I-A Football

Thomas Callaghan, Mason A. Porter, and Peter J. Mucha
Georgia Institute of Technology



The above figure represents the expected distribution of model random-walker votes cast for each NCAA Division I-A football team in 2001 pre-bowl rankings. The organization of the teams and the lines connecting them represent the community structure hierarchy, of which the conferences are one level of organization. The colors represent the expected percentage of votes cast per team at each level in the hierarchy, from individual teams up through intra-conference organization, the conferences, and the connections between conferences. The biased probability for voting for the winners of games in the data represented in this figure is $p = 0.65$. *Details are in the enclosed article.*

Additional commentary with current rankings is online
(http://www.math.gatech.edu/~mucha/BCS).

Random Walker Ranking for NCAA Division I-A Football

Thomas Callaghan*, Mason A. Porter*†, and Peter J. Mucha*‡
*School of Mathematics, Georgia Institute of Technology
Atlanta, GA 30332-0160; and
†Center for Nonlinear Science, School of Physics
Georgia Institute of Technology, Atlanta, GA 30332-0430.

November 11, 2004

We develop a one-parameter family of ranking systems for NCAA Division I-A football teams based on a collection of voters, each with a single vote, executing independent random walks on a network defined by the teams (vertices) and the games played (edges). The virtue of this class of ranking systems lies in the simplicity of its explanation. We discuss the statistical properties of the randomly walking voters and relate them to the community structure of the underlying network. We compare the results of these rankings for recent seasons with Bowl Championship Series standings. To better understand this ranking system, we also examine the asymptotic behaviors of the aggregate of walkers. Finally, we consider possible generalizations to this ranking algorithm.

‡Author to whom correspondence should be addressed, mucha@math.gatech.edu, 404/894-9235, 404/894-4409(fax); additional commentary with current rankings is available online (http://www.math.gatech.edu/∼mucha/BCS).

1 Introduction

The Bowl Championship Series (BCS) agreement was created in 1998 in an effort to match the top two NCAA Division I-A teams in an end-of-season National Championship game. The BCS system takes the champions from the six major conferences—the Pac-10, Big 12, Big Ten, ACC, SEC and Big East, plus Notre Dame—and two at-large teams to play in four BCS bowls, with the National Championship game rotating among those bowls on a yearly basis. The ranking system used to determine which teams play in the championship games has significant financial ramifications, with the Rose, Fiesta, Sugar and Orange bowls generating more than $100 million a year for the six BCS conferences. There is also a general belief that the schools playing in the National Championship game reap immediate increases in financial contributions and student applications.

Attempts to determine the "top two" teams also have enormous personal ramifications for college football fans nationwide. The BCS attempts to rank teams fairly in order to determine which two most deserve to play in the National Championship game and which teams should play in the other major bowl games. Before the BCS was instituted, teams were selected for major bowl games according to tradition. For example, the Rose Bowl had traditionally featured the conference champions from the Big 10 and Pac-10. Consequently a match between the #1 and #2 teams in the nation rarely occurred in bowl games. This often left multiple undefeated teams and many co-National Champions. On other occasions, a single team with an arguably easy schedule might be undefeated and therefore be declared National Champion by polls without ever having played a "major" opponent (e.g., BYU in 1984).

The BCS system attempts to address these problems and simultaneously maintain the tradition of finishing the season with bowl games. It consists of a combination of four equal factors: two polls (coaches and sportswriters), algorithmic rankings, strength of schedule, and losses. Bonus points are also awarded for defeating highly-ranked teams. Seven sources currently provide the participating ranking algorithms, selected by the BCS from over 50 applicants. The best and worst computer rankings are removed and the remaining five are averaged.

The fundamental difficulty in accurately ranking or even agreeing on a ranking algorithm for the NCAA Division I-A college football teams lies in two factors—the dearth of games played by each team and the large disparities in the type and difficulty of the schedules of individual teams. With 117 Division I-A football teams, the 10–13 regular season games (including conference tournaments) played by each team severely limits the quantity of information relative to, for example, college and professional basketball and baseball schedules. Even the 32 teams in the professional National Football League (NFL) each play 16 regular season games against 13 distinct opponents, including multiple games against divisional opponents. The NFL then uses the regular season outcomes to seed a 12-team playoff. Moreover, most of the Division I-A football teams play the majority of their games within their conferences, with significant variations in the level of play from one conference to another further complicating attempts to select the top two teams from the available information. To make matters worse, it is not even clear what the phrase "top two" means: Should these teams be the two with the best overall seasons or the two playing the best at the end of the season?

Despite the obvious difficulties in determining a "best" algorithm, many systems for ranking college football teams have been promoted by mathematically and statistically inclined fans (see, for example, [1]). Many of these schemes are relatively complicated mathematically, making it virtually impossible for the lay sports enthusiast to understand the ranking methodology and the assumptions made therein. Worse still, the essential formulas of many of the algorithms currently employed by the BCS are not even publicly declared. This state of affairs has inspired the creation of software to develop one's own rankings using a collection of polls and algorithms [2] and comical commentary on "faking" one's own mathematical algorithm [3].

Because of the essential difficulties and controversies involved, we sought to investigate whether a simply-explained algorithm constructed by crudely mimicking the behavior of a collection of voters could provide reasonable rankings. We define a collection of voting automatons (random walkers) who can each cast a single vote for the team they believe is the best. Because the most natural arguments relating the relative rankings of two teams come from the outcome of head-to-head competition, each of these voters routinely examines the outcome of a single game played by their favorite team, selected at random from that team's schedule, and makes a new determination of which team to vote for based on the outcome of that game, preferring but not absolutely certain to go with the winner. In the simplest definition of this process, the probability p of choosing the winner is defined to be the same for all voters and games played, with $p > 1/2$ because on average the winner should be the better team and $p < 1$ to allow one of these artificial voters to argue that the losing team is still the better team. (Additionally, the $p = 1$ limit can be mathematically more complicated in certain scenarios, as addressed in section 4.2). That is, if the voter's preferred team won the examined game, the walker's vote remains with that team with probability p; however, if the current favorite team lost the game, then with probability p the walker changes its vote to the winning opponent. Conversely, a given walker will vote for the losing team with probability $0 < (1 - p) < 1/2$, corresponding to any reason one might rationalize that the losing team is still the better team.

The behavior of each synthetic voter is driven by a simplified version of the "but my team beat your team" argument that one commonly hears about why one team should be ranked higher than another. For instance, much of the 2001 BCS controversy centered on the fact that BCS #2 Nebraska had lost to BCS #3 Colorado, and the 2000 BCS controversy was driven by BCS #3 Miami's win over BCS #2 Florida State and BCS #4 Washington's win over Miami. While developing a reasonable ranking is an incredibly difficult task because of the relatively few games between many teams, most fans would agree that Team A winning a game against Team B more often means that Team A should be the better team (thus $p > 1/2$). Despite that, various arguments are always made about why a losing team may still have been the better team, for instance due to weather, officiating, injuries, or sheer luck (thus $p < 1$).

The voting automatons thus act as independent random walkers on a graph ("network") defined with biased edges ("connections") between teams that played games head-to-head. This algorithm is easy to define in terms of the "microscopic" behavior of individual-voting walkers who randomly change their votes based on the win-loss outcomes of individual games. The random behavior of these individual voters is, of course, grossly simplistic. Indeed, under the specified range of p, a given voter will never reach a certain conclusion about which team is the best; rather, it will forever change its allegiance from one team to another, ultimately traversing the entire graph and thus eventually casting a vote for every single team. In practice, however, the "macroscopic" total number of votes cast for each team by an aggregate of random-walking voters quickly reaches a statistically-steady ranking of the top teams according to the quality of their seasons.

We do not claim that this random-walker ranking algorithm is superior to other algorithms, nor do we review the vast number of ranking systems available, as numerous such reviews are available (see, for example, [1] for a large listing of various schemes, [4, 5, 6, 7] for reviews of different ranking methodologies, and the "Bibliography on College Football Ranking Systems" maintained at [8]). Rather, our intention is to determine whether such a simply-defined random-walker ranking yields reasonable results. We do not even claim that this ranking algorithm is wholly novel; indeed, the resulting linear algebra problem is related to the "direct methods" discussed and referenced by Keener [4].

Nevertheless, we propose this model on the strength of its simple interpretation of random-walking voters as a reasonable way to rank the top teams (or at least as reasonable as other available methods, given the scarcity of games played relative to the number of teams). In contrast, the definitions of many ranking algorithms are more mathematically intensive, and some systems include a miasma of seemingly arbitrary parameters whose effects are difficult to understand and interpret. Some systems are even tweaked periodically in an attempt to yield the purportedly "most reasonable" ranking based on recent results. The advantage of the algorithm discussed here and its random-walking voter interpretation is that it has only one explicit, precisely-defined parameter with a meaningful interpretation that is easily understood at the level of single-voter behavior.

This article is organized as follows: The mathematical details of the ranking algorithm and statistical properties of the independent random walkers are presented in section 2, and we briefly compare recent historical outcomes from this algorithm with other ranking systems and polls. In section 3, we discuss some of the network properties of the graph defined by the games played, independent of the outcomes of those games, with special attention to community structure. In section 4, we investigate the asymptotic behavior of the random walkers for extreme values of the probability p. Finally, in section 5, we discuss possible generalizations of this ranking methodology.

2 Ranking with Random Walkers

Given the outcomes of the games played, we denote for each i the number of games played n_i, wins w_i, and losses l_i by team i. A tie (possible prior to the current NCAA overtime format) is treated as half a win and half a loss, so that $n_i = w_i + l_i$ holds. The number of random walkers casting their single vote for team i is v_i, with the condition that the total number of voters remains constant, $\sum_i v_i = Q$.

If team i beat team j, then the average rate at which a walker voting for team j changes allegiances to vote for team i is proportional to p, while the rate at which a walker already voting for team i switches to team j is proportional to $(1-p)$. We additionally need to define the rate at which individual walkers currently voting for team i consider the different n_i games played by that team. For simplicity, we assume that each of the n_i games is selected with equal weight. To avoid rewarding teams for the raw number of games played, we set the total rate at which games are considered by voters at the ith vertex to be proportional to the number n_i of edges coming out of that vertex.

The expected rate of change of the number of votes cast for each team in this random walk is thus quantified by a homogeneous system of linear differential equations,

$$\bar{\mathbf{v}}' = \mathbf{D} \cdot \bar{\mathbf{v}}\,, \tag{1}$$

where $\bar{\mathbf{v}}$ is the T-vector of the expected number $\bar{v}_i$ of votes cast for each of the T teams, and $\mathbf{D}$ is the square matrix given by

$$\begin{aligned} D_{ii} &= -pl_i - (1-p)w_i\,, \\ D_{ij} &= \frac{1}{2}N_{ij} + \frac{(2p-1)}{2}A_{ij}\,, \quad i \neq j\,, \end{aligned} \tag{2}$$

where N_{ij} is the number of head-to-head games played between teams i and j and A_{ij} is the number of wins of team i over team j minus the number of losses in those N_{ij} games. That is, if i and j played at most one head-to-head game, then

$$\begin{aligned} A_{ij} &= +1\,, \quad \text{if team } i \text{ beat team } j\,, \\ A_{ij} &= -1\,, \quad \text{if team } i \text{ lost to team } j\,, \\ A_{ij} &= 0\,, \quad \text{if team } i \text{ tied or did not play team } j\,. \end{aligned} \tag{3}$$

In the event that two teams play each other multiple times (which occasionally occurs because of conference-championship tournaments), we sum A_{ij} over each game. We again remark that the probability p of going with the winner is constrained to lie in the interval $(1/2\,,1)$, so that the winning team is rewarded for winning but that some uncertainty in voter behavior is maintained. The off-diagonal elements D_{ij} are therefore necessarily nonnegative and vanish if and only if teams i and j did not play directly against one another.

The matrix $\mathbf{D}$ encompasses all of the connections and win-loss outcomes between teams. The steady-state equilibrium

$$\mathbf{D} \cdot \bar{\mathbf{v}}^* = \mathbf{0} \tag{4}$$

of (1,2) gives the expected populations $\bar{\mathbf{v}}^*$ of the random walkers for each team. The critical point $\bar{\mathbf{v}}^*$ lies in the null-space of $\mathbf{D}$—that is, it is the eigenvector associated with a zero eigenvalue. It is important to realize that the appropriate equations to solve are *not* those indicating that each individual connection has zero flow at equilibrium. To illustrate this, consider a schedule of three teams in which each plays only two games such that i beats j, j beats k, and k beats j. In this cyclic situation, the votes flow on average around the triangle (no detailed balance), but a statistical equilibrium occurs when each team receives the same number of votes.

Recall that the ranking system presented here examines the NCAA Division I-A football teams. As many of these teams play Non-Division I-A teams, we have added an additional vertex to the graph representing all the Non-Division-I-A teams that play against Division I-A opponents. Because these teams usually do not fare well against Division I-A competition, this new "team" ends up ranked very low, and does not significantly affect the random walker populations except to penalize losses against Non-Division-I-A teams and to maintain the constraint that the total number of votes Q remains constant (i.e., votes do not leave the graph). Alternatively, one could elect to ignore games against Non-Division-I-A opponents, but one may prefer to recognize losses against such opponents as serious demerits in ranking teams. The ideal situation would be to study the connected component of all football games that includes all of Division I-A as a subgraph; but we do not consider this here because it was easier to obtain data (from [9]) and investigate the algorithm on the smaller graph consisting only of Division I-A football teams and a single node collectively representing the Non-Division I-A teams.

Given the schedule and win-loss outcomes, the long-time statistical behavior of these randomly walking voters is determined by the single parameter p, the probability of a random walker choosing the winner of a given game. These votes can be used directly to rank the teams. Despite the simplistic behavior of an individual random walker, the behavior of an aggregate of voters (or equivalently, because of the independence assumption, the long-time average of a single voter) seems to yield reasonably robust orderings of the top teams. It is not surprising that the detailed results describing the number of votes cast for a given team vary substantially for different values of p; nevertheless, the relative ranking of the top few teams does not vary much for different p, as indicated for the 2001 and 2002 seasons in figure 1. While we investigated college football rankings going back to 1970; we focus much of our discussion here on 2001 and 2002, since these seasons represent two extreme situations in attempting to rank college football teams.

The only certainty in the 2001 pre-bowl game rankings [10] was that Miami belonged in the National Championship game as the only undefeated team in Division I-A football. Indeed, both polls and all eight algorithms used by the BCS picked Miami as the #1 team going into the bowl games at the end of the 2001 season. The controversy commonly presented outside the Pacific Northwest in December 2001 concerned Nebraska's selection as the #2 BCS team, narrowly surpassing BCS #3 Colorado despite the latter's late-season rout of Nebraska, with smaller mention of the fact that BCS #4 Oregon had been ranked #2 in both polls. After the bowl games, in which Miami defeated Nebraska and Oregon defeated Colorado, it was Oregon's absence from the Championship game that became the centerpiece of national controversy. The random walkers select Oregon #2 for $p > 1/2$ up to $p \approx .62$, above which Nebraska takes #2 in a narrow range up to $p \approx .68$. Above that value, the random walkers select Tennessee as #2. The entire controversy that season would have been avoided had Tennessee not lost the SEC Championship game to LSU at the end of the season. Without any intention of being deliberately nonconformist [3] or regionally biased, the random walking voters (figure 1a) choose Tennessee as #2 over the widest range of probabilities (in contrast, Tennessee fell to #5 in the BCS and #8 in both polls following their loss to LSU). This ranking is partially explained by the fact that the simplest random-walking voter algorithm presented here does not distinguish games based on the date played. It is also influenced by the fact that the SEC as a whole is highly ranked that season; indeed, in the $p \to 1$ limit, Florida is ranked #3 and LSU is ranked #4 (neither shown in the figure), with Oregon falling to #5 and both Nebraska and Colorado falling completely out of the top 10. At smaller values of p (closer to $1/2$) Florida falls to #8, and LSU is not even in the top 10 for $p \lesssim 0.77$.

In contrast, the BCS system worked virtually without controversy in selecting teams for the National Championship game at the end of the 2002 season [11], as there were precisely two undefeated teams, both from major conferences, and no other obvious top contenders. Both polls picked Mi-

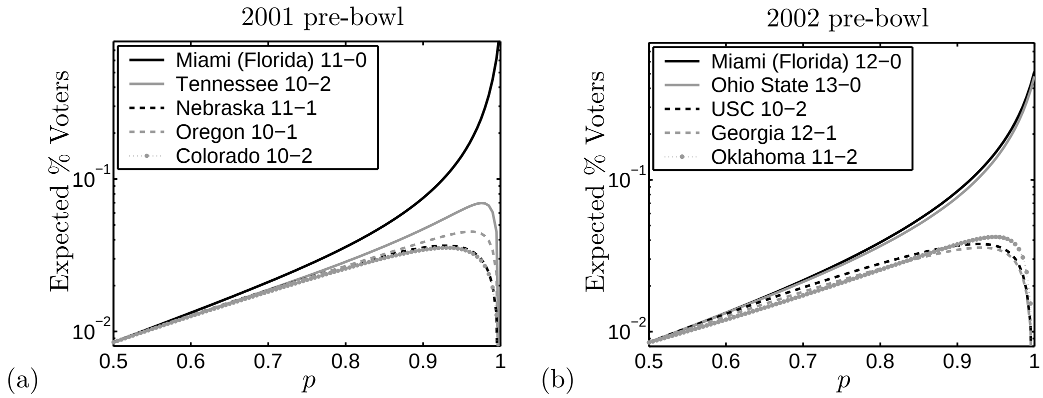


Figure 1: Expected percentages of voters $\bar{v}_i^*/Q$ for highest-ranked teams at different values of the probability p of selecting the winner of a given game, for (a) 2001 pre-bowl and (b) 2002 pre-bowl standings (that is, up to but not including the bowl games).

ami and Ohio State as the top two teams, in agreement with almost all of the seven ranking algorithms used that year (with some changes from the previous year because of just-instituted BCS rules forbidding margin of victory from contributing to the algorithms). The only nonconformist was the New York Times ranking system, which picked Miami and USC as the top two teams—the latter presumably in part due to its difficult schedule. For $p \approx 1/2$, the random walkers also put USC in the top two based on the strength of its schedule (not discernible in the figure), but they agree across most values of p that the top two teams are Miami and Ohio State.

The differences between the 2001 and 2002 pre-bowl results can be further explored by exploiting the statistical properties of expected vote totals to express a measure of confidence in the resulting rankings. An important property of these random walkers is that the matrix $\mathbf{D}$ gives a single attracting equilibrium $\bar{\mathbf{v}}^*$ for a given p, provided the underlying graph representing games played between teams consists of a single connected component. We can see this in three steps:

- The column sums of $\mathbf{D}$ vanish because the sum of the populations remains constant, $0 = Q' = \sum_i v_i' = \sum_i \sum_j D_{ij} v_j$ (that is, the dynamics of $\bar{v}_j$ are confined to a hyperplane of codimension 1).

- The off-diagonal elements of $\mathbf{D}$ are nonnegative and—once the graph becomes a single connected component—all of the off-diagonal elements of D^d are positive, where d is the diameter of the graph, so that vertices with $\bar{v}_j = 0$ necessarily have growing populations. In other words, all average flows go "into" the hyperquadrant in which $\bar{v}_j > 0$ for all j.

- Finally, because (1) is linear, the only possibility is that there is a single attracting sink, $\bar{v}^*$, in the hyperplane.

This argument breaks down when the graph is not connected, as one can then have a disconnected subgraph with zero voter population that remains zero. Alternatively, one can recast this matrix equation in terms of an eigenvalue problem and apply the Perron-Frobenius theorem, as described in [4]; however, given the specific random-walker interpretation built into the rate matrices here, the above arguments already ensure that the expected populations achieve a uniquely determined attracting state, provided that the graph is a single connected component. In the absence of such connectedness, the Perron-Frobenius theorem cannot be applied because the resulting matrices are no longer irreducible.

Any initial voter distribution eventually randomizes completely, and the steady-state distribution of the number of votes v_i cast for the ith team is therefore binomial (for Q trials) with probability $\bar{v}_i^*/Q$, mean $\bar{v}_i^*$, and variance $\bar{v}_i^*(1 - \bar{v}_i^*/Q)$. The joint probability density function of two vertices is not perfectly independent, as the sum over all vertices equals the number of random walkers Q. However, it is nevertheless obtained from distributing Q random trials across two vertices, with probabilities $\bar{v}_i^*/Q$ and $\bar{v}_j^*/Q$, so it is joint binomial. We can exploit this fact to measure confidence in the relative ranking of two teams in terms of the minimum number of voters Q_{min} required to ensure that the expected difference between the number of votes cast for each team is larger than the standard deviation of that difference. This Q_{min} required to successfully distinguish between rank-ordered pairs of teams is plotted for the 2001 and 2002 pre-bowl rankings in figure 2. In particular, we note that the relative numbers of voters required to successfully distinguish #1 from #2 in 2001 and #2 from #3 in 2002 are significantly smaller (have a higher degree of certainty) than, in particular, the distinction between #2 and #3 in 2001.

Because the statistical properties of the random walking voters follow directly from the linear algebra problem (1,2,4), there is no need to simulate independent random walkers to obtain the actual rankings. This simplicity disappears if interactions between random walkers are included, as considered briefly in section 5.

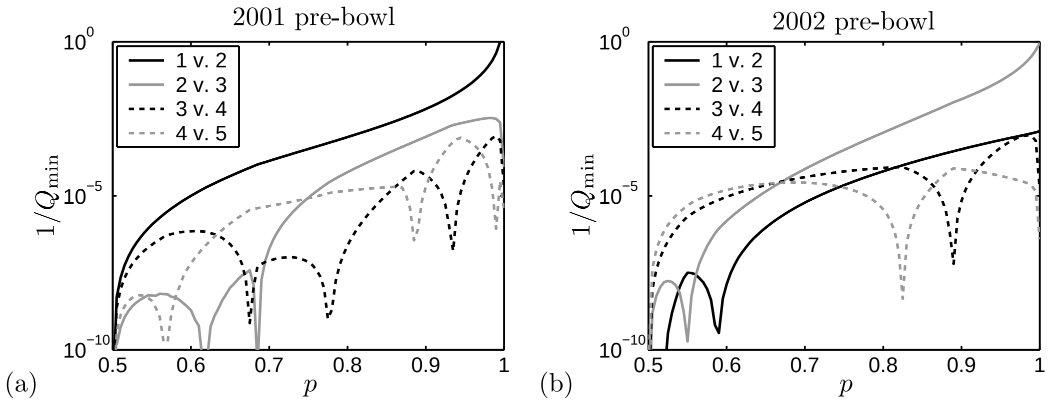


Figure 2: Degree of confidence in the ordering of pairs of teams in the (a) 2001 and (b) 2002 pre-bowl rankings, as quantified by $1/Q_{min}$, where Q_{min} denotes the minimum number of independent random walkers necessary to ensure that the standard deviation of the difference between the expected populations of two teams is smaller than the expected difference.

3 Network Structure

The proposed automatons randomly walk on a graph consisting of the Division I-A football teams (vertices) connected by the games already played between the teams that season (edges). Each season consists of 650–750 games between approximately 115 teams. In every season since 1990, the graph becomes a single connected component by the third or fourth week of the season (not counting faux connections via the single "Non-Division I-A" node that is inserted into the ranking system in order to include games played against Non-Division I-A opponents). The relative quickness in achieving a single connected component (after only about 200 games have been played) is due in part to schools typically playing many of their non-conference games at the beginning of the season.

The degree of each vertex (that is, the number of games played by each team) varies in a narrow range, depending on the point in the season analyzed, with between 10 and 13 connections each in the graph prior to the bowl games and between 10 and 14 connections for the post-bowl games graph. The diameter of the graph, defined as the maximal number of edges needed to travel from one vertex to another (that is, the longest geodesic path), was 4 in every post-bowl graph since 1970. We also calculated the local clustering

Conference	# Teams	2002 Mean ± Std.Dev.	2001 Mean ± Std.Dev.
ACC	9	0.4627 ± 0.0557	0.5364 ± 0.0574
Big 12	12	0.3458 ± 0.0490	0.3859 ± 0.0510
Big East	8	0.4019 ± 0.0592	0.4465 ± 0.0632
Big 10	11	0.3465 ± 0.0469	0.4061 ± 0.0245
CUSA	10	0.4315 ± 0.0500	0.3145 ± 0.0361
MAC	13	0.3716 ± 0.0374	0.3798 ± 0.0548
MWC	8	0.3775 ± 0.0283	0.4415 ± 0.0557
Pac-10	10	0.4282 ± 0.0458	0.4933 ± 0.0490
SEC	12	0.3643 ± 0.0436	0.3769 ± 0.0574
Sun Belt	7	0.3221 ± 0.0520	0.3669 ± 0.0781
WAC	10	0.4054 ± 0.0387	0.447 ± 0.0700
Notre Dame	1	0.1667	0.1273

Table 1: Means and standard deviations of local clustering coefficients of the teams in each Division I-A conference.

coefficient (C_i) for the ith team (team$_i$) in the network, given by [12, 13]

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}\,. \tag{5}$$

In particular, it is important to note the strong heterogeneity of this network. Each conference typically has different average local clustering coefficients that also differ significantly from the coefficients for independent teams such as Notre Dame (see table 1).

Several other network properties [13] can be calculated (including average path lengths, most central teams, and most connected teams), but such computations do not necessarily help explain the random-walker statistics. An important exception is the community structure [14] of the graph, which is useful for understanding the nature of the conference scheduling and the resulting effect on the random-walker statistics. Community structure relies on the notion of betweenness, which measures the number of geodesics on a graph that traverse a given subcomponent (such as a node or an edge) of the graph. In particular, the edge betweenness of an edge is defined as the number of geodesics that traverse it. The algorithm for computing this community structure is given in [14]. Briefly, the edge with the highest betweenness is removed from the graph, and the betweenness is recalculated for the resulting graph before another edge is removed. This process is subsequently repeated until no edges remain. The removal of some of these edges breaks a single graph component into two parts, thereby yielding information that is used to group the network into a hierarchy of communities. Girvan & Newman [14] demonstrated that their community structure algorithm closely reproduced the predefined conference structure of the 2000 football season; we found similar results for other years.

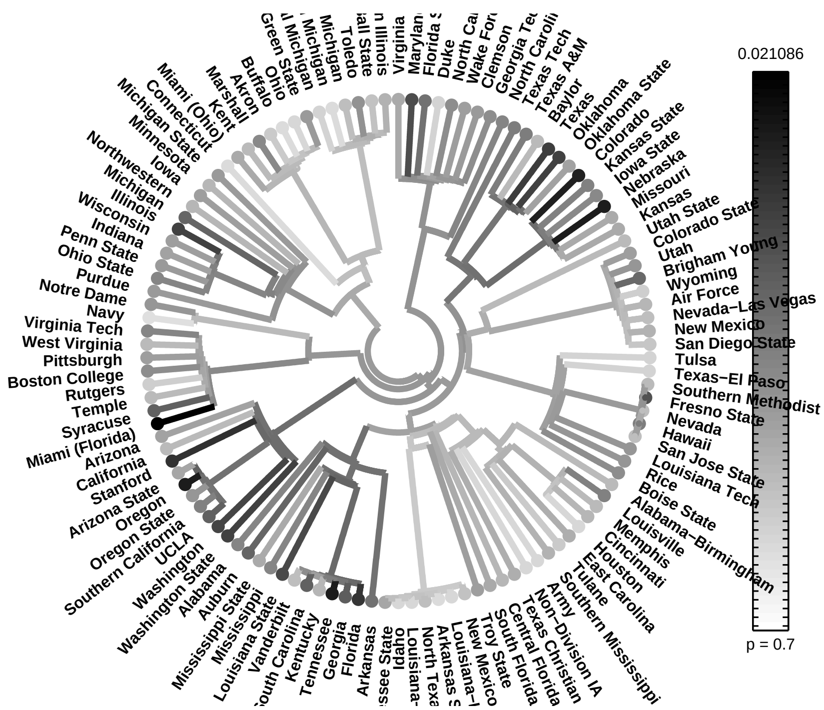


Figure 3: Graphical depiction of the community structure of the 2001 prebowl network at $p = 0.7$, with color-coding of the average votes per team for each community.

Using the 2001 schedule as an example, figure 3 portrays the college football conference structure ("community structure"). As was already demonstrated in [14] for the 2000 Division I-A schedule, the conferences and even their subdivisions are reasonably constructed based on their relative strength of community. One can even interpret qualitatively the relative closeness of different conferences and of the independent teams. Observe, for example, the close connections between Notre Dame, Navy, and the Big East.

This community structure is intimately linked to the dynamics of the random-walking voters, because the specific pairings of the interconference games and the outcomes of those games strongly influence the percentage of time that the voters spend inside given conferences or more general structures such as the divisions inside large conferences or larger groups of nearby conferences. The 2001 pre-bowl community structure in figure 3 is color-coded according to the average percentage of voters per team (for $p = 0.7$) at each level of the hierarchy of communities. Such a plot demonstrates the relatively high average vote counts given here to the SEC, Pac-10, and Big 12, in contrast to the significantly lower average number of votes per team in the Big East despite Miami's first-place standing in this ranking.

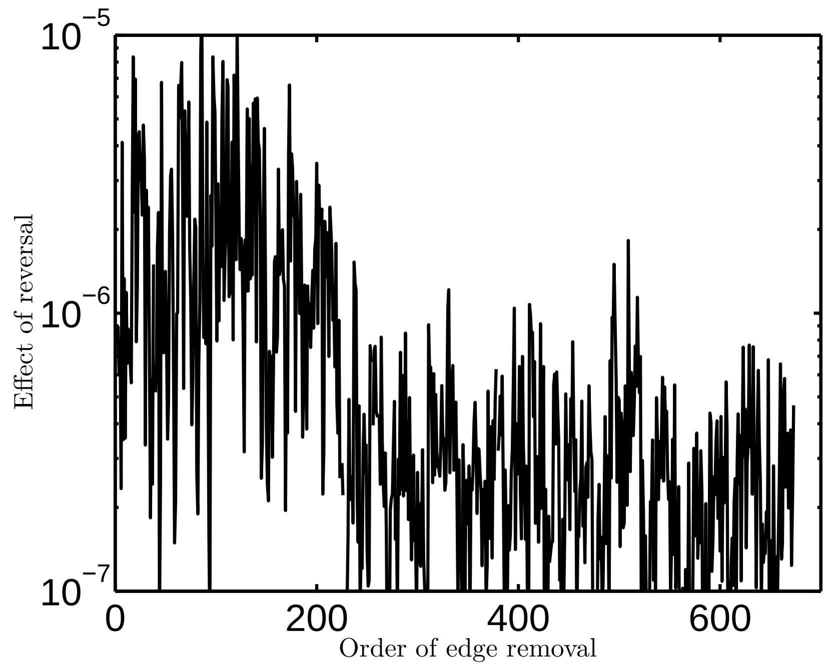
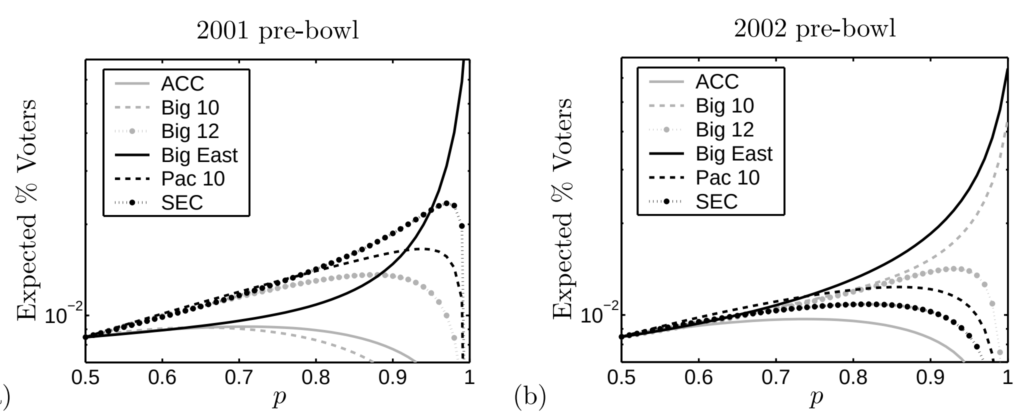


Figure 4: The effect of reversing the outcome of a single game played between teams j and k, quantified by $\sum_{i \neq j,k} |\bar{v}_i^*|^2$, plotted versus the order of that game according to edge removal in the community structure determination.

We further quantified the importance of the relationships between conferences by directly measuring the effect of reversing the outcomes of individual games. Keeping the games in the order in which their respective edges were removed in developing the community structure, we measure the difference between the original voter populations and the new populations calculated with the win-loss outcome of that single game reversed. Because the dominant effect of such reversals is to change the rankings of the two teams involved in that game, we calculate the change to the global distribution of votes by the quantity $\sum_{i \neq j,k} |\bar{v}_i^*|^2$, where the edge removed corresponds to a game between teams j and k. Plotting this quantity (figure 4) versus the order of the edge removed, we observe a sharp transition in average magnitude between the first approximately 200 edges and those that follow. This corresponds roughly to the number of edges removed in the community structure determination at which the hierarchy breaks up into different conferences. That is, these first couple hundred edges in the community structure determination are predominantly non-conference games, while those that follow are intraconference.

The fact that the SEC as a conference in 2001 does slightly better than the Pac-10 and Big 12 is quantified by the average conference votes per team plotted in figure 5a. This presumably results from the interconference games won by SEC teams that year. Only in the $p \to 1$ limit of this plot does the Big East dominate, because every random walker votes for Miami in this limit. As with almost every other measurement here, the 2002 results are completely dominated by the undefeated pair of Miami and Ohio State, with most of the votes going to their respective conferences (Big East and Big 10) in figure 5b.



(a) (b)

Figure 5: Rankings of the major conferences by average votes per team versus the probability p of going with the winner: (a) 2001 pre-bowl and (b) 2002 pre-bowl.

4 Asymptotics at Large and Small p

For a given probability p, the expected populations that go with the winner depend in a complex manner on the details of game schedules and outcomes. In an attempt to better understand the effects of selecting a given value of p, it is instructive to analytically investigate the limiting behaviors near $p = 1/2$ and $p = 1$. We demonstrate, in particular, that the main contributions near $p = 1/2$ include a measurement of strength of schedule, whereas values near $p = 1$ are dominated by undefeated teams.

It is worth noting that this ranking system inherently rewards teams both for good win-loss records and for playing difficult schedules, as do essentially all good ranking schemes, whether determined via an algorithm or a poll. One might thus question whether it makes sense for the BCS system to continue to explicitly add additional contributions for strength of schedule and for number of losses, each currently making up 25% of the BCS formula. One should also keep in mind that in order to reward a team for playing a difficult schedule, one is inherently penalizing—explicitly or implicitly—other teams for playing weaker opponents even when they win those games.

4.1 On Any Given Saturday

Consider $p = 1/2 + \varepsilon$, $\varepsilon \ll 1$. The rate matrix can be rewritten in the form

$$D_{ij} = \frac{1}{2}\Delta_{ij} + \varepsilon \tilde{D}_{ij}\,,$$

where $\mathbf{\Delta}$ is the graph Laplacian, with $\Delta_{ij} = 1$ for $i \neq j$ if nodes i and j are connected (the two teams played each other) and the ith element on the diagonal is equal to the negative of the degree of vertex i. The matrix $\tilde{\mathbf{D}}$ is then the same as the $\mathbf{A}$ matrix off the diagonal, with the values $w_i - l_i$ on the diagonal. The steady-state expected probability at each site, $\bar{v}_j^* = \bar{v}_j^{(0)} + \varepsilon\bar{v}_j^{(1)} + \varepsilon^2\bar{v}_j^{(2)} + \mathcal{O}(\varepsilon^3)$, then obeys

$$0 = \frac{1}{2}\Delta_{ij}\bar{v}_j^{(0)} + \varepsilon\left[\frac{1}{2}\Delta_{ij}\bar{v}_j^{(1)} + \tilde{D}_{ij}\bar{v}_j^{(0)}\right] + \varepsilon^2\left[\frac{1}{2}\Delta_{ij}\bar{v}_j^{(2)} + \tilde{D}_{ij}\bar{v}_j^{(1)}\right] + \mathcal{O}(\varepsilon^3)\,,$$

subject to the normalization condition, $\sum_j \bar{v}_j^{(k)} = Q\delta_{k0}$, for Q voters

The leading-order contribution yields the requirement that $\bar{v}_j^{(0)} = Q/T$ for each j, with the Q votes distributed equally across the T nodes of the graph.

The first order condition then simplifies, according to $\tilde{D}_{ij}\bar{v}_j^{(0)} = 2(w_i - l_i)Q/T$, to give

$$\Delta_{ij}\bar{v}_j^{(1)} = -\frac{4Q}{T}(w_i - l_i)\,.$$

That is, the first correction $\bar{v}_j^{(1)}$ at p close to 1/2 is the potential that solves a discrete Poisson equation (subject to the $\sum_j \bar{v}_j^{(1)} = 0$ constraint) with charge proportional to the win-loss record of each team. It thus incorporates the record of the jth team and is also heavily influenced by the records of the nearest neighbors and other close teams in the graph. That is, the first correction $\bar{v}_j^{(1)}$ is strongly influenced by a "strength of schedule" notion (though this is not the precise strength of schedule used in the BCS), as evidenced by USC's high 2002 pre-bowl ranking at small p. Finally, it is only with the second-order term $\bar{v}_j^{(2)}$ that information pertaining to specific games a given team won and lost begins to be incorporated, as $\bar{v}_j^{(1)}$ considers only net records.

It is insightful to compare this "strength of schedule" dominance near $p = 1/2$ with those obtained from randomly-generated round-robin results. If T teams each play $T-1$ games, one against each of the available opponents, then the graph Laplacian is the matrix of 1's off the diagonal and $-(T-1)$ on the diagonal. Hence, this strength of schedule contribution is also determined completely from a given team's win-loss record. The expected populations of the team vertices for round-robins at $p = 1/2 + \varepsilon$ are therefore linear in p, ordered strictly according to win-loss records.

4.2 Winner Takes All

The asymptotic behavior when $p = 1 - \varepsilon$ is more complicated because the $p = 1$ limiting state depends on the number of undefeated teams and the details of their schedules. For instance, it is not surprising that the 2001 pre-bowl rankings (figure 1a) gives all walkers voting for Miami in the $p \to 1$ limit, since Miami is the single undefeated team. In contrast, the 2002 pre-bowl rankings (figure 1b) only marginally separates the two undefeated teams, Miami and Ohio State, in the $p \to 1$ limit. Miami received roughly 52% of the votes in the limit and Ohio State got the remainder.

While numerically exploring this limit, we found several interesting cases, including two extreme final post-bowl-game random walker rankings using data from the 1971 and 1973 seasons. In 1971, both Nebraska and Toledo went undefeated and untied, but Nebraska (the 1971 Champion in both the AP and UPI polls) receives more than 95% of the votes in the $p \to 1$ limit, because of its stronger schedule. As an aside, it should of course be recognized that Toledo's 1973 season is far from the only historical instance of a team going undefeated on a relatively easy schedule, and any algorithmic or opinionated ranking system has a difficult time figuring out precisely how to rank such squads. In 1973, three teams went undefeated and untied and another three teams went undefeated with one tie each. In the $p \to 1$ limit, five of these teams maintain positive populations of random walkers, because two of the teams with ties, Michigan and Ohio State, got their tie in a game played against each other. With the edge between these two teams equally carrying votes back and forth, Michigan and Ohio State tie for second place in the $p \to 1$ limit, behind Notre Dame's undefeated and untied squad (#1 AP, #4 UPI) but ahead of two undefeated teams with no ties.

The simplest situation to consider asymptotically occurs when a single undefeated team garners all random walker votes for $p = 1 - \varepsilon$ as $\varepsilon \to 0$. The transition rates are then written $D_{ij} = D_{ij}^{(0)} - \varepsilon \tilde{D}_{ij}$ and $\bar{v}_j = \bar{v}_j^{(0)} + \varepsilon \bar{v}_j^{(1)} + \mathcal{O}(\varepsilon^2)$. Here, $\mathbf{D}^{(0)}$ is equal to $(\mathbf{N} + \mathbf{A})/2$ off the diagonal and the negation of the number of losses absorbed by each team on the diagonal (remembering that a tie counts as half a win and half a loss), while $\tilde{\mathbf{D}}$ remains as defined in section 4.1. The limiting state $\bar{v}_j^{(0)} = Q\delta_{ju}$ equals unity for the single undefeated team ("u") and zero otherwise. If there is only one undefeated team with no ties, then this is the single linearly-independent null-vector of $\mathbf{D}^{(0)}$, except for the unlikely but possible scenarios in which two one-loss (tie) teams lost (tied) each other (recalling again the 1973 final rankings). Even with this simplifying assumption, the first-order perturbative analysis requires one to solve the linear algebra problem, $\mathbf{D}^{(0)} \cdot \bar{\mathbf{v}}^{(1)} = \mathbf{b}$, with $b_j = Q\tilde{D}_{ju}$, which neither simplifies nor offers any intuition beyond the original rate equations (1,2,4).

5 Possible Extensions

Generalizations of the ranking algorithm are possible, ranging from trivial redefinitions of the rate matrix $\mathbf{D}$ to fundamental shifts in the mathematical tools required to investigate the random-walking voters. A complete description of the probability distributions of the random walkers given in terms of a single linear algebra problem relies on the independence of the dynamics (vote changes) of the individual voters. Generalizations that violate this independence can drastically change the facility of solution.

The simplest generalizations of the random walker rankings are those that modify the rate matrix $\mathbf{D}$ without changing the independence of the random walkers themselves. For instance, margin of victory, home-field advantage, and the date of the game can easily be incorporated into the definition of $\mathbf{D}$ by replacing the constant probability p of going with the winner with a function that includes these components. The resulting transition rates of walking each direction on a single edge continue to be defined by the outcome of the game represented by that edge, and the associated linear algebra problem determines the probability distribution of each state.

For instance, it would be natural to make the probability of going towards the winner along a given edge higher for a larger margin of victory or for a game won on the road. However, turning this qualitative assertion into a quantitative ranking requires adding more parameters to the ranking algorithm, in stark contrast to the simplifying philosophy considered here. Adding such variables would require at least some discussion about what "good" values of those parameters might be, and tweaking any of these parameters would of course change the resulting rankings. Moreover, starting with 2002 the BCS only uses ranking algorithms that do not include margin of victory, to avoid rewarding unsportsmanlike running-up of the score. Finally, whether or not one wishes to include date of game rests on whether one believes that the National Championship game should be played between the two teams who had the best seasons or the two teams who are playing the best at the end of the season (again, compare this with the favorable way that the random walkers rank Tennessee in 2001). There are thus good reasons to prefer the simpler system with a single probability p, and we have therefore not further explored generalizations that add additional parameters.

A generalization that makes the resulting calculations only slightly more difficult is to give each random walker two votes instead of only one. This change is particularly sensible if the point is to select the two teams to play head-to-head in a National Championship game. We considered such rankings generated from random walkers who each hold two equal votes (as opposed to a #1 vote and a separate #2 vote), using the same probability parameter p of selecting the winner along each edge, subject to the additional constraint that a given voter must cast their two votes for two different teams. This two-vote constrained random walk is most easily understood in terms of independent random walks on the significantly larger network in which each vertex represents a possible pair of votes and the edges between vertices include games played between teams representing one of the two votes, with the other vote held fixed. This again immediately reduces to a linear algebra problem for the expected percentage of votes garnered by each team, albeit of much higher rank ($T[T-1]/2$ for T teams). Clearly, further increases in the number of votes held by each random walker would quickly make the state space so large that the linear algebra solve is no longer feasible.

The expected vote percentages for the 2001 pre-bowl ranking with two votes per walker are presented in figure 6a. Comparing with the single-vote case, figure 1a, Miami of course can only obtain half the votes in the $p \to 1$ limit here because of the constraint that each walker casts the two votes for two different teams, leaving 50% of the votes available to select a #2 team. Interestingly, Tennessee does not do quite as well here relative to the other 2nd-place contenders (especially Oregon) as in the single vote case. The reasons for this shift remain unclear but presumably could be related to Tennessee's relatively close distance to Miami (one common opponent, both teams having defeated Syracuse that season, whereas Oregon is farther away from both Miami and Tennessee in the graph). Another likely possibility is that the constraint that each voter casts two votes for two different teams creates an effective pressure that reduces the numbers of votes cast for the closely-connected top-ranked teams of the SEC (Tennessee, Florida, LSU). Nevertheless, Tennessee maintains the 2nd place spot at most values of p.

Unsurprisingly, the 2002 two-vote rankings (not shown) are very similar to the single-vote values, with Miami and Ohio State splitting the votes in the $p \to 1$ limit.

One can also consider generalizations that destroy the independence of individual random walkers. One possibility is to consider voters that are influenced in their decisions by the number of other walkers voting for each team in a head-to-head game. Whether they are inclined to follow the crowd or to try to be nonconformist, such dependence between the random walkers makes the mean-field calculation nonlinear and removes most of our knowledge of the probability distributions about the mean percentages of votes per team. Another such generalization is to more strongly weight the effects of upsets by increasing the flow of votes towards a lower-ranked team that beats a team with larger numbers of votes. Of course, this flow increase might reverse the ordering of the two teams, thereby removing the upset character and reducing the flow towards the winning team of that game, so there may not even be a statistically-steady ordering of the two teams. A similar com-

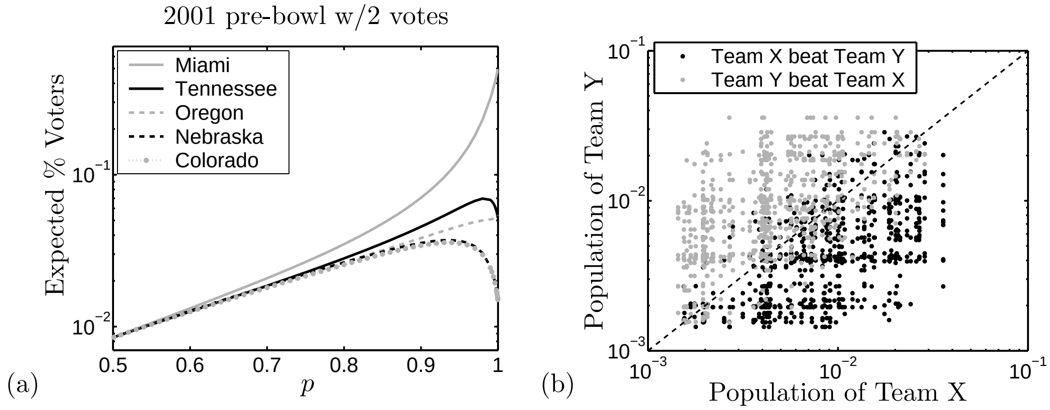


Figure 6: (a) Generalized 2001 pre-bowl rankings with each voter having two nonoverlapping votes. (b) Wins and losses plotted according to the 2001 pre-bowl rankings of each team at p=0.8.

plication occurs if one breaks the independence of the walkers by adding a bonus for beating a team ranked in (for example) the top 10 in the form of reducing the probability of voting for the loser below $(1 - p)$. This reduced flow towards the losing team can knock them out of the top 10, thereby causing the bonus to disappear so that the losing team rises back into the top 10 again. Obviously, the study of any of these interacting random walkers is significantly more difficult than the independent walkers considered in this article.

Finally, we should close with at least some discussion about what values of the probability p yield “good” rankings. Perhaps neither the dominance of strength of schedule near $p = 1/2$ (for example, USC in 2001) nor the emphasis on undefeated seasons near $p = 1$ (for example, BYU in 1984) are appropriate. Rather, one might argue that a more “balanced” value of p may be more appropriate, perhaps at or around $p = 0.75$ (naively split between the two extremes). We consider briefly whether there is an optimal p. For instance, figure 6b shows the organization of the win-loss outcome of every game in the 2001 pre-bowl schedule according to the random walker populations with $p = 0.8$. The votes are not strictly ordered according to the winner and loser of each individual game, but the winners do of course have more votes on average. A possible optimization procedure is to select p to maximize the quality of this ordering. This optimal p might vary from year to year. It may be worth investigating whether such optimization would connect this class of "direct methods" [4] to the more statistically sophisticated maximum likelihood "paired comparisons" originally introduced by Zermelo [15]. Of course, the fact that different values of p can give different orderings of the teams only underscores yet again the inherent difficulty in obtaining uncontroversial rankings because of the relatively few games played between the large number of NCAA Division I-A football teams.

6 Summary

In this article, we developed a simply-defined ranking algorithm in which random walkers individually and independently change their votes between two teams that meet head-to-head, with a single explicit parameter describing the preference towards selecting the winner. We applied this algorithm to the past 33 years of NCAA Division I-A college football, primarily discussing the 2001 and 2002 seasons here. We explored the relationship between the resulting rankings and the underlying network of games played by Division I-A teams. We investigated the asymptotic behavior of this ranking system in the two extremes of the probability parameter. We also considered possible generalizations of this ranking scheme—including the incorporation of additional game factors and models with multiple votes per walker—and discussed complications due to interactions between the random-walking voters.

We gratefully thank Mark Newman for several discussions over the course of this project and Michael Abraham for developing some of the code used in our analysis. This work was funded in part by National Science Foundation Grant DMS-0135290, a Vertical Integration Award funding the Research Experiences for Undergraduates program in which some of this research was performed.

References

[1] Wilson, D. L. (updated 2003) *http://www.cae.wisc.edu/~dwilson/rsfc/rate/*.

[2] Whitaker Software Applications. (2003) *http://www.footballrankulator.com/*.

[3] Kirlin, B. (as downloaded in 2003) *http://www.cae.wisc.edu/~dwilson/rsfc/history/kirlin/fake.html*.

[4] Keener, J. P. (1993) *SIAM Rev.* **35**, 80–93.

[5] Stefani, R. T. (1997) *J. App. Stats.* **24**, 635–646.

[6] Connor, G. R & Grant, C. P. (2000) *Eur. J. App. Math.* **11**, 225–247.

[7] Martinich, J. (2002) *Interfaces* **32**, 85–94.

[8] Wilson, D. L. (updated 2003) *http://www.cae.wisc.edu/∼dwilson/rsfc/rate/biblio.html.*

[9] Howell, J. (as downloaded 2003) *http://www.cae.wisc.edu/∼dwilson/rsfc/history/howell/.*

[10] CFP. (December 9, 2001) *http://www.collegefootballpoll.com/2001_archive_bcs.html.*

[11] CFP. (December 8, 2002) *http://www.collegefootballpoll.com/2002_archive_bcs.html.*

[12] Watts, D. J & Strogatz, S. H. (1998) *Nature* **393**, 440–442.

[13] Newman, M. E. J. (2003) *SIAM Rev.* **45**, 167–256.

[14] Girvan, M & Newman, M. E. J. (2002) *Proc. Nat. Acad. Sci.* **99**, 7821–7826.

[15] Zermelo, E. (1926) *Math. Z.* **29**, 436–460.