Fair and Efficient Ranking in Incomplete Tournaments

Fernando Leiva Bertran [1]

February 13, 2022

Abstract

I discuss basic desirable fairness and internal consistency standards for the case of incomplete tournaments. I present a parsimonious family of scoring methods that satisfies these standards. It includes the win percentage method as a special case. I analyze this family of scoring methods in terms of efficiency, defined as how close a scoring method comes to capturing what the teams' win percentages would have been, in a complete tournament. I show that efficient scoring methods are typically unfair. Finally, using data on betting odds, I calibrate the family of scoring methods to match, as closely as possible, the actual rankings that were used to determine the teams that would go on to compete for the championship of the NCAA division 1 football tournament between 2011 and 2017. I find that the rankings used by the NCAA were generally efficient and unfair, and I quantify the biases present in each year's ranking.

Keywords: Incomplete paired comparisons, tournament ranking, scoring methods.

1. Introduction

In May of 2016, against all odds (5,000 to 1 to be precise), Leicester City won the English Premier League football title. It did so by winning 23 games, drawing 12 and losing only 3, despite fielding a roster of players whose aggregate wage bill was the fifth-lowest in the 20-team league. The team that came in second had won 20, drawn 11 and lost 7. The title was uncontroversial. Just a year later another underdog (this time 1,000 to 1 odds), the University of Central Florida, did *not* win the NCAA division 1 (American) football title despite winning all 13 of its games in a league of 130 teams. The team that was awarded the title had won 13 games and lost 1. The title remains highly controversial. There are several factors that both fuel the controversy surrounding the outcome in the US while justifying the outcome in England, but the most glaring one is given by the very basic structures of the two tournaments: In the first one there are 20 teams that play 38 games each (twice against every opponent) and in the second one

[1]Department of Economics, Arizona State University. CPCOM 465-C, Main Campus, Tempe, AZ 85287. Phone: 480-965-2997. E-mail: fernando.leivabertran@asu.edu.

there are 130 teams that play less than 15 games each (and teams don't even play the same number of games). The first is a case of a twice-complete tournament and the second a case of an incomplete tournament. The fundamental question is straightforward: How to establish a final ranking of the teams that play in a given tournament in a fair and reasonable way. Intuitively, it seems like a very simple task to do this for the first case but not necessarily to do so for the second.[2]

In this work I present basic criteria of fairness, internal consistency and efficiency for developing scoring methods that, through rewards and punishments for winning and losing, assign each team a final score which is then used to establish the final ranking of teams. I also present a particular family of scoring methods that is intuitively simple and satisfies the fairness and consistency criteria. It is then calibrated to match (as closely as possible) the actual rankings used by the NCAA (between 2011 and 2017) in order to quantify, among other things, how fair, efficient and/or biased these rankings were.[3]

A Benchmark Scoring Method

In a complete tournament all teams play each other once. As a result, there is no advantage by any team over any other in terms of the quality of opposition faced. For this reason, the scoring method that has been used in all types of competitions where a complete tournament is played is a system of points where all wins count equally and all losses count equally, but less than a win.[4] The final scores are simply given by the sum of all points assigned to each team. Additionally, if, for example, we want to compare the scoring of two different complete tournaments where the number of teams is different, then a simple solution is to use points per game as a measure instead. This is also known as (or isomorphic to) the win/loss record, or, more precisely, the win percentage.

The simplicity of the win percentage scoring method is its main strength. When thinking about creating a scoring method that is reasonable, our first instinct is to reward winners and punish losers. Because we want to be unbiased and fair, we want our reward for a win to be the same for any team that wins and our punishment for a loss to be the same for any team that loses. Adding

[2] A different issue altogether is the optimal design of a tournament, which is not addressed in this work. For example, the playoff (or knock-out) format is the most efficient way to determine a champion, but a very poor mechanism to determine a complete final ordering of teams.

[3] Allowing, for example, to objectively weigh-in on the 2017 NCAA Football controversy (a non-trivial matter financially, with millions of dollars at stake).

[4] Typically one point is assigned for a win, half a point for a draw and zero for a loss (or its equivalent 2, 1 and 0 respectively) as, for example, in the case of chess tournaments. For the specific case of association football this system was changed to 3, 1 and 0 points respectively in 1995.

up rewards and punishments gives us a very natural way to compare different teams that played the same amount of games. Normalizing by the number of games played is just a minor adjustment that is reasonable when teams have not played the same amount of games.

However, a key to feeling comfortable with this simple method is that at the end of the tournament, all teams will have faced the same competition. Thus, if one team faces weak competition early on and as a result ranks high on the scoring table, we feel at ease because the remaining schedule will either pull it back down where it belongs or prove that this team is at least as good as it's current position suggests. In other words, our approval of the win percentage scoring method is directly tied to the notion that a complete tournament is itself a fair type of tournament.

When moving away from a complete tournament, the win percentage scoring method loses its appeal immediately. The tournament itself is no longer fair. A team that faced weak competition got an unfair advantage and a team that faced tough competition received an unfair disadvantage so, naturally, the win percentage will provide a biased measure of performance. The problem becomes compounded by the fact that the strength of the competition is itself an unknown variable that must be obtained using the same information (the results of all the games played) as the scoring of each team.

With the understanding that when it comes to incomplete tournaments, the very nature of such tournaments is biased (and therefore unfair), the family of scoring methods studied here seeks to correct the bias as much as possible, while preserving the very simple structure of additive rewards and punishments. In other words, I will discuss methods that assign points to teams for wins and losses, but these points are not necessarily confined to being 1 for every win and 0 for every loss.[5]

A Non-Trivial Scoring Problem

To be clear, the games considered here are always between two teams and the objective for any given team is to beat its opponent. That is, when one team wins, the other team necessarily loses and if no team wins then no team loses, which (if allowed) defines a draw.[6] There is no added information that will be used to score teams, that is, there will be no way in which a win can be qualified as *better* or *worse* other than from knowing which team won and which team

[5]For more on the reasons behind this approach see the Extended Introduction in Online Appendix A

[6]for the remainder of this work we will avoid draws for ease of exposition, but they can be easily included in the analysis.

lost. Interestingly there seems to be overwhelming consensus across different leagues of different sports and competitions that, despite there being multiple ways of further qualifying a given win (goals or points difference, judges scores, speed of victory, etc), none of these qualifiers should be used other than to break a tie in win percentage at the end of a complete tournament.[7] In other words, the way in which a win is secured does not matter. At the end of the match, one team walks away with the win and the other with a loss. Whether this is done to give appropriate closure to a match or to avoid teams running up scores against weak opponents or simply as a way to discourage cheating is irrelevant. The operating assumption in this work is that the only information that can be used to score teams is which teams played against each other and who won each game.

The results of all the games played in a given tournament of n teams can be summarized by an $n \times n$ matrix $\mathbf{W}$, labeled the *win matrix* and also referred to as the matrix of *tournament results*, where any entry w_{ij} represents the total number of times i beat j. Thus, any game between any two teams i and j that has been played in the tournament gets recorded either as a win by i (adding 1 to w_{ij}) or a win by j (adding 1 to w_{ji}).[8] This also implies that any row i shows all wins by team i, any column j represents all losses by team j and the diagonal entries are all zero because teams don't play against themselves. A matrix that records the games played by each team but not the results of those games is referred to as the *games matrix* and is defined as $\mathbf{G} \equiv \mathbf{W} + \mathbf{W}^\top$ so that every entry g_{ij} shows the number of times that team i plays against team j.[9] It will also be referred to as the *tournament schedule* in reference to a tournament that has not been played yet.

Other useful matrices and vectors are the *games played* diagonal matrix $\mathbf{D}_G$ which records the total games played by any team i along its diagonal entry d_{ii}. Multiplying $\mathbf{W}$ and $\mathbf{G}$ by a vector of ones $\mathbf{u}$ gives us the total wins vector $\mathbf{w}$ and the total games played vector $\mathbf{g}$ respectively. And if we pre-multiply the wins vector by the inverse of the games played diagonal matrix, each total wins entry gets divided by the total number of games played by the respective team, which gives us the win percentage vector $\widehat{\mathbf{w}}$.

[7] There are few exceptions like, for example, in Rugby Union which has a system of bonus points for scoring a sufficient number of tries or losing by a sufficiently low points margin.

[8] If the tournament allows for draws then this can be captured by adding ½ to both w_{ij} and w_{ji}, with a slight loss of generality that would come from not being able to distinguish two draws from a win and a loss against a given team.

[9] Also, by definition, row i transposed is equal to column i and all rows (and columns) have at least one entry that is not zero (or else the corresponding team would not play any games and therefore not be part of the tournament).

Teams must be scored using only the win matrix as a source of information so the win matrix can also be interpreted as the *scoring problem*. A *scoring function* is a multivariate function $M_n(\cdot)$ that assigns any $n \times n$ scoring problem an $n \times 1$ vector of scores $\mathbf{v}_n$. A *scoring method* is a collection of scoring functions $\{M_n(\cdot)\}_{n=2}^{\infty}$, where each function $M_n(\cdot)$ is applied to any scoring problem with n teams.[10] Finally, we may want the scoring functions of a scoring method to be based on the sum of points received as a result of each win or loss. For this we define a scoring method to be *points-additive* if all its scoring functions can be expressed through a points system where the points assigned to any team i can be decomposed as the following sum:

$$p_i \equiv \sum_j [w_{ij}\ F_{ij}(\mathbf{W}) + w_{ji}\ G_{ij}(\mathbf{W})]$$

where F_{ij} and G_{ij} are functions that assign a number to any scoring problem $\mathbf{W}$ (with F_{ij} representing the points assigned to i for every win against j and G_{ij} the points assigned to i for every loss against j). The score v_i assigned to team i is simply its assigned points p_i divided by the number of games played. In other words, we can obtain the scores vector $\mathbf{v}$ by pre-multiplying $\mathbf{p}$ by $\mathbf{D}_G^{-1}$.

It is important to note that without further restrictions, any scoring method can be expressed as a points-additive method by appropriately selecting F_{ij} and G_{ij} for each of its scoring functions.[11] But this could easily result in violating fundamental common-sense properties.[12] This is why, while there is technically no loss of generality in studying only points-additive scoring methods, if a scoring method isn't purposely designed to score teams through additive rewards and punishments for their victories and losses then it will likely fail multiple common-sense properties when re-expressing it in a points-additive way.

One example of a fundamental common-sense property is *anonymity*. It requires scoring methods to survive re-labeling of teams and it is discussed at length in the literature but taken for granted here.[13] The intuition for it is very simple: Re-labeling team i as team j and vice-versa (plus appropriately changing the win matrix to accommodate this re-labeling of teams) should always result in the exact same scoring of all teams (including i and j but where the new v_i would equal the old v_j and vice-versa). Otherwise the scoring method is fundamentally

[10] For the remainder of the work, unless it is strictly necessary, I will drop the subscript n for easeness of exposition.

[11] For example, for any $M(\mathbf{W})$, define $M'(\mathbf{W})$ as a points-additive method such that $F_{ij}(\mathbf{W}) = G_{ij}(\mathbf{W}) = M_i(\mathbf{W})$. This results in $M(\mathbf{W}) = M'(\mathbf{W})$ for all $\mathbf{W}$.

[12] The example in the previous footnote clearly violates (to name just one) a common-sense notion of a victory being rewarded with more points than a loss against the same opponent

[13] See for example Slutzki and Volij (2005) or Chebotarev and Shamis (1998).

unfair and of no practical use.[14]

This work concentrates specifically on anonymous points-additive scoring methods. I first develop intuitively appealing standards of fairness and internal consistency that are necessary for any such scoring method to satisfy. Second, I present a family of scoring methods that is intuitively simple and satisfies these standards. Third, I analyze this family of scoring methods by developing and applying a measure of efficiency, defined as how close a scoring method comes to capturing what the teams' win percentages would have been had the tournament been complete. It allows for discriminating between the different scoring methods within the family and showcasing the clash between fairness and efficiency that arises. Finally, I apply all the above findings to the NCAA division 1 football tournament by calibrating the family of scoring methods explored here to match as closely as possible the actual rankings that were used to determine the teams that would compete for the championship in seven different years (ranging from 2011 to 2017). This allows me to answer three basic questions about the actual rankings used: How efficient they were, how fair they were and whether there were any biases present that were consistently in favor of certain teams.

2. Fairness Criteria:

Absent any other information to be used, it would not be fair to assign two different teams a different amount of points for beating the same opponent. For the same reason it would not be fair to assign two different teams a different amount of points for losing to the same opponent. Thus, the first two fairness criteria are:

1. **Win fairness**: *A win against opponent j is assigned the same points to any team that beat j.*

Formally, this means that the points-additive method must satisfy the following: $F_{ij}(\mathbf{W}) = F_j(\mathbf{W})$ for all i, j.

2. **Loss fairness**: *A loss against opponent j is assigned the same points to any team that lost to j.*

Formally, this means that the points-additive method must satisfy the following: $G_{ij}(\mathbf{W}) = G_j(\mathbf{W})$ for all i, j.

Notice that the win percentage method satisfies these two criteria, but the criteria leave the door open to assigning different points for beating (or losing to) different opponents. This will be crucial to a scoring method that is applied

[14]Formally, a scoring method $M(\cdot)$ will satisfy anonymity if for any re-labeling (i.e: one-to-one) function $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ and its appropriately re-labeled win matrix $\mathbf{W}^\sigma$, the score of any team i, $v_i = v^\sigma_{\sigma(i)}$, where $\mathbf{v} = M(\mathbf{W})$ and $\mathbf{v}^\sigma = M(\mathbf{W}^\sigma)$

to incomplete tournaments, which requires qualifying a win and/or loss through the only sensible way: The identity of the opponent.

Allowing wins and losses to be assigned different points requires some caution: If we want to interpret a win as a positive signal (in the sense of being superior to the opponent) and a loss as a negative signal (inferior to the opponent) then the former should result in more points assigned than the latter. Thus, our third fairness criterion is:

3. **Win dominance**: *Any win by any team should be assigned more points than any loss by any team.*

Formally, it means that the points-additive method must satisfy the following: $F_{ij}(\mathbf{W}) \geq G_{kl}(\mathbf{W})$ for all i, j, k, l.

This criterion will be key to avoiding nonsensical scoring outcomes like a team that loses all its games (presumably to very strong opponents) being scored higher than a team that wins all its games (to weak opponents nonetheless). There is also a very practical reason for this type of criterion to be applied: If teams are allowed to schedule opponents then without win dominance there would be no incentives to schedule an a-priori very weak opponent against the possibility of scheduling a very tough one that would guarantee a higher score regardless of the result.[15]

The above argument on scheduling incentives is relevant because we implicitly assume that the scoring method will give higher rewards for beating better opponents. If, instead, a victory against a weak opponent were to be assigned more points than a victory against a stronger opponent then the scoring method would be giving an unfair advantage to teams playing against the weaker opponent with respect to those playing against the stronger one. This leads us to our last two fairness criteria. They require the scoring method to assign points for victories that are non-decreasing in opponent strength and to assign points for losses that are also non-decreasing in opponent strength. Notice that a reasonable measure of the strength of an opponent is given by the opponent's score, after all, the scoring method is designed to assign scores that reflect the relative importance of teams in order to rank them from best to worst. So we naturally use the scores as measures of strength. As a result, the strength of an opponent that is used for these two criteria is endogenous to the scoring method that the criteria are being applied to. This concept is referred to in the literature as *self-consistency*. Thus, our last two fairness criteria are:

4. **Self-consistent win fairness**: *The points assigned for victories are non-*

[15] The question of scheduling incentives in incomplete tournaments is explored in a separate working paper (see Leiva Bertran 2019)

decreasing in the opponent's score

Formally, this criterion requires for all i, j, k that $F_{ki}(\mathbf{W}) \geq F_{kj}(\mathbf{W}) \Leftrightarrow v_i \geq v_j$, where v_i and v_j are the i^{th} and j^{th} entries of the scores vector $\mathbf{v} = M(\mathbf{W})$.

5. **Self-consistent loss fairness**: *The points assigned for losses are non-decreasing in the opponent's score.*

Similarly, this criterion requires for all i, j, k that $G_{ki}(\mathbf{W}) \geq G_{kj}(\mathbf{W}) \Leftrightarrow v_i \geq v_j$.

If a points-additive scoring method satisfies all five fairness criteria I will label it a *fair* points-additive scoring method.

3. Consistency Criteria:

Having established basic notions of fairness, it is important to make sure that a scoring method exhibit internal consistency. The criteria presented and discussed here fit within the broader literature on pairwise comparisons. This literature analyzes different scoring methods through properties that may or may not be desirable for a given scoring problem that arises through a set of pairwise comparisons. Competitive tournaments are just one possibility, others being, for example, web-pages that link to each other or research works that cite each other. Of course, each example of a scoring problem will have its own idiosyncracies that result in some of the properties analyzed being either desirable, irrelevant or undesirable. In some instances a given scoring method can be uniquely defined by a set of independent properties. In this work I do not seek out to establish axioms for the family of scoring methods presented here. Instead, in what follows, I go through the main properties analyzed by the literature in addition to new ones I develop and discuss their desirability or lack thereof for the specific case of incomplete tournaments between teams or players that compete in head-to-head fashion each time.

In the literature on pairwise comparisons, similar or even the same property may be given different names by different authors. For consistency I use the work by Gonzalez Diaz, Hendrickx and Lohmann (2014) as a guide since first, their work is comprehensive as it includes an analysis of multiple scoring methods through more than 12 different properties and second, it is also written with the application to competitive tournaments in mind.[16] The only difference between the work here and their work is that they assume that the win matrix is irreducible, which ensures that all the scoring methods they study are well-defined, whereas the application in mind in this work must allow for reducible

[16] Another comprehensive analysis of scoring methods, with a broader scope towards preference aggregation in general, is the work by Chebotarev and Shamis (1998)

win matrices. However, it must be noted that the properties themselves are mostly independent of whether the win matrix is reducible or not.[17]

The most basic properties analyzed by the literature on pairwise comparisons (besides *anonymity* which was discussed here) are the following:

Homogeneity asks for the ranking of teams implied by the scoring method to not change if the win matrix is scaled. A stricter version would also require the scores to be the same. This is very intuitive: Assume that all the games of a given tournament are played for a second time and that the results are the same, then this should have no effect on the final scores. In other words, if we scale all results proportionally then scores should remain the same.

6.1. **Win-scaling consistency**: *Scores are invariant to scaling of the win matrix.*

Formally, this means that for any $k > 0, M(k\mathbf{W}) = M(\mathbf{W})$, where k is a positive integer (although one could conceive of allowing k to be any positive real number).

An even stricter version of this criterion would ask that if the tournament schedule is repeated but with possibly different results each time then the resulting scores should be the average of the ones assigned for each individual tournament (thus, if the results were to be the same then the scores would too). However, this does impose unnecessary linearity into the method's mechanism for obtaining scores, so it is not a property that must hold for the scoring method to show consistency. Nevertheless, if a scoring method does satisfy it then it can be considered a welcome addition:

6.2. **Game-scaling consistency**: *Scores are average in the scaling of the games matrix.*

Formally, for k scoring problems $\mathbf{W}_1, ..., \mathbf{W}_k$ where $\mathbf{G}_i = \mathbf{G}_j$ for all i, j the following must hold: $M(\sum_i \mathbf{W}_i) = (1/k) \sum_i M(\mathbf{W}_i)$.

A related property asks that if two different scoring problems with the same teams result in all players receiving the same scores -an outcome labeled *flat* scores- then the same should be true of the combined problem.

7. **Flatness preservation**: *Flat scores are preserved when combining scoring problems.*

Formally, for two scoring problems $\mathbf{W}_1, \mathbf{W}_2$ where $M(\mathbf{W}_1) = \gamma_1 \mathbf{u}$ and $M(\mathbf{W}_2) = \gamma_2 \mathbf{u}$ with γ_1 and γ_2 real numbers and $\mathbf{u}$ a vector of ones, it must be the case that $M(\mathbf{W}_1 + \mathbf{W}_2) = \delta \mathbf{u}$ for some real number δ.

[17] For more detail on related literature that is not directly applicable to scoring in incomplete tournaments see the Related Literature section in Online Appendix B.

Symmetry asks that the scores be the same for all teams if no team has beat another team more than that team has beat it (a symmetric win matrix). Define net wins for i against j as the number of wins minus the number of losses against that opponent (that is, $w_{ij} - w_{ji}$). Then

8.1. **Symmetry**: *If all net wins are zero then scores are flat.*

Formally, define the *net wins* matrix $\mathbf{N} \equiv \mathbf{W} - \mathbf{W}^\top$. Then if $\mathbf{N} = \mathbf{0}$, $M(\mathbf{W}) = \gamma\mathbf{u}$, for some real number γ.

Symmetry is implied by a stricter property that requires the ordering of teams to flip if all the results in the tournament are flipped (from victories to losses and vice-versa):

8.2. **Inversion**: *Scores are inversely related to those resulting from replacing all wins with losses.*

Formally, if $\mathbf{v} = M(\mathbf{W})$ and $\mathbf{x} = M(\mathbf{W}^\top)$ then for all $i, j, v_i \geq v_j \Leftrightarrow x_j \geq x_i$. Thus, if $\mathbf{W}$ is symmetric then it must be the case that $\mathbf{v} = \mathbf{x}$ and symmetry holds.

Next is a set of properties that I consider *desirable* in that they satisfy basic common sense notions along the following dimensions: The merging of tournaments, in relation to the specific roles that victories and losses play within a scoring method and in comparison to a complete tournament.

Regarding the merging of tournaments, assume that we can partition the set of teams in a given tournament into two subsets where no team in the first subset played against any team in the second subset. Then we want the scores assigned to be the same, whether they are obtained by treating the information as two separate scoring problems or as a single *merged* scoring problem. The intuition is simple: No additional information results from merging two tournaments (with different teams) so we should expect no changes in the scores if we do. Thus, we have:

9. **Merging consistency**: *The scores of the union of two scoring problems with different teams is equal to the union of scores.*

To formalize this requirement we briefly bring back the subscripts that indicate the number of teams in a given tournament. If we have a tournament $\mathbf{W}_1$ with n_1 teams in it and another tournament $\mathbf{W}_2$ with n_2 teams in it then the union of the scores is equal to the scores of the union if

$$\begin{bmatrix} M_{n_1}(\mathbf{W}_1) \\ M_{n_2}(\mathbf{W}_2) \end{bmatrix} = M_{n_1+n_2}\left(\begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix}\right).$$

At a more fundamental level, this consistency requirement forces a scoring method's

scoring functions to perform the same types of operations on the respective scoring problems in order to assign final scores.[18]

Regarding the treatment of victories and losses, a well studied property requires the score of any team to go up if all we do is turn any one of its losses into a victory.

10. **Positive response to the beating relation**: *A team's score is non-decreasing in replacing a loss with a victory against the same opponent.*

Formally, for two scoring problems $\mathbf{W}, \mathbf{W'}$ such that $\mathbf{W} = \mathbf{W'} + \mathbf{E}_{ij} - \mathbf{E}_{ji}$ where for any k, l, $\mathbf{E}_{kl}$ is a matrix of zeroes with a one at entry (k, l), if $\mathbf{v} = M(\mathbf{W})$ and $\mathbf{x} = M(\mathbf{W'})$ then $v_i \geq x_i$.

Another property that is very intuitive, has been well studied by the literature and falls within the self-consistency category is labeled *self-consistent monotonicity*. It has many possible versions that deserve to be analyzed in detail. In its weakest version,[19] it states that for any two teams (that play the same number of games), if we can create a one-to-one relation between the opponents of the first and second teams such that it is always the case that the first team fared no worse than the second and the first teams's opponent's score was no worse than the second's opponent's score then the first team's score should be higher than the second team's score. To summarize:

11.1 **Weak self-consistent monotonicity**: *For two teams that play the same number of games, better results against stronger opposition leads to higher scores.*

Formally, for two teams i and j that played the same number of games we must first define the following multi-sets:[20]

$$\begin{aligned} O_i^+ &\equiv \text{Multi-set of all teams that } i \text{ beat,} \\ O_i^- &\equiv \text{Multi-set of all teams that } i \text{ lost to,} \\ O_j^+ &\equiv \text{Multi-set of all teams that } j \text{ beat,} \\ O_j^- &\equiv \text{Multi-set of all teams that } j \text{ lost to,} \end{aligned}$$

Let there be a one-to-one relation $\sigma : O_i^+ \cup O_i^- \to O_j^+ \cup O_j^-$ such that for all $k \in O_i^+ \cup O_i^-$, $v_k \geq v_{\sigma(k)}$ and for all $k \in O_i^-$, $\sigma(k) \in O_j^-$. Then weak self-consistency holds if $v_i \geq v_j$.

[18] In other words, without this requirements the scoring method could be comprised of scoring functions that are wildly different from each other just because the number of teams in the tournaments are different.

[19] Referred to plainly as *self-consistency* in Chebotarev and Shamis (1998) and in Csato (2019) when discussing an impossibility theorem.

[20] A multi-set can have the same element multiple times in it.

This is an intuitively appealing requirement that would belong as a fairness criterion if it were not for the fact that it is already implied by a points-additive scoring method that satisfies the five fairness criteria discussed above. However, one could ask for more: For example, if a team has better results against better opposition for a number of games and the rest of the team's games are all victories whereas the rest of the other team's games are all losses, then it seems reasonable to ask that the first team be ranked above the second in this case too. Such a requirement would go hand-in hand with the idea of any victory assigning more points than any loss in a points-additive method, provided this version of self-consistent monotonicity still applied to two teams playing the same number of games. For lack of a better word, I will refer to it as *regular* self-consistent monotonicity.

11.2 **Regular self-consistent monotonicity**: *For two teams that play the same number of games, better results against stronger opposition plus added wins versus added losses leads to higher scores.*

Formally, using the same multi-sets as in (11.1), let there be a one-to-one relation $\sigma : O_i^+ \cup O_i^- \rightarrow O_j^+ \cup O_j^-$ such that for all $k \in O_i^-$, $\sigma(k) \in O_j^-$ and $v_k \geq v_{\sigma(k)}$ and for all $k \in O_i^+$ such that $\sigma(k) \in O_j^+$, $v_k \geq v_{\sigma(k)}$. Then regular self-consistent monotonicity holds if $v_i \geq v_j$.

This stricter version is also implied by the five fairness criteria as the following proposition states:

Proposition 1: *A fair points-additive scoring method satisfies regular self-consistent monotonicity*

The intuition for this result is straightforward: If a team plays against better opposition with better results then it will receive a higher score because the points awarded are increasing in the quality of competition due to self-consistent fairness and in the result of its matches due to win domination. In the latter case this is true regardless of the strength of the opposition faced, so if the first team has additional wins and the second team has additional losses then this does not affect the result, regardless of the strength of the opposition played in those additional games. In appendix A I discuss a stricter version of this property that allows for the two teams to play a different amount of games.

Finally, it was argued in the introduction that the win-percentage method is the benchmark for fairness when the tournament is complete. Thus, we don't want to use a scoring method that, in complete tournaments, delivers a different ranking of teams than the one resulting from using win percentages. Notice that the scores vector and the win percentage vector don't have to match in order

to deliver the same ranking, only the implied ordering of teams have to match. More generally we define scores vectors $\mathbf{v}$ and $\mathbf{x}$ as *rank equivalent* if there exists a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$ such that for all $i,\ v_i = h(x_i)$. Intuitively, the ranking that results from using the scores from $\mathbf{x}$ will not be altered by applying the same strictly increasing function to all x_i, so if such a function exists then both scores vectors $\mathbf{v}$ and $\mathbf{x}$ produce the same rankings. Consequently, the last consistency requirement can be expressed as:

12.1. **Win percentage consistency**: *In a complete tournament, scores and win percentages are rank-equivalent.*

Formally, if $\mathbf{G}$ is such that for all i, j if $i \neq j$ then $g_{ij} = k$ where k is a positive integer and $\mathbf{v} = M(\mathbf{W})$ then there exists a strictly increasing $h : \mathbb{R} \to \mathbb{R}$ such that for all $i,\ v_i = h(\widehat{w}_i)$.

A stricter version of this property, labeled *homogenous treatment of victories* requires win percentages to determine which team is scored higher when any two teams face the same opponents (other than possibly facing each other). Thus, define two schedules to be *equivalent* if they include the same opponents, possibly including each other as opponents respectively as well. Then we have

12.2. **Homogenous treatment of victories**: *Scores are increasing in win percentages when teams play equivalent schedules.*

Formally we have that for any two teams i, j, if $\mathbf{g}_i = \mathbf{g}_j + \gamma(\mathbf{e}_j - \mathbf{e}_i)$ where $\mathbf{g}_k$ is the k^{th} column of $\mathbf{G}$ and γ is a non-negative integer, then $v_i \geq v_j \Leftrightarrow \widehat{w}_i \geq \widehat{w}_j$.

Of course, in a complete tournament any two teams always play equivalent schedules so under homogenous treatment of victories, scores and win percentages must be rank-equivalent.

All of the above are reasonable properties to require of a scoring method within the context of an incomplete tournament. There is one additional property that we could ask a scoring method to exhibit that is of practical applicability even though it is not necessary for the scoring method to exhibit internal consistency: The existence of a benchmark score that never changes, regardless of the scoring problem. For example, if the average score is constant for all scoring problems $\mathbf{W}$ then a distance-to-average metric can be created to compare the relative merit of a given team's score. In other words, having a constant average score would anchor a very popular notion, that of *mediocrity*. The average score suffices as a measure of mediocrity for scoring problems where all teams play the same amount of games. However, for the cases where this is not true, a more generalized average score that weighs scores in proportion to the number of games played by the respective team is better suited to repre-

sent mediocrity (and conveniently collapses to the regular average score when all teams play the same amount of games) because it places a higher weight on scores that used more information (more games played) to be obtained. Define the game-weighted average score as $\overline{v}_{\mathbf{W}} \equiv \mathbf{u}^{\top}\mathbf{G}\mathbf{v}/\mathbf{u}^{\top}\mathbf{G}\mathbf{u}$. Then mediocrity is well defined if the game-weighted average score does not vary among different scoring problems:

13. **Well-defined mediocrity**: *The game-weighted average score is constant for all* $\mathbf{W}$.

To be clear, it would be difficult to justify labeling a scoring method that fails this property as inconsistent. However, the usefulness of this property will become apparent as it is further discussed in the section on the practicality of using normalized scores for Monte-Carlo simulations in applied analysis.

As for the properties that are discussed by the literature that I consider undesirable for incomplete tournaments, they are: *Order preservation, Independence of irrelevant matches, bridge-player independence, negative response to losses* and *strict self-consistent monotonicity.* I define and discuss each of these properties in appendix A, where I also provide an example that explains my reasoning for considering them undesirable.

In the following section a family of scoring methods is presented. It is shown to satisfy, under certain conditions, all five fairness criteria and all other desirable properties presented here. In addition, it does not satisfy the properties that were considered undesirable.

4. The generalized points method family:

The functions $\{M_n(\mathbf{W};\alpha)\}_{n=2}^{\infty}$ of the *generalized points* (GP) family of scoring methods assign scores $\mathbf{v}$ according to the following recursive equation:

$$\mathbf{v} = \alpha\widehat{\mathbf{w}} + (1-\alpha)\,\mathbf{D}_G^{-1}\cdot\mathbf{G}\cdot\mathbf{v}, \quad \text{where } \alpha\epsilon\,(0,1]\,. \tag{1}$$

In it, α is a free parameter that defines this family of scoring methods.[21]. To gain some intuition, we can focus on how the score of an individual team is obtained: The recursive formulation for team $i's$ score v_i is:

$$v_i = \alpha\widehat{w}_i + (1-\alpha)\sum_j \frac{g_{ij}}{g_i}v_j.$$

Thus, a team's score is a weighted average of it's win percentage and the

[21] This family of scoring methods is closely related to the generalized row sum method proposed in Chebotarev (1994) and for $\alpha = 1/2$ produces a close match to the scoring method in Colley (2004)

strength of schedule, where the strength of schedule is defined as the weighted average score of all opponents with weights defined by the percentage of all games by i that were played against a given opponent j.

For the recursive formulation in 1 to define a scoring method there should be a unique and finite scores vector $\mathbf{v}$ that satisfies the equation for all $\mathbf{W}$. Otherwise at least one of its scoring functions $M_n(\cdot)$ would not exist and the scoring method would not be well defined.

Proposition 2: *The GP method is a scoring method.*

The proof to proposition 2 shows that the recursive formulation can be re-arranged as an explicit, well-defined function. Additional re-arranging can also show that the GP method is points-additive and satisfies the following:

Corollary 1: *The GP method is a points-additive method that satisfies win and loss fairness and self-consistent win and loss fairness.*

The proof shows that the recursive formulation of the GP scoring method can be rearranged to fit the point-additivity requirement where $F_{ij}(\mathbf{W}) = \alpha + (1-\alpha)v_j$, $G_{ij}(\mathbf{W}) = (1-\alpha)v_j$, and where we know from proposition 2 that v_j is unique and finite for all j. Additionally, since both expressions are independent of i then both win and loss fairness are satisfied and since $\alpha < 1$ then F_{ij} and G_{ij} are increasing in v_j so self-consistent win and loss fairness are satisfied as well.

Expressing the GP method as a points-additive system provides us with a very intuitive interpretation of its underlying structure: For every game played between team i and team j, whenever i beats j the following points are assigned:

Points assigned to winning team $i = \alpha \times 1 + (1-\alpha) \times v_j$

Points assigned to losing team $j = \alpha \times 0 + (1-\alpha) \times v_i$

That is, the winning team receives a weighted average between 1 and the score of the losing team and the losing team receives a weighted average between zero and the score of the winning team. Notice as well that for $\alpha = 1$, this method assigns one point per win and zero per loss, so it collapses to the simple win-percentage method.

This intuitive way of expressing the GP method is useful for finding a sufficient condition on the possible values of α for the GP method to satisfy win dominance (which states that no loss should ever award more points than a win). That sufficient condition is that $\alpha \geq$ ½.

Corollary 2: *If $\alpha \geq$ ½, then the GP method satisfies win dominance*

If $\alpha \geq$ ½ then losing to a team of score $v_i = 1$ would award the loser $(1-\alpha) \leq$ ½ points whereas beating a team of score $v_j = 0$ would award the victor $\alpha \geq$ ½

points. The proof shows that no team can ever achieve a score higher than 1 or lower than 0.

In section 6 the question of a lower bound for α is further discussed. First, to show that if we want this criterion to be satisfied for all $\mathbf{W}$ (and regardless of the number of teams) then ½ is the appropriate lower bound. In other words, $\alpha \geq$ ½ is both a sufficient and a necessary condition. I will refer to this sub-set of GP methods as *globally fair*. Nevertheless, it is easy to show that when the number of teams in a given tournament is low, the corresponding lower bound is not as high. Thus, global fairness can be too restrictive for specific applications. For example, in a two or in a three team tournament, the lower bound is zero. Unfortunately, when the number of teams is high, the total number of possible combinations of results grows exponentially and calculating the actual lower bounds on α very quickly becomes impossible. Instead, the analysis in section 6 turns the question around by making explicit a practical way of *ruling out* values of α that do not satisfy this criterion.

Next we turn to the consistency criteria: Notice that the GP method can also be expressed as an infinite weighted average of the win percentages of all teams. As a result, when scaling the win matrix, as long as the weights remain the same (they do), the scores will also remain the same because the win percentages do so as well. Moreover, linearity in the win percentages ensures that if we scale the games matrix, the resulting scores will be the average of the scores of the individual tournaments. Thus,

Proposition 3: *The GP method satisfies game-scaling consistency.*

Linearity in the win percentages also guarantees that combining scoring problems with flat scores results in the same flat scores.

Proposition 4: *The GP method satisfies flatness preservation.*

The key to this proof is to show that for scores to be flat then it must be the case that win percentages are all one half. Since this is true of both tournaments then this must be true of the combined tournament as well.

Proposition 5: *The GP method satisfies inversion.*

The proof shows that if all wins are turned into losses and vice-versa then the result is for all teams to be assigned 1 minus their original score.

Recall that the GP method uses only wins and losses to generate both the win percentage and the strength of schedule. This means that when two sets of teams don't play each other, there is no mechanism present that would modify the win percentages or the strength of schedule as compared to treating the tournaments separately. Thus,

Proposition 6: *The GP method satisfies merging consistency.*

Key to this proof is that the win matrix of the merged scoring problem is block-diagonal.

For any scoring problem, if the result of one match is changed then the direct effect for the team with an extra win (and one loss less) is greater than any indirect effects that this may generate through the other teams' changes in their respective scores and their possible influence in the original team's strength of schedule.

Proposition 7: *The GP method satisfies positive response to the beating relation.*

This result is non-trivial in that the indirect effects can vary wildly, generating big swings in any team's strength of schedule.

If two teams share the same schedule (with the possible exception of facing each other) then when computing the difference in scores, the strength of schedule cancels out (and if they play each other it is proportional to the same difference in scores) so it is easy to show that the difference in scores is proportional to the difference in win percentages.

Proposition 8: *The GP method satisfies homogenous treatment of victories.*

This also means that when the tournament is complete, the GP method is linear in the win percentage method, not just rank-equivalent.

Corollary 3: *In a complete tournament the GP scores are a linear function of the win percentages.*

The proof of this proposition shows that in a complete tournament, scores $\mathbf{v}$ satisfy the following:

$$\mathbf{v} = \frac{\alpha\,(n-1)}{(n-\alpha)}\widehat{\mathbf{w}} + \frac{(1-\alpha)\,n}{(n-\alpha)2}\mathbf{u} \tag{2}$$

This property is not only important because it implies that the GP method satisfies win percentage consistency but also because it allows us to develop a normalization of the scores, by simply solving for the win percentages in equation (2), that can be used to directly compare the normalized scores of an incomplete tournament to the win percentages of a hypothetical complete tournament. The following sub-section discusses this possibility and its usefulness in, for example, Monte-Carlo simulations.

Normalization of scores

A natural step that follows from defining the family of GP methods using a free parameter α is to determine which value of α to use on a given scoring

problem. In order to do this, the scores assigned by different values of α should be made meaningfully comparable. As a first step, we can show that the game-weighted average score is constant and independent of α.

Proposition 9: *The GP method's game-weighted average score is one half.*

The fact that the game-weighted average score is always equal to ½ allows us to use this metric as an anchor that defines mediocrity and does not change for different values of α.[22]

Unfortunately the above property does not suffice to create a meaningful way of comparing scores that result from using different values of α. Equation (2) shows that as α approaches 0, the scores of all teams approach ½. This is because even though the GP method assigns the same number of total points for all α, it does so in an extreme way when $\alpha = 1$ (by assigning one point to the winner and zero to the loser of any game) but as α goes to zero, the points assigned to each team approaches ½ regardless of who wins. Thus, while the game-weighted average score is constant, the game-weighted standard deviation of scores will differ, depending on the value of α used. Naturally, the game-weighted standard deviation will be higher for higher values of α. This would result in comparisons that always overvalue very good teams and undervalue very bad teams for a higher value of α with respect to a lower one. A normalization of scores that preserves the weighted average score and closes the gap between the weighted standard deviations is required.

Solving for $\widehat{\mathbf{w}}$ in equation (2) defines $\widetilde{\mathbf{v}}$, a linear transformation of $\mathbf{v}$, that can be applied to any ranking problem, regardless of whether it is derived from a complete or an incomplete tournament:

$$\widetilde{\mathbf{v}} \equiv \frac{(n-\alpha)}{\alpha\,(n-1)}\mathbf{v} - \frac{(1-\alpha)\,n}{\alpha\,(n-1)}\frac{1}{2}\mathbf{u}$$

With this normalization, by definition, $\widetilde{\mathbf{v}} \to \widehat{\mathbf{w}}$ if $\mathbf{g} \to (n-1)\mathbf{u}$. Additionally, the normalization is a game-weighted mean-preserving spread of the GP method scores, as the following corollary shows:

Corollary 4: *Normalized scores are a game-weighted mean-preserving spread of the GP method's scores.*

What the normalization does is blow up the scores of teams with higher than mediocre scores and push down the scores of those with lower than mediocre scores without changing the game-weighted average score of ½. For practical purposes, the resulting game-weighted standard deviations are very similar. This

[22] And whenever all teams play the same amount of games the (non-weighted) average score, defined as $\overline{v} \equiv \mathbf{u}^{\top}\mathbf{v}/n$, is ½ regardless of α as well.

feature, along with that of convergence to win percentages when the tournament is complete, is very helpful for comparing the relative merits of the different possible values of α in Monte-Carlo simulations that contrast the scores that result from using the GP method in an incomplete tournament to the win percentages of a benchmark complete tournament.[23] The following section does precisely that.

5. Obtaining the most efficient α:

One way of addressing the question of which α to use is by comparing the rankings that result from different values of α in the incomplete tournament to the ranking that results from the win percentages had the tournament been complete. As explained previously, normalized scores are centered around ½ and have a similar standard deviation for all α. Thus, in this section we take advantage of this feature by directly comparing normalized scores to win percentages instead of indirectly comparing the rankings that each method generates. More precisely, we proceed as follows: Start by simulating a complete tournament and computing the corresponding win percentages. These will be referred to as the *true* win percentages. Next, generate a number of incomplete tournaments using randomly generated incomplete game matrices, but populating the corresponding win matrices with the results from the complete tournament win matrix. Then, for every possible value of α and each win matrix we can obtain the corresponding normalized scores. Finally, we can assess how *efficient* a given value of α is at approximating the win percentages of the complete tournament by calculating the sum, across all n teams, of the squared differences between the win percentage of the complete tournament and the average of the normalized scores of the incomplete tournaments (referred to as the *expected* normalized score of the incomplete tournaments). That is:

$$SS \equiv \sum_{i=1}^{n} [\widehat{w}_i - E(\widetilde{v}_i)]^2 .$$

The sum of squared differences is simply a natural way of evaluating the goodness of fit. Also, if we divide the sum of squares by n and take the square root, we can interpret it as the standard deviation from the true win percent-

[23] There is a caveat to using the normalization, namely that it may result in some normalized scores that are either greater than 1 or less than zero, which complicates any attempt to interpret them as implied win percentages. For example, if $n = 4, g = 2$ and team 1 beat 2 and 3 and teams 2 and 3 beat team 4, then it is easy to show that the resulting scores are: $\mathbf{v}' = [1+\alpha, 1, 1, 1-\alpha]/2$. After normalizing, the scores are $\widetilde{v}' = [1+(n-\alpha)/(n-1), 1, 1, 1-(n-\alpha)/(n-1)]/2$. In this example $\widetilde{v}_1 > 1$ and $\widetilde{v}_4 < 0$. Still, examples where this occurs are infrequent in simulations and in those cases scores are only marginally away from 0 and 1.

age. Finally, it is important to note that the expected normalized scores of the incomplete tournament $E(\widetilde{v}_i)$ will depend on the way in which we randomly create the incomplete tournaments. This last step is done through Monte-Carlo simulations because, other than for very simplistic randomization methods, it is practically impossible to obtain expected scores theoretically. Thus, the most efficient α will depend on the following: The number of teams n, the results of the complete tournament $\mathbf{W}$, the number of games played by each team $\mathbf{g}$ in the incomplete tournament (in the simulations all teams play the same amount of games, so $\mathbf{g} = g\mathbf{u}$) and the random process chosen to assign games in the incomplete tournaments. In the Monte-Carlo simulations, this random process is governed by parameter ρ, related to how likely it is that teams of similar complete tournament win percentage play against each other in the incomplete tournament. It is loosely labeled the *correlation parameter* because a value of $\rho = 1$ represents an incomplete tournament where teams almost exclusively face other teams of similar complete-tournament win percentage, a value of $\rho = 0$ represents a uniformly random chance of playing different opponents and $\rho = -1$ represents an incomplete tournament where teams almost exclusively face opponents that have opposite complete-tournament win percentages.[24]

Simulation results:

A robust feature of the Monte-Carlo simulations is that they produce sums of squared differences that are U-shaped in α. Thus, a *most efficient* α^* exists and is unique, for a given $(n, \mathbf{W}, g, \rho)$. Moreover, the simulations show that for a given (n, g, ρ), α^* is fairly constant in $\mathbf{W}$ as long as the standard deviation of the win percentages $\sigma_{\mathbf{W}}$ is constant. For example, Figure 1 shows 15 different sets of 200 simulations. Each set of simulations has $n = 130$, $g = 11$, $\rho = 0$ and a unique complete-tournament win matrix $\mathbf{W}$ that shares an almost identical $\sigma_{\mathbf{W}}$ with that of the other sets of simulations.[25] In each individual simulation within a set, a new incomplete tournament $\mathbf{G}_T$ is generated (and its corresponding win matrix $\mathbf{W}_T$ populated using win matrix $\mathbf{W}$) by selecting g games to be played by each team with the random pairing of teams governed by parameter ρ. It is clear upon inspection that, each time, α^* is close to 0.35 (the average α^* over

[24] Another way to interpret the correlation parameter ρ is to compare it to the average distance between teams playing in the tournament. If we relabel the teams from highest to lowest true win percentage then one can compute the average distance between teams as

$$\sum_i \sum_j \frac{|i-j|\,\mathbf{G}_{ij}}{\mathbf{u'Gu}}.$$

ρ can then be defined as an appropriate monotonic transformation of the average distance measure

[25] The range of values is between 0.203 and 0.209

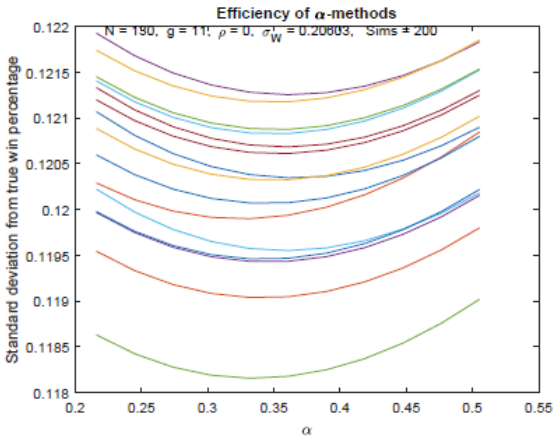
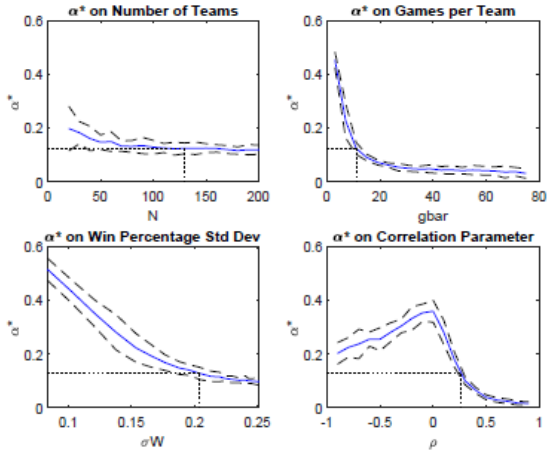


Figure 1: Efficiency



7

Figure 2: Comparative Statics

the 15 sets of simulations is, more precisely, 0.3473).

The simulations allow us to perform comparative statics on the main parameters of the incomplete tournament. Each of the following graphs in figure 2 shows what the most efficient α is as a function of the selected parameter, given benchmark values of the other ones.[26] More precisely, a point along the full line represents, for the corresponding parameter value, the average over all 15 sets of 200 simulations of the most efficient α within each set. The dashed lines display +/- two standard deviations from the average and the point marked with dotted lines represents the benchmark value of the given parameter.

[26]The benchmark values are $n = 130$, $g = 11$, $\rho = 0.265$ and $\sigma_{\mathbf{W}} = 0.204$. They correspond to the average values when calibrating to the NCAA football tournament, as explained in the applied section.

The first graph shows how changing the number of teams n affects the most efficient α. It is clear upon inspection that having more teams has almost no impact on the most efficient α (with the exception of having very few teams). The second graph shows how changing the number of games played by each team affects the most efficient α. It is decreasing in the number of games played because if teams play very few games then the strength of schedule itself is not very useful but as we increase the number of games played it becomes more reliable, so the true win percentage of a team is better predicted by giving the strength of schedule a higher weight. The third graph shows how changing the win percentage standard deviation affects the most efficient α. The bigger the difference between the stronger teams and the weaker teams the higher the win percentage standard deviation will be. This enhances the importance of the strength of schedule in correcting the bias that results from some strong teams playing other strong teams and weak ones playing other weak ones. Thus, the most efficient α declines with a higher win percentage standard deviation. Finally, and most importantly, the fourth graph shows how changing the correlation parameter affects the most efficient α. High and low values of ρ result in an α^* that is considerably lower than when $\rho = 0$. Intuitively, this is because when $\rho = 0$ opponents are drawn uniformly randomly, so this is when correcting for strength of schedule becomes least useful. But as the correlation increases in either direction (towards drawing teams of more similar true win percentages or towards drawing teams of less similar true win percentages) the usefulness of the strength of schedule adjustment increases, hence α^* decreases.

The important takeaway from these results is that it is critical to understand the way in which the incomplete tournament is created in order to determine which α is most efficient. It varies widely depending on how many games each team plays, how evenly matched the teams are, and the way in which games between teams are randomly assigned. Thus, it is important to have a grasp on these factors before deciding on what α to use under this efficiency metric. Moreover, random assignment of teams, whether uniformly or correlated in any way may not even be the way in which teams are assigned opponents in an actual application. In such instances this definition of efficiency would not apply and a new one would have to be developed.

6. Win dominance and the game matrix:

We know that the globally fair lower bound for α that guarantees that every win will award the victor more points than any loss will award the loser is ½. It relies on assuming that it is possible to have a team with a score of 1 and another with a score of 0. The analysis that follows shows that with a sufficient

number of teams n, the maximum possible score for a team does indeed approach 1 and the minimum possible score for a team approaches 0. Thus, $\alpha = ½$ is the appropriate lower bound if the intention is to satisfy win dominance for all $n, \mathbf{g}$.[27] However, even for fairly high values of n (for example 100), a sufficiently low number of games played by each team g leads to much lower bounds on α. This is because teams cannot achieve scores close enough to 0 and 1 even under very favorable circumstances. The constructive approach that generates the favorable circumstances used to prove that the maximum and minimum scores do indeed approach 0 and 1 for high enough n and g can then be used as a practical tool for obtaining lower threshold values of α below which win dominance is not satisfied and above which win dominance may be satisfied.

In order to show that the maximum and minimum possible scores approach 1 and 0, it suffices to show that this is true for a given class of tournament results that share the same property (that is, where for a given n, g, $\mathbf{W}$ has a pre-determined structure): What follows is a class of tournament results that creates a very high distance between the scores of the best and worst teams. The purpose is both to make it easier to prove that the maximum and minimum scores approach 1 and 0 as the number of teams $n \to \infty$ and to create lower bounds for α that are below ½ for low values of n. Assume that each team plays g games. The best team will be defined as the *type-1* team and it beats g second-best teams defined as *type-2* teams who each in turn beats $(g-1)$ *type-3* teams and so on until we reach *type-m*, a mediocre type. The mediocre teams each lose $g/2$ games to *type-(m-1)* teams and beat $g/2$ *type-(m+1)* teams if g is even.[28] Then, each *type-(m+1)* team loses to $(g-1)$ *type-m* teams and beats one *type-(m+1)* team. This goes on until we reach the last type, the single worst team, *type-t*, which loses all g games to type-$(t-1)$ teams. This structure generates a system of t equations and t unknowns where it is shown in Appendix C that the score of best team (the type-1 team) must satisfy:

$$v_1 > (1-\alpha)^{(t-1)/2}\frac{1}{2} + \frac{\alpha}{g} + \frac{(g-1)}{g}\left[1-(1-\alpha)^{(t-1)/2}\right]$$

Notice that for any α, as both t and g go to infinity, v_1 approaches 1. Also, due to symmetry, $v_1 + v_t = 1$, so the score of a type-t team approaches 0 as well.

The class of tournament results used here was designed to create a big distance between the top and bottom teams' scores. Thus, we can create tourna-

[27] In that light, the Colley method can be seen as the GP-method that minimizes α while guaranteeing that win dominance is met for all n.

[28] If g is odd, replace $g/2$ with $(g-1)/2$ and add a draw against a mediocre team to maintain symmetry

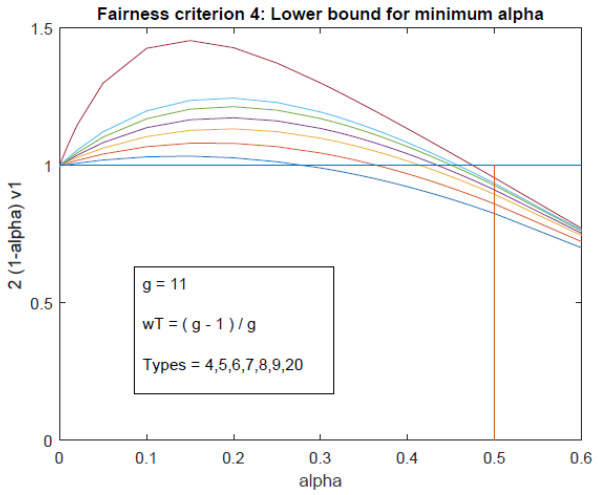


Figure 3: Fairness: Unfair α Methods

ments that fall under this class and obtain the resulting scores, giving us a very good approximation of the relation between the best/worst possible scores and α, n and g.[29] This, in turn, allows us to compute values of α that result in win dominance being met as a function of the number of types t and the number of games g that each team plays. Knowing that v_1 is a function of α and that $v_1 + v_t = 1$, then win dominance will be satisfied under this class of tournament results if

$$(1-\alpha)v_1(\alpha) \leq 1 - (1-\alpha)v_1(\alpha)$$

because the left-hand side is the number of points awarded to a team that has lost to the best team and the right-hand side is the number of points awarded to a team that has beat the worst team. Rearranging, we have

$$2(1-\alpha)v_1(\alpha) \leq 1$$

Figure 3 shows, for different number of types t, whether the left-hand side of the above inequality is indeed below 1 or not as a function of α.

As expected, for high values of α the inequality always holds. For lower values of α it is a non-trivial issue whether the inequality holds or not because as α decreases two opposing forces are at play: The direct effect is that the left-hand side increases due to an increase in $(1-\alpha)$ but the indirect effect is that $v_1(\alpha)$ is converging towards ½. However, Figure 3 shows that it is the direct

[29] It is only an approximation in that the structure described may not represent the structure that *maximizes* the distance in scores between the best and worst teams

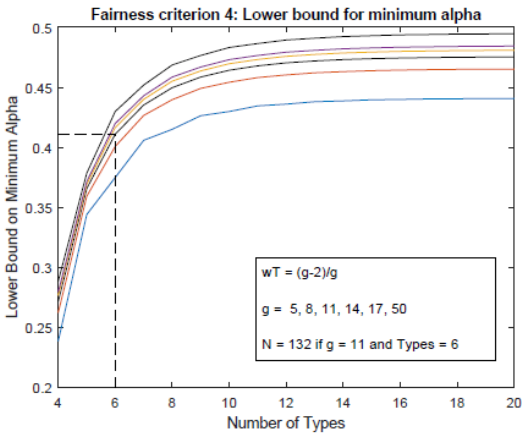


Figure 4: Fairness - Lower Bounds

effect that dominates, so there always exists a threshold level of α below which the GP method does not satisfy win dominance.

It is important to recall that because we are using a specific class of tournament results, the threshold α's found here are only lower bounds for the values of α that satisfy win dominance for a given number of types t and games played per team g. That is, if a better tournament-results structure exists that achieves more distance between the best and worst team using no less teams in total then it simply means that the minimum α that satisfies win dominance is higher than the ones found using this particular structure. In other words, we can only use this specific class of tournaments to *rule out* values of α.[30]

Figure 4 shows the lower bounds for different values of t and g. An interesting feature that can be extrapolated from the graph is that as $t \to \infty$ the lower bound does not converge to ½ if g remains fixed, nor does it converge to ½ as $g \to \infty$ with t fixed. Both $t \to \infty$ and $g \to \infty$ are required in order for the lower bound to be ½. More importantly, the graph shows that win dominance will typically not be satisfied for values of α lower than 0.25 (all that is required is 12 teams playing 5 games each) and for applications where the number of teams is not much higher (for example, more than 25 teams playing 5 games each) a value of α that is greater than 0.35 will be required (for reference, the dotted lines in figure 4 show the lower bound for α in a case of 132 teams playing 11 games each).

[30] In the application to the NCAA football rankings, this approach to obtaining a lower bound will become relevant as it will rule out any ranking method that is consistent with a value of α that is lower than the lower bound that is specific to the number of teams and number of games played by each team in an NCAA football season.

The takeaway from these results is that there is typically a conflict between what is fair and what is most efficient, as defined here. Fairness dictates that we look at values of α that are mostly above 0.35 but maximum efficiency rarely occurs at those levels of α. Thus, if we want to maintain fairness while being as efficient as possible it becomes important to determine the *minimum fair* α because it will most likely be the most efficient one.

7. Application: College Football Rankings

Every year between 2011 and 2017 there have been over 120 teams playing in the upper division of college football. They play between 11 and 14 games each season, depending on their success on the field. In order to advance to a bowl game, a team must finish with a non-losing record. In order to play for the championship (semi-finals and a final in the College football playoff era and just a final during the BCS era) a team must be selected to participate. We are currently in the College Football playoff era, where a committee selects the final four teams that will compete for the national championship. Prior to that, during the BCS era the Associated Press, Coaches and a set of ranking algorithms were weighted in order to determine the two teams that would play for the title.

Whether explicitly or not, all these rankings took into account (among other things) the win percentage and the strength of schedule of a team in order to score and/or rank it. But the rankings were created by means of aggregating the individual subjective rankings of human beings.[31] As a result one would expect biases and/or individual preferences for information other than results and strength of schedule to color the outcomes.[32] Even then, we can establish how close any ranking is to that of some GP method's ranking. This will give us an objective way of determining, at a fundamental level, which of the two main components (win percentage or strength or schedule) is favored more by each individual ranking. Then, having done that for the rankings by a given entity over multiple years, we can ask three questions for that ranking entity: How fairly it ranks, how efficiently it ranks and whether its rankings include specific biases in favor of or against certain teams or conferences.

Best Fit metric used

In most of the rankings that are (or were) used by the NCAA only the top

[31] The cases of the computer rankings are not studied here because the algorithms allegedly used were maintained secret. This offense to science is resolved through indifference. The only exception is the Colley Matrix algorithm which was already mentioned. An analysis of the computer rankings is left for future work.

[32] Good examples are the timing (early in the season vs. late in the season) and/or the margin (by how many points) of a victory or a loss.

25 teams are ranked.[33] The metric favored in this work (labeled LR) in order to assess which GP method comes closest to matching a given ranking is the following:

$$LR \equiv \sum_{i=1}^{25} |\ln(x_i + \kappa) - \ln(y_{i\alpha} + \kappa)|$$

where x_i represents the ranking position of a given ranked team (as ranked by a given entity), $y_{i\alpha}$ is the ranking position of that same team according to the GP method used here and κ is a non-negative number. Notice that this means that $x_i \in \{1, ..., 25\}$ whereas $y_{i\alpha} \in \{1, ..., n\}$ because the GP methods result in a complete ranking of all teams. Notice as well that absolute values are used instead of squares. This is done to minimize the effect of outliers on the total sum, which avoids turning the best fit metric into a metric that best fits to the one or two outliers (with only 25 observations, this is a non-trivial matter). For robustness I also calculated the sum of squares in each case. Finally, notice that as $k \to \infty$ the metric is equivalent to using the sum of the absolute value of the differences and when $\kappa = 0$ it is equivalent to using the difference in natural logarithms. Neither of these extremes is best. The difference in values gives the same weight to all ranking positions. But there is a sense in which a team that was ranked 24th by one method and 21st by another method was indeed closely matched by the two methods but a team that was ranked 1st by one and 4th by the other was not closely matched by the two methods. This is because teams at the top are on the tail of the distribution whereas teams at the bottom (25th out of more than 120 teams) are closer to the median which presumably means that if two ranking methods are truly similar, it is more difficult to misalign the rankings of teams at the top of the 25 team ranking than it is those at the bottom.[34] On the opposite extreme, using the difference in log values gives too much weight to the top ranking positions because very small misalignments there would be equivalent to extreme misalignments at the bottom of the 25-team ranking. In order to strike a balance we can calibrate κ to treat equally the average misalignment in each position. The calibrated value is $\widehat{\kappa} = 2.5$. With few exceptions, the results obtained are robust to changes in κ.

The Fairness Question

[33] The exception is given by some of the computer rankings which not only rank more teams, but also provide a score for each team that can, in principle, be compared to the scores assigned by the family of scoring methods presented here. However, as George Berkeley once said: "*If an algorithm is used but not made public, does it really exist?*"

[34] But there is a case to be made about preventing teams that are ranked near the bottom to influence the metric in the same way as those at the top (after all, the main objective of these rankings is to select only the top 2 or 4 teams that will play for the national championship).

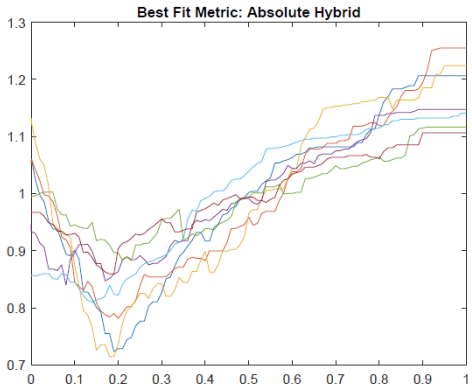


Figure 5: Sum of absolute $\kappa-$log differences for $\kappa = 2.5$

After analyzing the rankings of three different entities (BCS, AP and CFP) over 7 different years (14 different rankings because CFP came into existence as a replacement of the BCS) the results on the question of fairness were conclusive: Not a single one of the 14 rankings (even under any of the best fit metrics used as robustness checks) met the win dominance standard. In other words, not a single one of the rankings was fair. This is because in order for a ranking to be considered fair it would have to have given no more than a 59% weight to the strength of schedule ($\alpha \geq 0.41$).[35] Figure 5 shows the LR sum of absolute differences at $\kappa = 2.5$ between the ranking by the BCS/CFP and that of the GP method for different values of α for each of the 7 years with data.

The stepwise nature of the graphs are the result of small differences in α not always changing the ordering of teams. The best fit occurs where the sum of absolute differences is at its lowest. For all years the minimum occurs at values of $\alpha < 0.41$.

The Efficiency Question

In order to properly assess whether any ranking is efficient we must know the true win percentage of all teams. However, this is only possible through simulations. In any application, teams only play a limited number of games so it is impossible to know what the true win percentage is. To get around this problem (in this application) we can use data on betting odds for every individual game played in a given season. This allows us to obtain an implied

[35]However, one can take a more relaxed approach to the fairness question and only ask a ranking to be *ex-post fair* in the sense of not *effectively* assigning more points for a loss than a win (once the games have been played) as opposed to not *potentially* assigning more points for a loss than a win (before the games are played), the latter being the definition favored in this work. Somewhat surprisingly there are cases that can't even meet this much lower standard, but at least it is a standard that many rankings do meet.

strength vector $\widehat{\mathbf{s}}$, under a very simple strength-based winning probability model discussed in Stern (1991). With this information in hand we can then obtain two important measures: First, an implied $\widehat{\rho}$ [36] and second, through simulations, an implied $\widehat{\sigma}_W$ [37] which can be contrasted to the actual σ_W from the data in order to validate or not the simple strength model used. Finally we can simulate the model, calibrated to each individual season, to obtain the most efficient α in order to contrast it to the one implied by any given ranking.

The strength-based winning probability model

Assume that every team has strength s_i and that the probability that a team i beats opponent j is given by the relative strengths of team i with respect to team j's relative strength. That is,

$$P(w_{ij} = 1/g_{ij} = 1) = \frac{s_i}{s_i + s_j}.$$

Also assume that this random process follows the same structure for any two teams and they are all independently distributed. We don't know the strength of each team but we do know the probability that team i beats team j because we have information on the betting lines from gambling sites. We can solve for team $i's$ strength as a function of team $j's$ strength

$$s_i = \frac{p_{ij}}{1 - p_{ij}} s_j$$

The betting lines from gambling sites will not be perfectly consistent with the $\mathbf{u}^\top \mathbf{g}/2$ equations (between 600 and 750 games in this application) because there are only N unknowns (between 120 and 130 teams). We can recursively define the strength of any given team i as the relative probability-weighted geometric average of the strength of its opponents. This leads to an algorithm that updates the strength vector in iteration k as follows:

$$s_i(k) \equiv \prod_{j \in \mathbf{G}_i} \left[\frac{p_{ij}}{p_{ji}} s_j(k-1) \right]^{\frac{1}{\mathbf{g}_i}}$$

where $s_i(0) = 1$ for all i. For every season, the algorithm converges to a unique (fixed-point) strength vector. The resulting strength vector defines the implied strengths of each team during that season.

Once the implied strength vector $\widehat{\mathbf{s}}$ has been established for a given season, I

[36] Recall that ρ is a measure that assesses how likely it is that two teams of similar true win percentages will play each other. Applied to the strength model this is equivalent, in expectation, to assessing how likely it is that two teams of similar strength will play each other.

[37] Recall that σ_W is the expected standard deviation of the win percentages of each team.

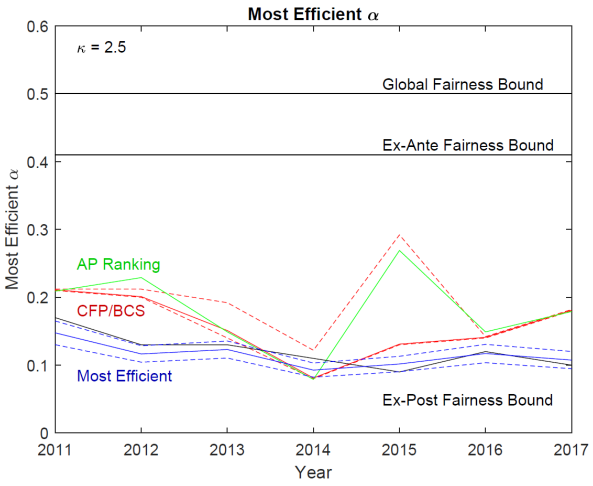


Figure 6: Most Efficient α by year

use it to compute the implied strength of schedule of each team as the geometric mean of the strengths of its opponents in that season. A linear regression of the logarithm of the strength of a team on the logarithm of its strength of schedule will result in an intercept of zero (by definition) and a slope that will depend on how likely it is that teams play opponents of similar strength.[38] This is compared to the slopes that would result for different values of ρ in the simulated model in order to obtain the implied value $\widehat{\rho}$ for a given college football season.

After calibrating $\widehat{\rho}$, I compute the implied win-percentage standard deviation $\widehat{\sigma}_W$ and compare it to the actual σ_W for every given season. In all seasons the two are within 10% of each other (in fact there seems to be a consistent upward bias that may be the result of betting odds being consistently too favorable to the stronger team). Finally, I run 1500 simulations for each strength model calibrated to a given season and compute the corresponding most efficient GP method (i.e: the corresponding α^*). The results are shown in Figure 6.

The solid line labeled *Most Efficient* shows the most efficient levels of α for each year, given that year's calibrated parameter values (with the dashed lines representing +/- one standard deviation). The solid line labeled *CFP/BCS* shows the implied values of α by the BCS (for the years 2011 to 2013) and by the CFP (for the years 2014 to 2017) using $\kappa = 2.5$ as the best-fit parameter.[39]

[38] If they play opponents of similar strength almost exclusively then the slope will be high, if opponents are uniformly randomly assigned then the slope will be zero and if opponents are of opposing strengths then the slope will be negative

[39] For robustness on the best-fit selection of κ, the dashed lines represent the maximum and minimum calibrated values of α when considering all possible values of $\kappa \geq 0$. For robustness on the ranking entity itself, I add the same analysis for the ranking provided by the Associated Press, labeled *AP Ranking*.

Team	2011	2012	2013	2014	2015	2016	2017	Total	#Obs
Michigan State	1.76	-	1.77	1.84	1.00	-	0.85	1.37	5
Florida State	1.36	1.19	1.29	1.00	1.87	1.08	-	1.27	6
Southern California	1.67	-	-	0.96	-	1.26	1.19	1.25	4
Oklahoma State	0.82	-	1.06	-	1.00	1.77	1.72	1.22	5
Baylor	1.00	-	1.12	1.27	1.24	-	-	1.15	4
Louisville	-	0.96	1.10	1.09	-	1.46	-	1.14	4
Wisconsin	1.61	-	1.09	1.10	1.20	1.00	0.88	1.13	6
Oklahoma	0.67	0.93	1.15	-	1.00	1.32	2.11	1.12	6
Clemson	1.24	1.06	1.00	1.15	1.29	1.00	1.00	1.10	7

Figure 7: Most overrated teams between 2011 and 2017

The remaining three solid lines show the lower bounds on α given by the respective fairness criterion used, that is: *Global fairness*, *ex-ante fairness* and, for completeness, *ex-post fairness*. Figure 6 clearly shows that ranking entities are much closer to ranking teams in a way that is more consistent with efficiency rather than fairness.

The Bias Question

Having established an implied $\widehat{\alpha}$ for each ranking of any given season, we can then take the rankings for all seasons by the same entity to try and determine if it consistently over or under-ranks any given teams or conferences. Of course, with more than 120 teams each season and only 7 seasons studied it is inevitable to find certain teams that will be consistently over-rated or consistently under-rated. The interesting question is a quantitative one: By how much the overrated teams are ranked above what the implied GP method would have suggested[40] Figure 7 shows the 9 most overrated teams (of those ranked at least 4 times) in the past 7 years, with a quantitative assessment of how overrated they are: Using $\kappa = 2.5$ as our best-fit parameter, this figure shows how overrated each team was in every year that is was ranked by the entity that was in charge of determining which teams would qualify for the playoffs that year. The rankings used are those of the regular season (excluding bowl games and playoffs). Values above 1 mean that the team was overrated that year. More specifically, the ranking that a team should have received $r^* = (\gamma - 1)\kappa + \gamma r$, where r is the ranking it actually received and γ is the corresponding entry in the figure. Thus, for example, a value of 1.1 means that a team that was ranked 7th should have been ranked 8th or a team that was ranked 17th should have been ranked 19th. Similarly, a value of 1.35 means that a team that was ranked 6th should have been ranked 9th.

The same question can be asked of the different football conferences, which include anywhere from 10 to 16 teams each. Figure 8 shows how over or under-

[40]For underrated teams the same question applies but it is more difficult to answer in the context of this application because we are working with a biased sample of the top 25 teams. This means that the truly underrated teams are simply left out of the ranking.

Conference	2011	2012	2013	2014	2015	2016	2017	Total	#Obs
ACC	1.05	1.13	1.13	1.20	1.53	1.25	0.91	1.17	23
Big 12	0.74	1.22	1.11	1.01	1.04	1.38	1.71	1.11	23
Big 10	1.19	0.98	1.20	1.24	0.94	0.90	0.94	1.03	33
PAC	1.19	0.95	0.92	0.93	0.99	1.04	1.08	1.00	28
Indep	-	1.00	1.07	-	1.10	-	0.82	0.99	4
SEC	1.03	1.04	0.89	0.90	0.85	1.13	1.15	0.98	37
MWC	1.05	1.31	0.91	0.66	-	-	0.96	0.97	6
AAC	-	-	1.05	-	0.94	1.00	0.69	0.89	10
MAC	-	0.97	0.92	-	-	0.66	-	0.87	4

Figure 8: Most overrated conferences between 2011 and 2017

rated every conference (with at least 4 ranked teams in the seven years studied) was. An interesting finding is worth mentioning: All power five conferences[41] were overrated relative to all other smaller conferences. However, with the exception of the ACC and the Big12 the degree of overrating in comparison to the smaller conferences was not very strong.

Another interesting question that can be answered relating ranking biases is whether the right teams were chosen for each season's semi-finals (or just the final in seasons prior to 2014). The following teams were left out of contention for the championship as a result of ranking biases: Oklahoma State in 2011, Florida in 2012, Texas Christian in 2014 and Central Florida in 2017.[42]

Finally, and most glaringly, there was one of the above teams that despite being left out of the championship playoff, despite having had to play a lower-ranked opponent in its bowl game and despite the fact that the teams that did get to play for the championship got an extra boost from playing higher-ranked teams, still managed to rank number one among all teams as objectively measured by the implied ranking method used that year by the very ranking entity that chose to leave it out. That team was the 2017 Knights of the University of Central Florida.

9. Conclusion

Head-to-head match-ups in sports and other competitions conclude with the declaration of a winner and a loser (or the absence of both which defines a draw). After multiple matches involving multiple teams it is natural and customary to establish a final ordering of the teams. The win percentage scoring method is the most widely used, least disputed method to create such an ordering in complete tournaments. Using it as a benchmark, I presented a parsimonious

[41] The power five conferences are the most important conferences in college football due to their fan-base and the strength of their football programs. Naturally, they have an outsized proportion of ranked teams compared to the smaller conferences (84% vs. 16% of all ranked teams). Also, the row labeled *Indep* refers to the University of Notre Dame which is also considered to have a powerhouse football program.

[42] The teams that benefited were Alabama in 2011 and 2012, Ohio State in 2014 and Oklahoma in 2017.

family of scoring methods that satisfy basic fairness and consistency standards for the case of incomplete tournaments (the most important one being that no team should be awarded more points for a win than for a loss, regardless of the opponents). It includes the win percentage method as a special case. I then analyzed this family of scoring methods in terms of efficiency, defined as how close each scoring method comes to capturing what the teams' win percentages would have been had the tournament been complete. I showed that there is a clash between fairness and efficiency in that the most efficient scoring method will typically be an unfair one. Finally, using data on betting odds and results for the NCAA division 1 football tournament I calibrated the family of scoring methods to match as closely as possible the actual rankings that were used to determine the teams that would go on to compete for the championship in each of the years ranging from 2011 to 2017. The main findings are that the rankings used by the NCAA were generally efficient (if anything the strength of schedule component was under-utilized under this metric) but clearly unfair (the strength of schedule component was over-utilized under this metric) and there were quantifiable biases present in the rankings, the most glaring one occurring during the 2017 season where the best team in the country was left out of the four-team playoff that ultimately determined that year's champion.

10. Acknowledgements

I would like to thank Hector Chade, Alejandro Manelli, Claudiney Pereira, Galina Vereschagina, Kevin Reffett, Konstantin Von Beringe and other faculty and students at Arizona State University for their helpful comments. I would also like to give special thanks to Charles Knipp and Nicholas Delafuente for their excellent research assistance. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

[1] Chebotarev, P. [1989] "Generalization of the row sum method for incomplete paired comparisons" *Autom Remote Control* 50, 1103-1113.

[2] Chebotarev, P. [1994] "Aggregation of preferences by the generalized row sum method" *Mathematical Social Sciences* 27, 293-320.

[3] Chebotarev, P. and Shamis, E. [1998] "Characterizations of scoring methods for preference aggregation" *Annals of Operations Research* 80, 299-332.

[4] Chebotarev, P. and Shamis, E. [1999] "Preference fusion when the number of alternatives exceeds two: indirect scoring procedures" *Journal of the Franklin Institute* 336, 205-226.

[5] Colley, W. [] "Colley's Bias Free College Football Ranking Method" *unpublished.*

[6] Csato, L. [2019] "Some impossibilities of ranking in generalized tournaments" *International Game Theory Review* 21(1) 15pgs.

[7] Gonzalez Diaz, J., Hendrickx, R. and Lohmann, E. [2014] "Paired comparisons analysis: an axiomatic approach to ranking methods" *Social Choice and Welfare* 42, 139-169.

[8] Jackson, Matthew [2008] "Social and Economic Networks" *Princeton U. Press.*

[9] Keener, J. (1993), "The Perron-Frobenius Theorem and the Ranking of Football Teams" *, SIAM Review* 35 (1), 80-93.

[10] Leiva Bertran, F. [2019] "Scheduling incentives in incomplete tournaments" *working paper.*

[11] Palacios-Huerta, I. and Volij, O. [2004] "The measurement of intellectual influence" *Econometrica* 72(3), 963-977.

[12] Slutzki, G. and Volij, O. [2005] "Ranking participants in generalized tournaments" *International Journal of Game Theory* 33(2), 255-270.

[13] Slutzki, G. and Volij, O. [2006] "Scoring of web-pages and tournaments-axiomatizations" *Social Choice and Welfare* 26(1), 75-92.

[14] Stern, H. [1991] "On the probability of winning a football game" *The American Statistician* 45(3), 179-183.

Appendix A

The following are five properties discussed by the literature on pairwise comparisons that I consider undesirable for the case of incomplete tournaments:

1. *Order preservation:* It states that if a team's score is higher than another team's score in two different scoring problems then it should still be higher in the combined problem. This property has the undesirable feature of asking the scoring method to essentially ignore the fact that opponents that were deemed strong in the first scoring problem may have been considered as weak in the second tournament, affecting the scores of any teams that played against them. In other words, what is desirable is to have teams' relative scores changing in either direction when more information about the opponents is obtained. As a result, order preservation is unwarranted. To be more specific, when combining two tournaments we will encounter four types of teams: Type 1 are those that did well in both, type 2 are those that did well in the first but not the second, type 3 are those that did well in the second but not the first and type 4 are those that did poorly in both. We can easily imagine how a team that beat teams of type 2 in the first tournament and of type 3 in the second would have a very high relative score in both tournaments if they were scored separately but we would want it to have a mediocre score when both tournaments were combined into one. Order preservation would prevent this from happening.

2. *Independence of irrelevant matches:* It states that the relative scores of any two teams should not depend on the results of matches that don't involve at least one of these two teams. This property is undesirable because it runs counter to the idea of quality of opposition being an important factor in determining a team's score. It's not the same to beat a team that won all its other games than one that lost them. Thus, the relative scores should reflect this and not be independent of it.

3. *Bridge player independence:* If we partition the set of players into three subsets with one of them having a single player in it then that player is considered a bridge player if it is the only link between the other two sets of players. That is, players in one set never played against any player in the other set but at least one player of each set has played against the bridge player. Bridge player independence requires that the results of the bridge player in the second set should not influence the relative scores in the first set. This property is specifically tailored for tournaments with irreducible win matrices (where it is always possible to find a string of players that successively beat each other starting from any player and ending at any player) which is not necessarily the case here. But even if we restrict ourselves to this subset of scoring problems then there is still

a case to be made against it. The bridge player is an opponent of some players in the first set but not of others. Thus, its results against teams in the second set should indeed impact the relative scores of the opponents of the bridge player in the first set and not be independent of them. In a sense, this is a much weaker version of independence of irrelevant matches that is equally unappealing.

4. *Negative response to losses:* It asks that if we start out with all teams receiving the same score and we multiply each team's losses by different constants then the final scores should be inversely related to those constants. The problem with this property is that it ignores the fact that every extra loss implies an extra win too. So we can imagine a situation where one team's losses are multiplied by a relatively high constant but it happened to have given another team its only loss and that team's losses are now multiplied by an even higher constant. On average the first team is not in a bad position after all because it now receives many more new wins than losses. As a result it would be reasonable to score it above a third team whose losses got multiplied by a lower constant. Negative response to losses would prevent this from happening.

5. *Strict self-consistent monotonicity:* If a team has better results against better opposition than another team then it makes sense to score it above the second. This is the essence of self-consistent monotonicity as explained in section 3. The strongest version of this property would expand the requirement to teams that don't play the same number of games by simply allowing the set of all victories for the first team or the set of all losses for the second team to have different cardinality (including being empty sets). However, it can be argued that this may be too much to ask. We can imagine a team that beat one very good opponent versus another that beat a slightly better opponent and also a really bad one. Clearly, strict self-consistent monotonicity would want the scoring method to assign the second team a higher score than the first (whereas regular self-consistent monotonicity would have nothing to say). In essence, it wants the victory against weak opposition to never lower the team's score. Reversing the argument would lead us to also conclude that it wants a loss against stronger opposition to never increase a team's score. This would result in either having to ignore victories to anything other than the best opposition (and vice versa for losses), abandon win and loss fairness or abandon the point's-additive structure altogether. It would have as a consequence that teams that racked up victories against very weak opposition would always get rewarded for it or at least never punished. This is not necessarily an undesirable result, but it can be very constraining when the number of teams is high and the number of games played by each team is very low and we are trying to extract as much

information as possible from the few observations that exist for each team. In more practical terms, if we are comfortable giving a team a very high score because it beat only one really good team plus a number of very weak ones then this is a good property to have. Otherwise it is too strict and its weaker versions should be considered.[43]

Appendix B

Proposition 1: *A fair points-additive scoring method satisfies regular self-consistent monotonicity*

Proof: Let i and j be two teams that played the same number of games such that the following multi-sets can be defined:

$$\begin{aligned} O_i^+ &\equiv \text{Multi-set of all teams that } i \text{ beat,} \\ O_i^- &\equiv \text{Multi-set of all teams that } i \text{ lost to,} \\ O_j^+ &\equiv \text{Multi-set of all teams that } j \text{ beat,} \\ O_j^- &\equiv \text{Multi-set of all teams that } j \text{ lost to,} \end{aligned}$$

Let there be a one-to-one relation $\sigma : O_i^+ \cup O_i^- \to O_j^+ \cup O_j^-$ such that if $k \in O_i^-$ then $\sigma(k) \in O_j^-$ and $v_k \geq v_{\sigma(k)}$ and if $k \in O_i^+$ and $\sigma(k) \in O_j^+$ then $v_k \geq v_{\sigma(k)}$.

For team i we have

$$\begin{aligned} p_i &= \sum_{k=1}^{n} w_{ik} F_{ik}(\mathbf{W}) + w_{ki} G_{ik}(\mathbf{W}) \\ &= \sum_{k=1}^{n} w_{ik} F_k(\mathbf{W}) + w_{ki} G_k(\mathbf{W}) \\ &= \sum_{k \in O_i^+} F_k(\mathbf{W}) + \sum_{k \in O_i^-} G_k(\mathbf{W}) \end{aligned}$$

Similarly for team j we have

$$\begin{aligned} p_j &= \sum_{k=1}^{n} w_{jk} F_{jk}(\mathbf{W}) + w_{kj} G_{jk}(\mathbf{W}) \\ &= \sum_{k=1}^{n} w_{jk} F_k(\mathbf{W}) + w_{kj} G_k(\mathbf{W}) \\ &= \sum_{k \in O_j^+} F_k(\mathbf{W}) + \sum_{k \in O_j^-} G_k(\mathbf{W}) \end{aligned}$$

where in both cases the first equation is due to the method being points-additive,

[43] Interestingly, most of the arguments for and against the appeal of the rankings provided by different entities like the Associated Press, the coaches poll or the playoff committee for the NCAA Football tournament boil down to the treatment of such teams. That is, where to rank teams that win many games because they play very weak schedules.

the second due to win and loss fairness and the third from the definitions of the respective multi-sets as defined above. Subtracting j's points from i's points we have

$$\begin{aligned} p_i - p_j &= \sum_{k \in O_i^+} F_k(\mathbf{W}) + \sum_{k \in O_i^-} G_k(\mathbf{W}) - \sum_{k \in O_j^+} F_k(\mathbf{W}) - \sum_{k \in O_j^-} G_k(\mathbf{W}) \\ &= \sum_{\substack{k \in O_i^+ \\ \sigma(k) \in O_j^+}} \left[F_k(\mathbf{W}) - F_{\sigma(k)}(\mathbf{W})\right] + \sum_{\substack{k \in O_i^+ \\ \sigma(k) \in O_j^-}} \left[F_k(\mathbf{W}) - G_{\sigma(k)}(\mathbf{W})\right] + \sum_{k \in O_i^-} \left[G_k(\mathbf{W}) - G_{\sigma(k)}(\mathbf{W})\right] \end{aligned}$$

where the second line results from the existence of the one-to-one relation σ where we know that if $k \in O_i^-$ then $\sigma(k) \in O_j^-$. The first term on the right-hand side is non-negative because of self-consistent win fairness where we know that $v_k \geq v_{\sigma(k)}$. The second term is non-negative because of win domination . The third term is non-negative because of self-consistent loss fairness where we know that $v_k \geq v_{\sigma(k)}$. As a result, team i is assigned more points than team j and because they both play the same number of games then $v_i \geq v_j$.

Proposition 2: *The GP method is a scoring method.*

Proof:

$$\begin{aligned} \mathbf{v} &= \alpha \widehat{\mathbf{w}} + (1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G} \cdot \mathbf{v} \\ \mathbf{v} &= \alpha \left[\mathbf{I} - (1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G}\right]^{-1} \cdot \widehat{\mathbf{w}} \end{aligned}$$

where the matrix $\left[\mathbf{I} - (1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G}\right]$ is invertible because it has ones along its diagonal and non-positive numbers along off-diagonal entries that add up to $(1-\alpha) < 1$ along any row. Thus, $M(\mathbf{W})$ is well-defined.

Corollary 1: *The GP method is a points-additive method that satisfies win and loss fairness and self-consistent win and loss fairness.*

Proof: Define $F_{ij}(\mathbf{W}) = \alpha + (1-\alpha)v_j$ and $G_{ij}(\mathbf{W}) = (1-\alpha)v_j$ for all i, j. Then for all i we have

$$\begin{aligned} p_i &\equiv \textstyle\sum_j \left[w_{ij}\ F_{ij}(\mathbf{W}) + w_{ji}\ G_{ij}(\mathbf{W})\right] \\ &= \textstyle\sum_j \{w_{ij}\ \ [\alpha + (1-\alpha)v_j] + w_{ji}\,[(1-\alpha)v_j]\} \end{aligned}$$

that is,

$$\begin{aligned} \mathbf{p} &= \alpha\, \mathbf{W} \cdot \mathbf{u} + (1-\alpha)\mathbf{W} \cdot \mathbf{v} + (1-\alpha)\mathbf{W}^\top \cdot \mathbf{v} \\ &= \alpha\, \mathbf{w} + (1-\alpha)\mathbf{G} \cdot \mathbf{v} \end{aligned}$$

which means that

$$\mathbf{v} \equiv \mathbf{D}_G^{-1} \cdot \mathbf{p} = \alpha\, \widehat{\mathbf{w}} + (1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G} \cdot \mathbf{v}$$

The last equation shows that the v_js used to define F_{ij} and G_{ij} are indeed the scores of the GP method. Thus, the GP method is points-additive and since both F_{ij} and G_{ij} are independent of i then win and loss fairness are satisfied.

Corollary 2: *If $\alpha \geq$ ½, then the GP method satisfies win dominance*

Proof:

$$\begin{aligned} \mathbf{v} &= \alpha \left[\mathbf{I} - (1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G}\right]^{-1} \cdot \widehat{\mathbf{w}} \\ &= \alpha \left[\sum_{t=0}^{\infty} \left[(1-\alpha)\, \mathbf{D}_G^{-1} \cdot \mathbf{G}\right]^t\right] \cdot \widehat{\mathbf{w}} \\ &\leq \alpha \sum_{t=0}^{\infty} (1-\alpha)^t\, \mathbf{u} = \mathbf{u} \end{aligned}$$

The first equation comes from the last equation in the proof of proposition 1. The second equation can be derived from the fact that the inverse of any matrix $\mathbf{A}$ can be expressed as $\mathbf{A}^{-1} = \sum_{t=0}^{\infty} (\mathbf{I} - \mathbf{A})^t$. Finally, the inequality results from $\left(\mathbf{D}_G^{-1} \cdot \mathbf{G}\right)^t$ being a stochastic matrix for all t so every term in the infinite sum is a weighted average of the win percentages, discounted by $(1-\alpha)^t$. The infinite sum is positive and finite because $\alpha \in (0, 1]$.

Knowing that every $v_i \in [0, 1]$ then a win against team j awards $F_j = \alpha + (1-\alpha)v_j \geq \alpha$ points and a loss to team k awards $G_k = (1-\alpha)v_k \leq (1-\alpha)$ points. Thus, if $\alpha \geq 1/2$ then $F_j \geq G_k$ for all j, k.

Proposition 3: *The GP method satisfies game-scaling consistency*

Proof: For k scoring problems $\mathbf{W}_1, ..., \mathbf{W}_k$ where $\mathbf{G}_i = \mathbf{G}_k = \mathbf{G}$ for all i, j. Define $\mathbf{W}_\Sigma \equiv \sum_i \mathbf{W}_i$. Then $\mathbf{G}_\Sigma = k\mathbf{G}$ and $\mathbf{D}_{\mathbf{G}_\Sigma} = k\mathbf{D}_\mathbf{G}$. Then we know

$$\mathbf{v}_i = \alpha \mathbf{D}_\mathbf{G}^{-1} \mathbf{W}_i \mathbf{u} + (1-\alpha) \mathbf{D}_\mathbf{G}^{-1} \mathbf{G} \mathbf{v}_i \ , \ \text{for all } i$$

Then

$$\begin{aligned} \overline{\mathbf{v}} &\equiv (1/k) \sum_{i=1}^{k} \mathbf{v}_i \\ &= \alpha(1/k) \mathbf{D}_\mathbf{G}^{-1} \sum_{i=1}^{k} \mathbf{W}_i \mathbf{u} + (1-\alpha) \mathbf{D}_\mathbf{G}^{-1} \mathbf{G} (1/k) \sum_{i=1}^{k} \mathbf{v}_i \\ &= \alpha \mathbf{D}_{\mathbf{G}_\Sigma}^{-1} \mathbf{W}_\Sigma \mathbf{u} + (1-\alpha) \mathbf{D}_{\mathbf{G}_\Sigma}^{-1} \mathbf{G}_\Sigma \overline{\mathbf{v}} \end{aligned}$$

Thus, $\overline{\mathbf{v}} = M(\mathbf{W}_\Sigma)$.

Proposition 4: *The GP method satisfies flatness preservation*

Proof: Let $\mathbf{W}, \mathbf{W}'$ be such that $M(\mathbf{W}_1) = \gamma_1 \mathbf{u}$ and $M(\mathbf{W}_2) = \gamma_2 \mathbf{u}$. Then

$$\begin{aligned} \gamma_1 \mathbf{u} &= \alpha \mathbf{D}_G^{-1} \mathbf{W}\mathbf{u} + (1-\alpha)\, \mathbf{D}_G^{-1} \mathbf{G} \gamma_1 \mathbf{u} \\ &= \alpha \mathbf{D}_G^{-1} \mathbf{W}\mathbf{u} + (1-\alpha)\, \gamma_1 \mathbf{u} \end{aligned}$$

so

$$\gamma_1 \mathbf{u} = \mathbf{D}_G^{-1} \mathbf{W}\mathbf{u} = \widehat{\mathbf{w}}$$

Similarly for $\mathbf{W}'$ we have

$$\gamma_2 \mathbf{u} = \mathbf{D}_{G'}^{-1} \mathbf{W}'\mathbf{u} = \widehat{\mathbf{w}}'$$

As a result, $\gamma_1 = \gamma_2 = \frac{1}{2}$. Knowing this, then $\frac{1}{2}\mathbf{u} = \mathbf{D}_{G+G'}^{-1}(\mathbf{W} + \mathbf{W}')\mathbf{u}$, where

$$\frac{1}{2}\mathbf{u} = \alpha \mathbf{D}_{G+G'}^{-1}(\mathbf{W} + \mathbf{W}')\mathbf{u} + (1-\alpha)\, \mathbf{D}_{G+G'}^{-1}(\mathbf{G} + \mathbf{G}')\frac{1}{2}\mathbf{u}$$

so $M(\mathbf{W} + \mathbf{W}') = \frac{1}{2}\mathbf{u}$.

Proposition 5: *The GP method satisfies inversion*

Proof: Let $\mathbf{v} \equiv M(\mathbf{W})$. Then we know that $\mathbf{v}$ satisfies:

$$\mathbf{v} = \alpha \mathbf{D}_G^{-1} \mathbf{W}\mathbf{u} + (1-\alpha)\, \mathbf{D}_G^{-1} \mathbf{G}\mathbf{v}$$

then

$$\begin{aligned} \mathbf{u} - \mathbf{v} &= \alpha \left(\mathbf{u} - \mathbf{D}_G^{-1} \mathbf{W}\mathbf{u}\right) + (1-\alpha) \left(\mathbf{u} - \mathbf{D}_G^{-1} \mathbf{G}\mathbf{v}\right) \\ &= \alpha \left(\mathbf{I} - \mathbf{D}_G^{-1} \mathbf{W}\right) \mathbf{u} + (1-\alpha)\, \mathbf{D}_G^{-1} \mathbf{G}\, (\mathbf{u} - \mathbf{v}) \\ &= \alpha \mathbf{D}_G^{-1} \mathbf{W}^\top \mathbf{u} + (1-\alpha)\, \mathbf{D}_G^{-1} \mathbf{G}\, (\mathbf{u} - \mathbf{v}) \end{aligned}$$

As a result $M(\mathbf{W}^\top) = (\mathbf{u} - M(\mathbf{W}))$

Proposition 6: *The GP method satisfies merging consistency*

Proof: If we have an incomplete tournament $\mathbf{W}_1$ with n_1 teams and $\mathbf{W}_2$ with

n_2 teams then the union is $\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix}$, with $n = n_1 + n_2$ teams. Thus,

$$\begin{aligned} M_n(\mathbf{W}) &= \alpha \left(\sum_{t=0}^{\infty} (1-\alpha)^t \left(\mathbf{D}_G^{-1} \cdot \mathbf{G} \right)^t \right) \cdot \widehat{\mathbf{w}} \\ &= \alpha \left(\sum_{t=0}^{\infty} (1-\alpha)^t \begin{bmatrix} \mathbf{D}_{G_1}^{-1}\mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{G_2}^{-1}\mathbf{G}_2 \end{bmatrix}^t \right) \cdot \begin{bmatrix} \widehat{\mathbf{w}_1} \\ \widehat{\mathbf{w}_2} \end{bmatrix} \\ &= \begin{bmatrix} \alpha \left(\sum_{t=0}^{\infty} (1-\alpha)^t \left(\mathbf{D}_{G_1}^{-1}\mathbf{G}_1 \right)^t \right) \cdot \widehat{\mathbf{w}_1} \\ \alpha \left(\sum_{t=0}^{\infty} (1-\alpha)^t \left(\mathbf{D}_{G_2}^{-1}\mathbf{G}_2 \right)^t \right) \cdot \widehat{\mathbf{w}_2} \end{bmatrix} \\ &= \begin{bmatrix} M_{n_1}(\mathbf{W}_1) \\ M_{n_2}(\mathbf{W}_2) \end{bmatrix} \end{aligned}$$

Proposition 7: *The GP method satisfies positive response to the beating relation*

Proof: Let $\mathbf{W}, \mathbf{W}'$ be two scoring problems where $\mathbf{W} = \mathbf{W}' + \mathbf{E}_{12} - \mathbf{E}_{21}$, that is two identical scoring problems with the exception of a win by team 2 over team 1 in $\mathbf{W}'$ turning into a loss in $\mathbf{W}$. Define $\Delta\mathbf{v} \equiv M(\mathbf{W}) - M(\mathbf{W}')$, and $\Delta\mathbf{p} \equiv \mathbf{D_G} M(\mathbf{W}) - \mathbf{D_{G'}} M(\mathbf{W}')$. By definition of $\mathbf{W}$ we have $\mathbf{G} = \mathbf{G}'$ and $\mathbf{D_G} = \mathbf{D_{G'}}$, so $\Delta\mathbf{p} = \mathbf{D_G}\Delta\mathbf{v}$.

In order to prove that the GP method satisfies positive response to beating we must show that $\mathbf{e}_1^\intercal \Delta\mathbf{p} > \mathbf{0}$.

Define for any $\mathbf{v}$,

$$Z(\mathbf{v}) \equiv \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \hat{\mathbf{G}}^t \mathbf{v},$$

where $\hat{\mathbf{G}} \equiv \mathbf{G}\mathbf{D}_G^{-1}$. Notice that it follows from the definition that $Z(\gamma\mathbf{v}) = \gamma Z(\mathbf{v})$ and $Z(\mathbf{v} + \mathbf{w}) = Z(\mathbf{v}) + Z(\mathbf{w})$, for all $\mathbf{v}, \mathbf{w}$ and $\gamma > 0$.

Finally, define

$$\begin{aligned} \hat{g}_{ij} &\equiv \hat{\mathbf{G}}_{ij} = \mathbf{e}_i^\intercal \hat{\mathbf{G}} \mathbf{e}_j \\ \mathbf{v}_i &= \mathbf{E}_{ii}\mathbf{v}, \ (\text{so } \mathbf{v} = \mathbf{v}_1 + ... + \mathbf{v}_n) \\ v_i &= \mathbf{e}_i^\intercal \hat{\mathbf{G}} \mathbf{v} \end{aligned}$$

As a result, $\Delta\mathbf{p} = Z(\mathbf{e}_1) - Z(\mathbf{e}_2)$, where

$$\begin{aligned}
Z(\mathbf{e}_i) &= \alpha\sum_{t=0}^{\infty}(1-\alpha)^t\hat{\mathbf{G}}^t\mathbf{e}_i = \alpha\mathbf{e}_i + (1-\alpha)Z(\hat{\mathbf{G}}\mathbf{e}_i) \\
&= \alpha\mathbf{e}_i + (1-\alpha)\left[\mathbf{e}_1^{\intercal}\hat{\mathbf{G}}\mathbf{e}_i Z(\mathbf{e}_1) + ... + \mathbf{e}_n^{\intercal}\hat{\mathbf{G}}\mathbf{e}_i Z(\mathbf{e}_n)\right] \\
&= \alpha\mathbf{e}_i + (1-\alpha)\left\{\mathbf{e}_1^{\intercal}\hat{\mathbf{G}}\mathbf{e}_i Z(\mathbf{e}_1) + \sum_{k_1=2}^{n}\mathbf{e}_{k_1}^{\intercal}\hat{\mathbf{G}}\mathbf{e}_i\left[\alpha\mathbf{e}_{k_1} + (1-\alpha)Z(\hat{\mathbf{G}}\mathbf{e}_{k_1})\right]\right\} \\
&\vdots \\
&= \alpha\mathbf{e}_i + (1-\alpha)\hat{g}_{1i}Z(\mathbf{e}_1) + \alpha(1-\alpha)\sum_{k_1=2}^{n}\hat{g}_{k_1 i}\mathbf{e}_{k_1} + (1-\alpha)^2\sum_{k_1=2}^{n}\hat{g}_{1k_1}\hat{g}_{k_1 i}Z(\mathbf{e}_1) \\
&\qquad + \alpha(1-\alpha)^2\sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i}\mathbf{e}_{k_2} + (1-\alpha)^3\sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i}Z(\hat{\mathbf{G}}\mathbf{e}_{k_2}) \\
&= \left[(1-\alpha)\hat{g}_{1i} + (1-\alpha)^2\sum_{k_1=2}^{n}\hat{g}_{1k_1}\hat{g}_{k_1 i} + (1-\alpha)^3\sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\hat{g}_{1k_2}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i} + ...\right]Z(\mathbf{e}_1) \\
&\qquad + \alpha\left[\mathbf{e}_i + (1-\alpha)\sum_{k_1=2}^{n}\hat{g}_{k_1 i}\mathbf{e}_{k_1} + (1-\alpha)^2\sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i}\mathbf{e}_{k_2} + ...\right]
\end{aligned}$$

After replacing into $\Delta\mathbf{p}$ we have

$$\begin{aligned}
\mathbf{e}_1^{\intercal}\Delta\mathbf{p} &= \mathbf{e}_1^{\intercal}\left(Z(\mathbf{e}_1) - Z(\mathbf{e}_2)\right) \\
&= \left\{1 - \left[(1-\alpha)\hat{g}_{12} + (1-\alpha)^2\sum_{k_1=2}^{n}\hat{g}_{1k_1}\hat{g}_{k_1 2} + ...\right]\right\}\mathbf{e}_1^{\intercal}Z(\mathbf{e}_1)
\end{aligned}$$

where we know that for $\alpha = 0$, the sum in the square brackets $= 1$ because

$$\begin{aligned}
1 &= \sum_{k_1=1}^{n}\widehat{g}_{k_1 i} = \hat{g}_{1i} + \sum_{k_1=2}^{n}\hat{g}_{k_1 i} \\
&= \hat{g}_{1i} + \sum_{k_1=2}^{n}\hat{g}_{k_1 i}\left(\sum_{k_2=1}^{n}\hat{g}_{k_2 k_1}\right) = \hat{g}_{1i} + \sum_{k_1=2}^{n}\hat{g}_{k_1 i}\left(\hat{g}_{1k_1} + \sum_{k_2=2}^{n}\hat{g}_{k_2 k_1}\right) \\
&\vdots \\
&= \hat{g}_{1i} + \sum_{k_1=2}^{n}\hat{g}_{1k_1}\hat{g}_{k_1 i} + \sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\hat{g}_{1k_2}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i} + \sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\sum_{k_3=2}^{n}\hat{g}_{1k_3}\hat{g}_{k_3 k_2}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i} \\
&\qquad + \sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\sum_{k_3=2}^{n}\sum_{k_4=2}^{n}\hat{g}_{1k_4}\hat{g}_{k_4 k_3}\hat{g}_{k_3 k_2}\hat{g}_{k_2 k_1}\hat{g}_{k_1 i} + ...
\end{aligned}$$

Thus, for $\alpha > 0$ we have $\mathbf{e}_1^\intercal \Delta\mathbf{p} > 0$.

Proposition 8: *The GP method satisfies homogenous treatment of victories.*

Proof: Consider two teams i and j that have the same schedule. that is, they face the same oponents the same number of times in each case (and may or may not play each other as well). Then

$$v_i = \alpha\widehat{w}_i + (1-\alpha)\sum_{k=1}^{n}\frac{g_{ik}}{g_i}v_k$$

and

$$v_j = \alpha\widehat{w}_j + (1-\alpha)\sum_{k=1}^{n}\frac{g_{jk}}{g_j}v_k.$$

Subtracting one from the other gives us

$$\begin{aligned} v_i - v_j &= \alpha\left(\widehat{w}_i - \widehat{w}_j\right) + (1-\alpha)\sum_{k=1}^{n}\left(\frac{g_{ik}}{g_i} - \frac{g_{jk}}{g_j}\right)v_k \\ &= \alpha\left(\widehat{w}_i - \widehat{w}_j\right) + (1-\alpha)\left(\frac{g_{ij}}{g_i}v_j - \frac{g_{ji}}{g_j}v_i\right) \end{aligned}$$

where the second equality comes from the fact that they have the same schedule. Since $g_i = g_j$ and $g_{ij} = g_{ji}$ then

$$v_i - v_j = \frac{\alpha}{1 - \frac{g_{ij}}{g_i}(1-\alpha)}\left(\widehat{w}_i - \widehat{w}_j\right)$$

so the difference in scores is proportional to the difference in win percentages.

Corollary 3: *In a complete tournament the GP scores are a linear function of the win percentages.*

Proof: In a round-robin, $g_{ij} = 1$ for all i, $j \neq i$, $\mathbf{G}\cdot\mathbf{u} = (n-1)\,\mathbf{u}$ and $\mathbf{D}_G = (n-1)\mathbf{I}$. Thus,

$$\begin{aligned} \mathbf{v} &= \alpha\widehat{\mathbf{w}} + (1-\alpha)\,\mathbf{D}_G^{-1}\cdot\mathbf{G}\cdot\mathbf{v} \\ \mathbf{v} &= \alpha\widehat{\mathbf{w}} + (1-\alpha)\,\frac{1}{(n-1)}\mathbf{I}\cdot\left[(\mathbf{G}+\mathbf{I})\cdot\mathbf{v} - \mathbf{I}\cdot\mathbf{v}\right] \\ (n-\alpha)\mathbf{v} &= \alpha\,(n-1)\,\widehat{\mathbf{w}} + (1-\alpha)\,(\mathbf{G}+\mathbf{I})\cdot\frac{1}{(n-1)}\mathbf{I}\cdot\mathbf{p} \\ \mathbf{v} &= \frac{\alpha\,(n-1)}{(n-\alpha)}\widehat{\mathbf{w}} + \frac{(1-\alpha)\,n}{(n-\alpha)2}\mathbf{u} \end{aligned}$$

Proposition 9: *The GP method's game-weighted average score is one half.*

Proof: We know that

$$\mathbf{v} = \alpha \mathbf{D}_{\mathbf{G}}^{-1}\mathbf{W}\mathbf{u} + (1-\alpha)\mathbf{D}_{\mathbf{G}}^{-1}\mathbf{G}\mathbf{v},$$

so

$$\begin{aligned}
\mathbf{u}^{\top}\mathbf{G}\mathbf{v} &= \alpha \mathbf{u}^{\top}\mathbf{G}\mathbf{D}_{\mathbf{G}}^{-1}\mathbf{W}\mathbf{u} + (1-\alpha)\mathbf{u}^{\top}\mathbf{G}\mathbf{D}_{\mathbf{G}}^{-1}\mathbf{G}\mathbf{v} \\
&= \alpha \mathbf{u}^{\top}\mathbf{W}\mathbf{u} + (1-\alpha)\mathbf{u}^{\top}\mathbf{G}\mathbf{v} \\
&= \mathbf{u}^{\top}\mathbf{W}\mathbf{u},
\end{aligned}$$

where the second line results from $\mathbf{u}^{\top}\mathbf{G}\mathbf{D}_{\mathbf{G}}^{-1} = \mathbf{u}^{\top}$ and the third line from solving for $\mathbf{u}^{\top}\mathbf{G}\mathbf{v}$. As a result,

$$\frac{\mathbf{u}^{\top}\mathbf{G}\mathbf{v}}{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}} = \frac{\mathbf{u}^{\top}\mathbf{W}\mathbf{u}}{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}} = \frac{1}{2}$$

Corollary 4: *Normalized scores are a game-weighted mean-preserving spread of the GP method's scores.*

Proof: We know that

$$\widetilde{\mathbf{v}} = \frac{(n-\alpha)}{\alpha\,(n-1)}\mathbf{v} - \frac{(1-\alpha)\,n}{\alpha\,(n-1)}\frac{1}{2}\mathbf{u},$$

so

$$\begin{aligned}
\frac{\mathbf{u}^{\top}\mathbf{G}\widetilde{\mathbf{v}}}{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}} &= \frac{(n-\alpha)}{\alpha\,(n-1)}\frac{\mathbf{u}^{\top}\mathbf{G}\mathbf{v}}{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}} - \frac{(1-\alpha)\,n}{\alpha\,(n-1)}\frac{1}{2}\frac{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}}{\mathbf{u}^{\top}\mathbf{G}\mathbf{u}} \\
&= \frac{(n-\alpha)}{\alpha\,(n-1)}\frac{1}{2} - \frac{(1-\alpha)\,n}{\alpha\,(n-1)}\frac{1}{2} = \frac{1}{2}.
\end{aligned}$$

Also

$$\begin{aligned}
\widetilde{\mathbf{v}} - \frac{1}{2}\mathbf{u} &= \frac{(n-\alpha)}{\alpha\,(n-1)}\mathbf{v} - \left[\frac{(1-\alpha)\,n}{\alpha\,(n-1)} + 1\right]\frac{1}{2}\mathbf{u} \\
&= \frac{(n-\alpha)}{\alpha\,(n-1)}\left(\mathbf{v} - \frac{1}{2}\mathbf{u}\right)
\end{aligned}$$

As a result, for any $v_i \geq 1/2$, we have $\widetilde{v}_i \geq v_i$ because $(n-\alpha) > \alpha\,(n-1)$ and for any $v_i \leq 1/2$, we have $\widetilde{v}_i \leq v_i$ for the same reason.

Appendix C

The structure described generates the following system of t equations and t unknowns:

$$
\begin{aligned}
v_1 &= \alpha + (1-\alpha)v_2 \\
v_2 &= \alpha\frac{(g-1)}{g} + (1-\alpha)\left[\frac{1}{g}v_1 + \frac{g-1}{g}v_3\right] \\
&\vdots \\
v_{m-1} &= \alpha\frac{(g-1)}{g} + (1-\alpha)\left[\frac{1}{g}v_{m-2} + \frac{g-1}{g}v_m\right] \\
v_m &= \alpha\frac{1}{2} + (1-\alpha)\left[\frac{1}{2}v_{m-1} + \frac{1}{2}v_{m+1}\right] \\
v_{m+1} &= \alpha\frac{1}{g} + (1-\alpha)\left[\frac{g-1}{g}v_m + \frac{1}{g}v_{m+2}\right] \\
&\vdots \\
v_{t-1} &= \alpha\frac{1}{g} + (1-\alpha)\left[\frac{g-1}{g}v_{t-2} + \frac{1}{g}v_t\right] \\
v_t &= (1-\alpha)v_{t-1}
\end{aligned}
$$

In this scenario, $m = (t+1)/2$, where t is odd and the number of teams n is increasing in g and t. An interesting property of this class of tournament results is that it is symmetric in the types, that is, the number of type$-j$ teams is equal to the number of type-$(t-j+1)$ teams. The system of equations will thus lead to scores that are symmetric (around ½) as well. This means that for all j, $v_j + v_{t-j+1} = 1$.[44] Knowing this, the system of equations simplifies to

$$
\begin{aligned}
v_1 &= \alpha + (1-\alpha)v_2 \\
v_2 &= \alpha\frac{(g-1)}{g} + (1-\alpha)\left[\frac{1}{g}v_1 + \frac{g-1}{g}v_3\right] > \alpha + (1-\alpha)v_3 \\
&\vdots \\
v_{m-1} &= \alpha\frac{(g-1)}{g} + (1-\alpha)\left[\frac{1}{g}v_{m-2} + \frac{g-1}{g}v_m\right] > \alpha + (1-\alpha)v_m \\
v_m &= \frac{1}{2}
\end{aligned}
$$

[44] This can be easily proven by showing that if $v_{m+s+1} + v_{m-(s+1)} = 1$ and $v_{m+s-1} + v_{m-(s-1)} = 1$ then $v_{m+s} + v_{m-s} = 1$.

After successive replacing we obtain:

$$\begin{aligned} v_1 &> (1-\alpha)^{m-1}\frac{1}{2} + \frac{\alpha}{g} + \alpha\frac{(g-1)}{g}\sum_{k=0}^{m-2}(1-\alpha)^k \\ &= (1-\alpha)^{m-1}\frac{1}{2} + \frac{\alpha}{g} + \frac{(g-1)}{g}\left[1-(1-\alpha)^{m-1}\right] \\ &= (1-\alpha)^{(t-1)/2}\frac{1}{2} + \frac{\alpha}{g} + \frac{(g-1)}{g}\left[1-(1-\alpha)^{(t-1)/2}\right] \end{aligned}$$

Online Appendix A - Extended Introduction

Why win percentage is the standard in complete tournaments

The objective of this extended introduction is to expand on the reasons why the win percentage method is the golden standard when the main objective is to declare a performance-based final ranking of teams in a complete tournament.

1. Rewards and punishment methods make sense

The nature of a competition between two agents is such that as a result of this competition, one of the agents shows superiority over the other. This is called a victory. The natural consequence is that the victor be rewarded and the loser punished. So what other way than having a system of rewards and punishments would make sense? I certainly cannot think of anything and I haven't seen anything ever written that cannot be thought of as a system of rewards and punishments, irrespective of whether the tournament is complete or not.

2. The focus should be on win/loss outcomes exclusively

One of the most alluring aspects of competitions is the excitement that comes from the possibility that the weaker team may beat the stronger team. Weak teams tend to edge out their stronger rivals whereas strong teams tend to crush their weaker foes. By focusing only on who wins and who loses we are, in a sense, being respectful of this uneven relation, rendering both the convincing victory and the lucky one as equivalent. This idea is at the heart of why competitions (where a match is between two agents) all over the world, with incredibly few exceptions, keep their focus on the win/loss outcome and only use other information as a tie-breaker at the end of a tournament.

3. Additivity makes sense

Recall that the idea of additive rewards and punishments does not mean that the method must be linear. It just means that whatever (possibly non-linear) reward that is assigned for winning or punishment assigned for losing

accumulates additively for the team over the games it plays. Certainly there are non-linear ways to accumulate rewards and punishments. The question then becomes, would this be an improvement over a simple summation in a complete tournament? Multiplicative rewards and punishments can be log-linearized, so they are not any different from additive rewards and punishments. Other more sophisticated methods would require a very careful examination of their properties, as the non-linear accumulation could easily result in unintended biases.

4. Simplicity is key

The win percentage method cannot be beat on the simplicity metric. Thus, if we are to favor a different scoring method we would have to be able to at least come up with an example of a complete tournament where we have a solid case to make against declaring the team with the most wins the champion (or for that matter, the one with the second-most wins the runner-up, etc.). For example, maybe the runner-up beat a lot of very good teams and the one with the most wins beat mainly bad teams. This could be grounds for declaring the runner-up the champion. The problem with this argument is that for this example to hold true then it must be the case that the runner-up lost to many bad teams and at least the first place team lost to a bunch of good teams. In other words, for every *good win* by the runner up relative to the first place team there is always a counterbalancing relative *bad loss*. Otherwise the runner-up would have finished first in wins. So the argument for declaring the runner-up the champion would have to come from some sense of superiority of *what (good) it means to win* over *what (bad) it means to lose*. But a win and a loss are just two sides of the same coin, that is, the only possible outcomes of a binary event.[45] So a mechanism that is not symmetric in its treatment of wins and losses would be either a very silly mechanism or respond to a completely different objective.[46]

So the short answer is: The win percentage method takes care of the main issues regarding the creation of a performance-based ranking for complete tournaments in the simplest possible way.

Online Appendix B - Related Literature

What does not work for incomplete tournaments and why

[45] Even adding the possibility of a draw does not affect this symmetry because a draw simply represents the absence of a win and a loss. This is why it is most common to consider it as half a win and half a loss.

[46] The case in point is when FIFA encouraged football associations to switch from 2 points per win to 3 points per win, breaking the symmetry between wins, losses and draws. The argument was never that of accuracy or fairness; it was that this change would result in more offensive-minded games that would be more entertaining to fans.

The objective of this section is to explain why certain types of scoring methods that are widely used by the literature, for what is essentially a different purpose, fail to do an adequate job of scoring teams in incomplete tournaments.

1. Scoring methods that seek to tease out win probabilities for matches

There are literally hundreds of papers and manuscripts that are motivated by the pursuit of determining the odds that a given team beats another in a given match. Regardless of how elaborate the models are, ultimately they produce a probability p_{ij} that team i beats team j. It is then very easy to assign scores for each team in a way that is consistent with these probabilities. In principle, this can be used to produce the final ranking of teams. However, this is not a desirable avenue for both practical and fundamental reasons: From a practical perspective, the nature of the objective (assessing correct probabilities) lends itself to using as much information as possible, well beyond wins and losses. Thus, most of these models simply cannot be applied to the objective stated here. For example: It would make sense to use not just the information on who won or lost a game but also how many points each team scored in that game. Over multiple games, this would provide objective information on which team is stronger. This would be very useful to determine whether a team like Leicester City, despite winning more games than, say Manchester City, should objectively be favored to lose in a game between them. But, more fundamentally, this would not help in any way to tell us whether Leicester City should end up at the top of the ranking or not because the objective is not to declare a team as *most likely to win its matches*. The objective is to declare a *champion* based only on wins, losses and draws. In fact, in its incredible 2016 season, Leicester City scored less goals than Manchester City and had a lower goal difference than Tottenham. Betting odds consistently favored teams like Manchester City and Tottenham in almost all of their games and consistently favored the opponent of Leicester City in most of its games. No reasonable association football fan thought that Leicester City was *stronger* than Manchester City, yet no reasonable association football fan thought Leicester City was not deserving of being crowned champion.[47]

2. Perron-Frobenius-based eigenvector scoring methods:

Eigenvector methods are elegant. They can also be intuitively appealing.[48] Ultimately the scoring method is reduced to solving the system of equations $\mathbf{Av} = \lambda\mathbf{v}$, where the solution $\mathbf{v}$ is then used as the scores vector. Ignoring the fact that advocating for such methods puts the cart before the horse, there are

[47] I would even venture to say that anybody who sees a contradiction in this last statement doesn't really understand sports or competitions in general.

[48] For a very good summary see Keener (1993).

more practical problems with these scoring methods: In selecting the matrix $\mathbf{A}$, one must make sure that the matrix is irreducible or else the solution $\mathbf{v}$ will have zeros. Not zeros in the sense of a score that happens to be zero (think for example of a scoring method that produces scores between -1 and 1 where a zero is one of many possible scores). Zeros in the sense of an absorbing value that deprives the scoring method of the sufficient nuance. For example: If $\mathbf{A}$ is such that $a_{ij} = w_{ij}/n_i$, then our score v_i for team i has the interesting property that it is proportional to the average score of the teams that i beats.[49] However, this results in all winless teams receiving a score of zero because every row of $\mathbf{A}$ that represents a winless team will have only zeros in it (hence $\mathbf{A}$ is reducible). This is undesirable because in an incomplete tournament, the opposition faced by two winless teams will have been different, so a scoring method should have sufficient nuance built in so that it can distinguish between a team that lost to many good teams from one that lost to many bad teams. This eigenvector method lumps these two teams together with the absorbing score (of zero in the example). And this problem does not end with these two bad teams. One can easily imagine a case where the two top teams had similar schedules with one of them playing the first bad team and the other one the second. This scoring method would not have enough nuance built in to determine that the first team should be rewarded more for its victory over the not-so-bad winless team than the second one's victory over the really-bad winless team. So the lack of nuance trickles upward, biasing the scores of all teams. A similar argument can be made for the case of unbeaten teams (in this case their respective columns in $\mathbf{A}$ would be zeros).

A fix to the absorbing score problem typically involves an ad-hoc twist of the original scoring method. Of course, the problem with doing this is that the properties that made the original method appealing in the first place no longer hold once the twist is introduced.

3. Centrality Measures:

Centrality measures seek to determine the connectedness of a node within a network, where a node is typically associated with a location and a link between the nodes typically represents the existence of a path (in the colloquial meaning of this word) between them. But this description is not just geographical: Good examples are web-pages and their links or articles and their citations. So, a natural idea is to apply this to teams and their victories and use the centrality measure as a score for the teams. But the flaw of this approach is apparent:

[49] This is known as Katz prestige. For more, see Jackson (2008).

Connectedness and *importance* (or value, or prestige) are essentially very different because at a fundamental level, a path that connects two nodes (even if it is a one way path) is a positive signal for both nodes, whereas a victory is decidedly good for one node and decidedly bad for the other node. In other words, we would expect a good centrality measure to reward both locations with a higher *connectedness* score as a result of the existence of a path between the two but reward the victor and punish the loser as a result of a game played between two teams. So, what makes a good centrality measure is likely to produce a bad measure of prestige. Take for example *betweenness centrality:*[50] It rewards a node for being in the middle of many paths (in the graph-theory meaning of the word) between other nodes. While this clearly indicates locational connectedness, from the point of view of a tournament all it means is that this team has beaten and was beaten by many teams. This does not indicate prestige. Rather, it indicates mediocrity.

So the short answer here is that trying to force-fit a mechanism designed to address one issue (win probability, mathematical elegance, connectedness) in order to address a different one (a fair final ranking of teams) is doomed to fail.

[50] See Jackson (2008).