

# AN EVALUATION OF RANKING METHODS FOR MULTIPLE INCOMPLETE ROUND-ROBIN TOURNAMENTS

**P. Chang**, Department of Information Systems, New Jersey Institute of Technology,  
Newark, NJ 07102, peishih.chang@njit.edu, 973-596-5422

**D. Mendonça**, Department of Information Systems, New Jersey Institute of  
Technology, Newark, NJ 07102, mendonca@njit.edu, 973-596-5212

**X. Yao**, Department of Information Systems, New Jersey Institute of Technology,  
Newark, NJ 07102, xy3@njit.edu, 973-596-5655

**M. Raghavachari**, Department of Decision Sciences and Engineering Systems,  
Rensselaer Polytechnic Institute, Troy, NY 12180, ragavm@rpi.edu, 518-276-2962

## ABSTRACT

We investigate the ability of various methods to reveal the true ranking of players competing in multiple incomplete round-robin tournaments, which arise when not all tournament competitions are held. Statistical and graphical analyses in this simulation-based study reveal that a method based on players' average winning percentages is usually best. **Keywords:** multicriteria decision making, pair-wise comparisons, simulation

## 1 INTRODUCTION

The problem of how to evaluate a number of options along a variety of criteria—even when not all objects can be compared along all the criteria—arises frequently in theory and practice. In home buying, for example, a prospective buyer may evaluate homes according to the number of bedrooms and square footage. Yet if some houses have pools and some do not, developing a final ranking of the houses is more difficult. In a multiple round-robin tournament, players compete in a number of tournaments, with each tournament corresponding to a single criterion. A multiple tournament is said to be incomplete when not all players compete in all tournaments but all competitions within a particular tournament are held.

The central problem of this paper is to evaluate the efficacy of various ranking methods in revealing the true player ranking from the results of a multiple incomplete round-robin tournament. This paper extends prior work on evaluating the efficacy of ranking methods for such tournaments in three principal ways. First, it explores—via simulation-based experimentation—how variation in tournament structure and players' strengths can impact ranking method efficacy. Second, it introduces techniques for analyzing the experimental results that involve both statistical and graphical methods, thus supporting the ease with which the results of similar simulations can be interpreted. Finally, it develops heuristics to guide selection of the appropriate ranking method when some foreknowledge about players and/or tournament structure is available.

## 2 BACKGROUND

A model of pairwise comparisons attempts to explain the capability of one player to defeat another in a single contest. Assume that the strength of each player  $i$ ,  $i = 1, \dots, n$ , is modeled by a random variable  $X_i$  having mean  $V_i$ . The ranking of the mean strengths of the players may then be taken as the true ranking.

Two well-known models for player strength are (i) that of Mosteller and Thurstone [7] and (ii) that of Bradley and Terry [2, 3] and Mallows [11] (also attributed to Zermelo) [7]. If  $\pi_{ij}$  is the probability that player  $i$  defeats player  $j$ , then under the Mosteller-Thurstone model (denoted MT), each  $V_i$  is the mean of

normally distributed random variable  $X_i$  with variance  $\sigma^2$  and with  $X_i$  and  $X_j$  assumed to be independent. Therefore,

$$\pi_{ij} = \Phi\left(\frac{V_i - V_j}{\sigma\sqrt{2}}\right), \quad (1)$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. Under the second model (denoted BTM),

$$\pi_{ij} = \frac{V_i}{V_i + V_j}. \quad (2)$$

Several distributions of the  $X_i$ 's may yield BTM and it may be verified that equation (2) follows if each  $X_i$  is exponentially distributed with mean  $V_i$ . [7].

In the usual form of the single round-robin tournament, each one of  $n$  players plays the remaining  $(n-1)$  players exactly once. Results of a tournament between  $n$  players can be summarized in an  $n$ -by- $n$  incidence (i.e., tournament) matrix  $\mathbf{A}$  whose  $(i, j)^{\text{th}}$  element is 1 if player  $i$  defeats player  $j$  and is 0 otherwise (it is assumed that no draws are allowed). Various versions of the single round-robin tournament are possible [15], some of which are more efficacious than others in revealing the best player [8]. In a multiple round-robin tournament, several players compete in a schedule of  $T \geq 2$  tournaments, some of which may be incomplete, in that they are contested by  $n_t \leq n$  players, thus yielding  $n_t(n_t-1)/2$  outcomes for tournament  $t$ .

## 2.1 Ranking Methods for Multiple Tournaments

Methods for determining rankings in tournaments are a work on pairwise comparisons [7], tournaments generally [15], ranking methods [19] and models of rank data [6, 12]. Ranking methods for multiple complete tournament structures have also been presented [9, 13]. A ranking method for multiple incomplete tournaments allows the comparison of the  $n$  players when not all players participate in all tournaments. Such methods might be based upon numerical optimization [5, 14], the percentages of matches won by the players [10, 14], or properties of higher powers of the tournament matrices [14, 20]. This paper offers a further evaluation of the latter two types, since both have exhibited attractive qualities in prior research [14].

Before discussing these methods, some additional notation (adapted from [5]) must be introduced. Let  $K_{i_0}$  be the set of tournaments in which player  $i_0$  competes and let  $a_{i_0}^t(m)$  be the row sum for player  $i_0$  in the  $m^{\text{th}}$  power of the incidence matrix of tournament  $t$ . Let  $R_{i_0}$  be the score assigned to player  $i_0$  by a ranking method. The notation  $i_0$  is necessary since, for example, the second player in tournament may not be "Player 2." Finally, let  $|B|$  denote the cardinality of any set  $B$ . The methods are as follows.

*Kendall Modified (KM)*. A method proposed by Kendall [10] enables the determination of the final rank of a player in a single round-robin tournament, computed by calculating that player's percentage of matches won out of total matches played. Player  $i$ 's relative strength,  $R_i$ , is determined as:

$$R_i = \frac{a_i}{(n-1)}, \quad (3)$$

where  $a_i$  is the sum of the elements in the  $i^{\text{th}}$  row of  $\mathbf{A}$ ; that is, the number of matches won by player  $i$ . This method was later extended to the multiple tournaments context [14] as follows:

$$R_{i_0} = \sum_{t \in K_{i_0}} \frac{a_{i_0}^t(1)}{|K_{i_0}|}, \quad (4)$$

where  $a_{i_0}^t(1)$  is the row sum for player  $i$  in the incidence matrix for tournament  $t$ . Equation (4) therefore represents the average winning percentage of player  $i_0$  taken across the tournaments in which that player competes. We will refer to this method as KM for Kendall Modified.

*Eigenvector 1 (Eig1)*. Using the eigenvector of the incidence matrix from a single round-robin tournament as the basis of a ranking was first proposed by Wei [20]. Two such methods [14] are reviewed here. The first method, Eig1, is similar in construction to KM:

$$R_{i_0} = \sqrt{\frac{\sum_{t \in K_{i_0}} u_{i_0}^2(t)}{|K_{i_0}|}}, \quad (5)$$

with which is associated a vector

$$u^*(t) = \lim_{l \rightarrow \infty} \left( \frac{\mathbf{A}^t}{\lambda_t} \right)^l \times \mathbf{e}, \quad (6)$$

which is the non-normalized score vector for tournament  $t \in K_{i_0}$ , for which  $\lambda_t$  is the unique positive characteristic root having the largest absolute value for the incidence matrix of tournament  $t$  [14]. In equation (5),  $u_{i_0}^2(t)$  is the square of the elements of the normalized form of  $u^*(t)$  corresponding to player  $i_0$ . *Eig1* is therefore an average of the relative strengths across all tournaments in which  $i_0$  competes.

*Eigenvector 2 (Eig2)*. This method uses the distance from a vector whose elements are a player's relative strengths (i.e., the  $|K_{i_0}|$  values of  $u_{i_0}^2(t)$  where  $t \in K_{i_0}$ ) to some optimal vector [14]. In a schedule of three complete tournaments (i.e.,  $n_t = n$  for all  $t$ ), each of the three values of  $u_{i_0}^2(t)$  can at most be unity. So, in  $T$ -dimensional space, a vector of optimal scores can be represented as  $\mathbf{e}$ , the  $n \times 1$  vector of ones. Let the vector of player  $i_0$ 's relative strengths be represented by  $\mathbf{u}_{i_0}$  and let  $\mathbf{e}_{i_0}$  be a unit vector of dimensionality  $|K_{i_0}|$ . A player's relative strength may then be based on the head-to-head distance:

$$R_{i_0} = -\|\mathbf{u}_{i_0} - \mathbf{e}_{i_0}\|. \quad (7)$$

Two prior studies [5, 14] investigated the efficacy of ranking methods for a single schedule of multiple incomplete tournaments. The first [5] evaluated a number of optimization-based methods based on data envelopment analysis [4]. The second [14] evaluated two of the methods from the first study along with the three discussed in the previous section, using six different sets of players participating in the same tournament schedule as the first study. The results of this study revealed method KM to be the strongest overall performer, with Eig1 and Eig2 exhibiting some attractive properties (such as low variance in their correlations with the true ranking). Accordingly, the present paper extends this previous scope of inquiry by (i) considering how KM, Eig1 and Eig2 perform for a greater variety of tournament structures and player strengths and by (ii) presenting additional tools to support analysis of the results.

## 2.2 Visualization Techniques for Ranking Data

Visualizations of large data sets may improve understanding and exploration of the underlying phenomena that gave rise to the data sets. [1] A decision maker might want to know how sensitive the results might be to variation either in the schedules of the tournaments or the relative strengths of the players. Data visualizations can improve the comprehension of multidimensional data by representing three- and higher-dimensional data using color and/or texture. [18] So, when the number of players and/or tournaments grows large, visualizations may support rapid identification of (clusters of) top-ranked players. [17]

### 3 SIMULATION METHODOLOGY

A simulation-based methodology is used to test how well KM, Eig1 and Eig2 reproduce the true ranking of players competing in multiple incomplete round-robin tournaments. The methodology is run for the Bradley-Terry/Mallows (BTM) and Mosteller-Thurstone (MT) probability models and consists of the three stages of initialization, simulation, and analysis and visualization.

*Initialization.* A set of players having known mean strength is established. Six different sets—or profiles—of players are used, as shown in TABLE 2. Both relative and absolute mean player strengths are varied, allowing for a very full investigation of the relative performance of the ranking methods. For example, strengths are separated by one in profiles 2 and 5, by two in Profiles 3 and 6, and by 3 in Profile 4. When the difference is one or two, the Profiles are also of different magnitudes (e.g., Player 1 has strength of 1 in Profile 2, but strength of 7 in Profile 5).

One way to characterize multiple round robin tournaments is according to their degree of completeness, here denoted the density. The density is calculated by summing the number of players actually competing in each tournament and dividing by the product of the number of tournaments and the number of players in the profile. Table 1 shows a sample schedule. A shaded box indicates that a player participated in a particular tournament; an unshaded box indicates that a player did not participate in a particular tournament. The density for this sample schedule is  $17/(6 \times 4) = 0.71$ . To complete the initialization phase, a profile and a density are chosen from the set of all combinations of profiles and densities.

**TABLE 1: Sample Schedule**

Tournament	Player					
	1	2	3	4	5	6
1						
2						
3						
4						

**TABLE 2: Mean Player Strengths**

Profile	Player					
	1	2	3	4	5	6
1	3	3	3	3	3	3
2	1	2	3	4	5	6
3	2	4	6	8	10	12
4	3	6	9	12	15	18
5	7	8	9	10	11	12
6	8	10	12	14	16	18

*Simulation.* The players in a profile participate in multiple tournaments of various densities. The Simulate Strength process uses the governing probability model (i.e., either MT or BTM) to generate a sample of a player’s strength. In the current implementation, the variance in player strength under MT for all profiles is held constant at 3, which we regard as reasonably appropriate for the range of player strengths. However, note that the methodology can accommodate the case of unequal but known variances. Next, the Generate Schedule process produces a number of schedules of given density. For each schedule, 1000 incidence matrices are generated, with any ties resulting from the pairwise comparisons re-simulated until a winner emerges. To complete the simulation, the three methods are applied to the incidence matrices to produce the estimated rankings. The true ranking is generated by ranking the players according the mean value of their strength in the current profile.

*Analysis and Visualization.* The efficacy of the various ranking methods is evaluated both quantitatively and qualitatively across all profile and density combinations. The estimated ranking of the players is compared to the true ranking as determined from the mean strength, as follows. Let  $y_{jki}$  be the rank of player  $i_0$  according to method  $k$  for the  $j^{\text{th}}$  simulation, and let  $s_{i_0}$  be the true rank of player  $i_0$ . The statistic of interest is the well-known Spearman’s rank correlation coefficient [16], referred to in this paper as Spearman’s  $r$ . For a method  $k$ , the correlation measure for the  $j^{\text{th}}$  simulation is computed as

$$r_{jk} = 1 - \frac{6 \sum_{i_0=1}^n (y_{jk i_0} - s_{i_0})^2}{n^2 (n-1)}, \quad (8)$$

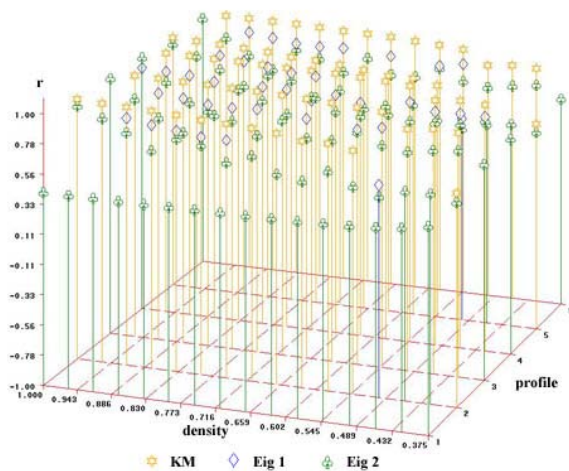
where  $r_{jk}$  is the Spearman's correlation between the rank orders obtained by method  $k$  and the true rank order for the  $j^{\text{th}}$  simulation run. Different methods are then compared with reference to the average Spearman's  $r$  over all the simulations. For a given profile of players a one-way analysis of variance (ANOVA) is performed to test equality of the average correlation coefficients among all three methods. If means are unequal, Tukey's multiple comparison procedure [16] is applied in order to cluster the methods into groups having statistically equal means, and then to rank the clusters according to their efficacy. Finally, two- and three-dimensional visualizations are used to represent the results in a concise way. The simulation is complete once all combinations of profiles and densities have been run.

The methodology is implemented as follows. Sixteen different densities were used, with ten schedules generated for all 16 density values except values 1, 0.417, and 0.375. Because at least two players must compete in a tournament, only one schedule is available when the density is one, and less than 10 schedules are available when the density is 0.417 or 0.375. In total, 138 schedules were generated. One thousand simulations were run for each combination of player profile and tournament schedule using MATLAB®6.5. All statistical tests and data visualizations were executed with SAS® 8e.

#### 4 RESULTS

*Mosteller-Thurstone.* The value of Spearman's  $r$  for each experimental condition is plotted in FIGURE 1 for each combination of density and profile under MT. The figure shows a rough topology of the values so that it is possible to identify the regions in which various methods are best. It may also be useful for decision making purposes to visualize the best (or second- or third-best) method and whether the differences between methods are statistically significant. FIGURE 2 provides these insights by plotting the results of Tukey's multiple comparison procedure ( $p$ -value  $\leq 0.0001$  for ANOVA and 0.05 for Tukey's test). If the symbol corresponding to a particular method is underlined, then that method is not significantly better than the one on the plane beneath it. FIGURE 2 depicts the best method for each density/profile combination. KM is the most efficacious method in Profiles 2 through 6, as was found in [14] for the single schedule case. In Profile 1, players are of equal strengths, so no method is particularly effective.

**FIGURE 1: 3-D Plot of Spearman's  $r$  (MT)**



**FIGURE 2: Best Method Plots (MT)**

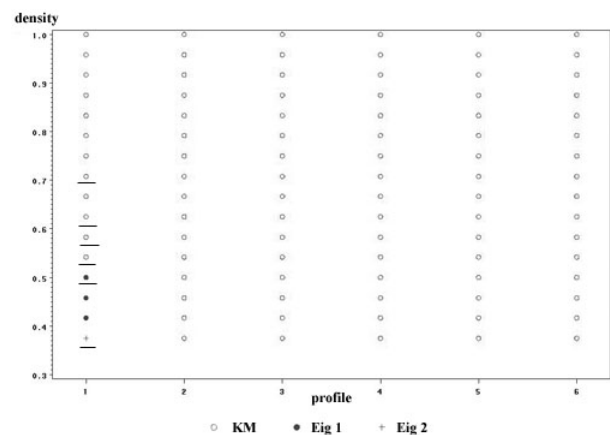


FIGURE 3: 3-D Plot of Spearman's  $r$  (BTM)

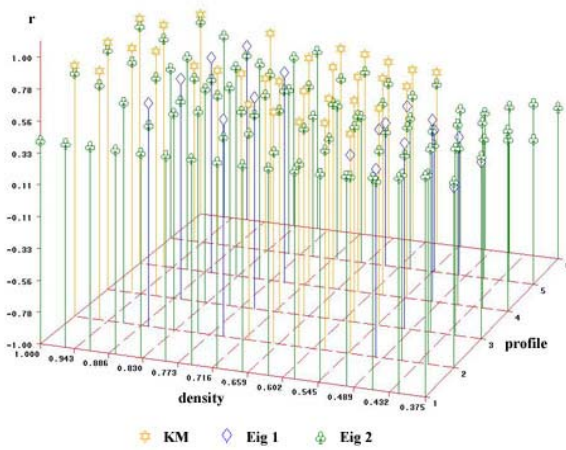
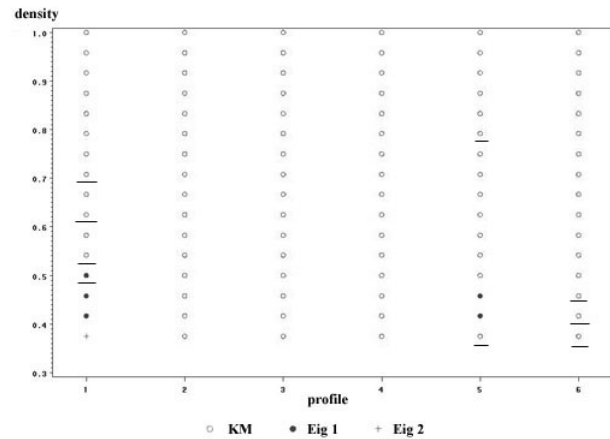


FIGURE 4: Best Method Plots (BTM)



*Bradley-Terry/Mallows*. The results under BTM are presented in Figures 3 and 4. As suggested by both figures, KM is the best method for Profiles 2 through 4. In Profiles 5 and 6, KM is also usually best, except for select instances, such as when density is low. Eig1 is usually second-best, except in cases of very low or very high density, when Eig2 is second-best. When density is near 0.6, Eig1 is sometimes no better than Eig2.

## 5 CONCLUSIONS

Under MT, the results suggest that profile and density do not influence which method is most likely to reproduce the true ranking. Eig1 is usually second-best, though there is some evidence that, when density is very low or very high, Eig2 is second-best. Profile does not seem to have an effect on the efficacy of the second-best method.

Under BTM, KM is best with some exceptions, but there does not seem to be either a profile or density effect. As under MT, Eig1 is usually second-best, except when density is very high or very low. Profile does not seem to impact the efficacy of the second-best method. It must be emphasized, however, that Spearman's  $r$  for all three methods was lower and more variable than under MT. With these caveats, KM may produce the ranking that is closest to the true ranking when conditions are similar to those described in this study. Overall, then, KM—which is a ranking method based on the percentage of matches won by a player—outperforms other methods for all but a few profile/density combinations. This conclusion can be reached—despite the large number of experimental conditions—by examining a series of visualizations developed for this study.

## 6 SELECTED REFERENCES

- [10] Kendall, M. G., "Further Contributions to the Theory of Paired Comparisons," *Biometrics*, vol. 11, pp. 43-62, 1955.
- [14] Mendonça, D. and Raghavachari, M., "Comparing the Efficacy of Ranking Methods for Multiple Round-robin Tournaments," *European Journal of Operational Research*, vol. 123, pp. 593-605, 2000.