Revisiting the Rankability of Unweighted Data from Pairwise Comparisons

PAUL ANDERSON, TIM CHARTIER, AMY LANGVILLE, and KATHRYN PEDINGS-BEHLING

Additional Key Words and Phrases: rankability, linear ordering problem

1 LOP AND HILLSIDE COUNT FOR UNWEIGHTED DATA

We designed the LOP and Hillside rankability methods for weighted matrices, yet they can also be used for unweighted matrices. Thus, LOP and Hillside provide alternatives to the method of Anderson et al. for unweighted graphs [1]. These three methods differ in their definition of k, the distance from perfection. The method of Anderson et al. defines k as the number of link additions and deletions required to transform the dominance matrix $\underset{\sim}{D}$ into a reordering of strictly upper triangular form, whereas the Hillside method defines k as the number of violations of the hillside constraints regarding ascending rows and descending columns. For unweighted data, Hillside Count finds a reordering that transforms the dominance matrix $\underset{\sim}{D}$ into a form that is as close to hillside upper triangular form as possible and then counts hillside violations from this as k. The following example demonstrates the differences between the Anderson et al. method and the Hillside unweighted method. We applied both methods to the *unweighted data* of the 1995-2012 seasons of the Big East conference of NCAA college football. Table 1 shows that k values of these two rankability methods are correlated (Pearson $r = 0.909$).

Table 1. Comparing rankability methods for unweighted data: Anderson et al. [1] vs. Hillside Count for 1995-2012 seasons of the Big East conference of college football.

	Anderson k, p	Hillside Count k, p
1995	2, 1	14, 4
1996	2, 3	6, 6
1997	8, 48	12, 4
1998	4, 1	28, 12
1999	4,1	28, 4
2000	2, 1	10, 4
2001	2, 1	10, 4
2002	2, 1	10, 4
2003	4,1	22, 4
2004	6, 1	40, 48
2005	4, 1	24, 12
2006	8, 4	36, 8
2007	12, 7	72, 24
2008	6, 3	32, 12
2009	4, 1	28, 24
2010	8, 3	60, 12
2011	8, 3	52, 24
2012	8, 1	52, 48

Since there is already a rankability method for unweighted data, what is to be gained by using the Hillside method for unweighted data? The 1999 and 2003 seasons show an advantage of the Hillside method. These two years have the same Anderson et al. rankability values ($k = 4$ and $p = 1$), yet the Hillside Count values differ ($k = 28$ and $p = 4$ for year 1999 and $k = 22$ and $p = 4$ for 2003). How is the Hillside Count method differentiating between these two years? Compare the 1999 and 2003 $\underset{\sim}{D}(\underset{\sim}{r}, \underset{\sim}{r})$ matrices below, which are dominance matrices symmetrically reordered according to optimal ranking $\underset{\sim}{r}$ given by the Hillside method.

Authors' address: Paul Anderson; Tim Chartier; Amy Langville; Kathryn Pedings-Behling.

$$\underset{\sim}{D}_{1999}(\underset{\sim}{r},\underset{\sim}{r}) = \begin{array}{c} \\ 7 \\ 2 \\ 1 \\ 5 \\ 8 \\ 3 \\ 6 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccccc} 7 & 2 & 1 & 5 & 8 & 3 & 6 & 4 \end{array} \\ \left(\begin{array}{cccccccc} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right) \end{array} \text{ and } \underset{\sim}{D}_{2003}(\underset{\sim}{r},\underset{\sim}{r}) = \begin{array}{c} \\ 8 \\ 2 \\ 3 \\ 7 \\ 1 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccccc} 8 & 2 & 3 & 7 & 1 & 4 & 5 & 6 \end{array} \\ \left(\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \end{array}.$$

The entries contributing to hillside violations are highlighted in red. Year 1999 has just two nonzeroes in its lower triangular, while year 2003 has four. Yet though year 1999 has fewer nonzeroes in the lower triangle than year 2003, it has more hillside violations, resulting in a slightly worse rankability score for k (28 vs. 22). This occurs because nonzeroes farther from the diagonal contribute more hillside violations than nonzeroes closer to the diagonal. In other words, big upsets (i.e., type 1 violations in the lower triangular that are far from the diagonal) naturally cost more than mild upsets (i.e., type 1 violations in the lower triangular that are near the diagonal). In this example, the Hillside Count method has determined that year 1999's two big upsets (the penultimate team beating the third place team and the last place team beating the fourth place team) are worse than year 2003's four mild upsets between neighboring teams (2^{nd} place over 1^{st} place, 4^{th} over 2^{nd}, 5^{th} over 4^{th}, and 7^{th} over 5^{th}). *Thus, the Hillside method is preferred over the method of Anderson et al. when the built-in accounting of rank violations by the severity of the violation is important.*

For unweighted data, another advantage of the Hillside method over the method of Anderson et al. is the simplicity, elegance, and history of the Hillside Model.

2 REVISED METHOD TO FIND p AND P FOR ANDERSON ET AL

A second rankability idea from this paper on weighted data that can be applied to unweighted data concerns the p half of the two rankability pieces k and p. As a result of Section 1, we now have three choices for rankability methods for unweighted data: the original Anderson et al. method, the LOP method, and the Hillside Count method. As mentioned above, these methods measure slightly different aspects of rankability. Suppose that a practitioner has some modeling reasons for preferring the method of Anderson et al. for an unweighted application. The most expensive part of the Anderson et al. rankability measure is the pruning tree for finding p. In this section, we replace that pruning tree with the more efficient ideas for finding and summarizing p and P. In order to do this, we must replace the original Anderson et al. original model with the alternative model, Model (1), shown below and first presented in [1].

$$\begin{array}{rcll} \max \sum_{i \neq j} d_{ij} x_{ij} & & & \quad (1) \\ x_{ij} + x_{ji} & = & 1 \quad \forall\, i < j & \text{(anti-symmetry)} \\ x_{ij} + x_{jk} + x_{ki} & \leq & 2 \quad \forall\, j \neq i, k \neq j, k \neq i & \text{(transitivity)} \\ x_{ij} & \in & \{0, 1\} \quad \forall\, i \neq j & \text{(binary)} \end{array}$$

Notice that the constraints of the LP-relaxed version of this alternative Model (1) are exactly the same classic LOP constraints that form the LOP polytope [3] and, thus, are exactly the same constraints and polytope for the Hillside Model. In other words, the LP LOP polytope, the LP Hillside rankability polytope, and the LP unweighted rankability polytope are identical. Only the objective functions differ. This means that theorems similar to those for weighted rankability models can be proven for this unweighted rankability Model (1) above. Namely, we have the following results.

Theorem 2.1. *Every ranking of an unweighted rankability problem (Model* (1)*) corresponds to a binary extreme point of the LP unweighted rankability polytope.*

Proof. Every ranking $\underset{\sim}{r}$ has a corresponding binary strictly upper triangular matrix $\underset{\sim}{X}(\underset{\sim}{r}, \underset{\sim}{r})$ which denotes $\underset{\sim}{X}$ after it has been symmetrically reordered according to $\underset{\sim}{r}$. The matrix $\underset{\sim}{X}$ is binary and clearly feasible since anti-symmetry and transitivity are easy to verify from the upper triangular form of $\underset{\sim}{X}(\underset{\sim}{r}, \underset{\sim}{r})$. It remains to show that $\underset{\sim}{X}$ is an extreme point, i.e., that $\underset{\sim}{X}$ cannot be written as a convex combination of other extreme points. We do this by contradiction. Suppose that there exists a scalar $0 < \alpha < 1$ and, without loss of generality, exactly two binary feasible matrices $\underset{\sim}{Y} \neq \underset{\sim}{Z}$ such that $\underset{\sim}{X} = \alpha \underset{\sim}{Y} + (1 - \alpha)\underset{\sim}{Z}$. Since $\underset{\sim}{Y} \neq \underset{\sim}{Z}$, there exists at least one element, say (i, j) such that $y_{ij} \neq z_{ij}$. Suppose, without loss of generality, that $y_{ij} = 1$ and $z_{ij} = 0$. Then $x_{ij} = \alpha y_{ij} + (1 - \alpha)z_{ij} = \alpha$, which means that $\underset{\sim}{X}$ is fractional, which contradicts the statement that $\underset{\sim}{X}$ is binary. Therefore, the assumption that $\underset{\sim}{X}$ is a convex combination of $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ is false and rather it is that $\underset{\sim}{X}$ is an extreme point. □

The corollary below follows from Theorem 2.1.

Corollary 2.2. *Every* optimal *ranking of an unweighted rankability problem of Model* (1) *corresponds to a binary extreme point on the* optimal face *of the LP unweighted rankability polytope.*

When the LP relaxation of the interior point solver applied to Model (1) terminates, there are two options for the optimal objective value δ^* (integer and non-integer) and two options for the optimal solution matrix $\underset{\sim}{Z}^*$ (binary and fractional) creating the following four outcomes.

0. δ^* is non-integer and $\underset{\sim}{Z}^*$ is binary.
1. δ^* is integer and $\underset{\sim}{Z}^*$ is binary.
2. δ^* is integer and $\underset{\sim}{Z}^*$ is fractional.
3. δ^* is non-integer and $\underset{\sim}{Z}^*$ is fractional.

Case 0 is actually not possible and therefore not an outcome because since $\underset{\sim}{D}$ being binary is integer and $\underset{\sim}{Z}^*$ is binary, then the objective value $\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\, z_{ij}^*$ must be integer. Case 1 means that $\rho = 1$, there is a unique optimal solution, and the LP solution is optimal for the IP. Case 2 is the most interesting to us and we will return to it with Theorem 2.3 below to build the set P of all optimal solutions for Model (1). Case 3 means that the LP solution is not optimal for the IP. Our experiments show that Case 3, though possible, is very unlikely. This is also supported by Anderson et al. [1] and Reinelt et al. [2, 3].

Theorem 2.3. *If the Interior Point solver of the LP relaxed unweighted rankability problem of Model* (1) *ends in Case 2 (δ^* is integer and $\underset{\sim}{X}^*$ is fractional) and the LP optimal face is the IP optimal face, then*

(1) *δ^* is the optimal objective value for the integer program,*
(2) *$\underset{\sim}{X}^*$ is on the interior of the optimal face (i.e., the convex hull of all optimal solutions) of the integer program, and*
(3) *fractional entry (i, j) in $\underset{\sim}{X}^*$ means that there exists at least one optimal ranking in P with $x_{ij}^* = 1$ (thus, $i > j$) and at least one with $x_{ij}^* = 0$ (thus, $i < j$).*

As a result, this means that the ideas can also be used for the unweighted case. That is, when an interior point solver applied to an unweighted rankability Model (1) concludes with an integer k^* and a fractional optimal solution $\underline{X}^*$, the reordered $\underline{X}^*(\underline{r}, \underline{r})$ can be analyzed to efficiently summarize P, the set of all optimal rankings.

Example 4 shows a demonstration applied to the *unweighted data* for the 2008 Big East men's college football season.

Example 4. The 2008 season has an integer $\delta^* = 6$ and the following optimal fractional $\underline{Z}^*$ matrix shown in Figure 1. The 3×3 fractional submatrix creates $3! = 6$ subrankings of the items 4, 8, and 5 that are evaluated for optimality. Of

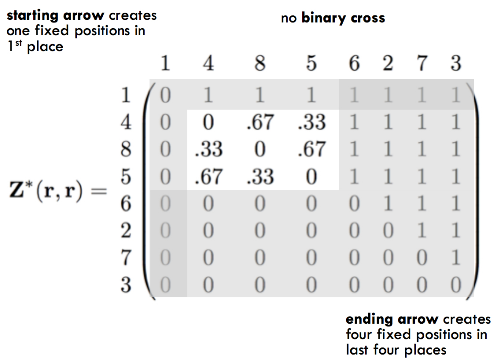

Fig. 1. Model can also be applied to unweighted data. The interior point solution of unweighted Example 4 is a fractional matrix $\underline{Z}^*(\underline{r}, \underline{r})$ with a starting arrow, ending arrow, and fractional submatrix.

these 6, only 3 are indeed optimal, meaning $p = 3$, and $P = [1\ 8\ 5\ 4\ 6\ 2\ 7\ 3], [1\ 5\ 4\ 8\ 6\ 2\ 7\ 3], [1\ 4\ 8\ 5\ 6\ 2\ 7\ 3]$.

$$\underline{Z}^*(\underline{r}, \underline{r}) = \begin{array}{c} \\ 1 \\ 4 \\ 8 \\ 5 \\ 6 \\ 2 \\ 7 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 4 & 8 & 5 & 6 & 2 & 7 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & .67 & .33 & 1 & 1 & 1 & 1 \\ 0 & .33 & 0 & .67 & 1 & 1 & 1 & 1 \\ 0 & .67 & .33 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

REFERENCES

[1] Paul Anderson, Timothy Chartier, and Amy Langville. 2019. The Rankability of Data. *SIAM Journal on the Mathematics of Data Science* 1 (2019), 121–143.

[2] Rafael Marti and Gerhard Reinelt. 2011. *The Linear Ordering Problem: exact and heuristic methods in combinatorial optimization.* AMS.

[3] Gerhard Reinelt. 1985. *The Linear Ordering Problem: Algorithms and Applications.* Heldermann Verlag.