A Judging System for Project Showcase: Rating and Ranking with Incomplete Information (DRAFT)

Jiangtao Gou and Shuyi Wu

June 10, 2020

Abstract

Mixed effects model is usually used to handle the situation where the collected data points are not truly independent. We develop a rating and ranking method based on mixed effects modeling, and compare it with various popular rating and ranking models in data science, including PageRank, HITS, Elo's method, etc. This proposed method can be used in a judging system where each qualified judge can only rate a portion of the participants or projects, and also in web page ranking, news recommendations, sports, etc. One advantage of using the mixed effects model over other methods like PageRank is that we can perform statistical inference naturally. An R library package for creating various judging systems for a research project showcase is made available on CRAN.

Keywords. Data science, Judging system, Mixed effects model, Sports ranking, Webpage ranking.

1 Background: High School Project Showcase

Each year the School of Education and Social Policy (SESP) at Northwestern University (NU) hosts a high school STEM[1] project showcase to engage the 9^{th}–12^{th} grade students from the Chicagoland area in project-based learning and creative problem solving to prepare students for their college-level education. High school students and their teachers and parents are invited to join Northwestern's Undergraduate Research Expo. Moreover, a key component of this event is a forty-five-minute poster presentation of high school student research projects. The poster presentations are judged onsite and award winners are chosen based on quality of research, oral and visual presentations. The volunteer poster judges are NU's PhD candidates who have participated in the Research Communication Training Program (RCTP) or the Reach for the Stars program (RftS) or both. The goals of RCTP are to improve graduate students' presentation and communication skills in order to form a connection with any audience. The RftS program provides the opportunity for NU's graduate students to serve as resident scientists in middle and high school classrooms. Therefore, all judges are capable to evaluate high school student research projects and provide constructive feedback.

In 2014, the first author of this article severed the high school project showcase as a poster judge out of eighteen. A total of thirty-three projects participated in this event, and the title and abstract of these projects were sent to the judges one week ahead of time, along with judging criteria and score sheet. A judge evaluated a project based on fourteen different components each with a rating from 0 to 5, so the perfect score is 70. All judges participated

[1]STEM is an acronym for "Science, Technology, Engineering and Mathematics".

in a one-hour judges training on the same day before the showcase to discuss the instructions of evaluation criteria and try to reach a similar grading scale. Due to the 45-minute time limit for the project showcase, it was not likely that a judge can evaluate more than four projects. During the judges' meeting before the showcase, 33 high school student projects were randomly assigned to 18 PhD candidate judges, where each project was planned to be evaluated by two judges, and each judge planned to evaluate three or four projects. The distribution of the number of evaluations for each project and that of the number of projects evaluated by each judge are summarized in Table 1. Most student projects were evaluated by at least two judges, except two projects which only received one score. This slight departure from the original plan was mainly because a few judges spent more time than expected with high school students in discussing their projects and the project showcase was only 45-minute long.

Table 1: Frequency tables of the number of evaluations for each project and the number of projects evaluated by each judge

Number of evaluations	1	2	3
Frequency	2	30	1

Number of projects evaluated	2	3	4	5
Frequency	2	6	7	3

Mean scores across all judges who evaluated the corresponding student projects were calculated, ranging from 33.0 to 64.5, and determined winners. Top ten students with highest mean scores received recognition at the award ceremony. In addition, since the gap between the fourth place and the fifth was relatively large, the four highest ranking students were named the first price winners and the next six were named the second price winners. During the award ceremony, a natural question arose in the first author's mind: is there a better way to rate and rank these high school student projects based on the current judging system?

This article is organized as follows. Section 1 tells a story of the first author who served as a poster judge in a high school student project show case. Section 2 brings in a mixed effects model solution to improve the scoring system. In Section 3, we extend the application of mixed effect models to sports and information science. Meanwhile, we apply the mathematical models rooted in sports and online business to the data set from the high school student project show case, and compare them with the mixed effect models via simulation in Section 4. Conclusions and discussions are provided in Section 5. An R library package `raincin` (ranking with incomplete information) to implement the ranking methods listed in this article is made available on CRAN.

2 Judging System: Mixed Effects Modeling

If each judge is able to visit each poster and discuss with the student during the project showcase, we can calculate a mean score over all judges or a robust average score with outlier excluded for each student, similar with the scoring systems in gymnastics and diving. However, it is common for a poster presentation competition to be scheduled for a limited time period, involve a large number participants, but not have enough well-qualified poster

judges. In addition, it is important that judges in the high school project showcase spend enough time with students in having constructive conversations to make students feel valued and recognized as researchers. Therefore, each judge can only evaluate a set of posters in such a situation, and the rating, ranking and inference have to be based on incomplete information.

As a PhD candidate in statistics at that time, the first author realized that the scoring method implemented in the high school student project showcase assumes the simplest type of an experiment which involves a single treatment (Tamhane, 2009). We write this standard model as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (1 \le i \le a, 1 \le j \le n_i), \tag{1}$$

where $a = 33$ denotes the number of projects, n_i denotes the number of judges who are randomly assigned to the i^{th} project, μ is the overall mean score, $\mu + \alpha_i$ are mean score for each project and are referred to as fixed treatment effects, and the ϵ_{ij} are independent and identically normal-distributed errors. Under the one-way analysis of variance (ANOVA) assumptions, the unbiased estimation of $\mu_i = \mu + \alpha_i$ is $\overline{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$, which is the average score across all judges who have evaluated the i^{th} student project. We compare the adjacent pairs of mean scores in the ranking list by the least significant difference (LSD) procedure (Fisher, 1935), and find the top three most significant gaps, which are between the 2^{nd} highest score and the 3^{rd} highest score with a p-value 0.578, between the 8^{th} and the 9^{th} with $p = 0.696$, and between the 4^{th} and the 5^{th} with $p = 0.760$. This explains the selection for awarding the first prize for top four students with highest mean scores.

The direct averaging method is simple and lucid. However, we may not ignore the effect of variability from judges to judges. Although the judge training before the student project showcase helps form a standard grading scale and reduce the variability, some judges may still tend to give higher scores than the others, since rating a poster presentation is not as simple as grading true or false questions. Therefore, a mixed-effects model that takes into account the judge variation can be applied as shown below.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (1 \le i \le a, 1 \le j \le b), \tag{2}$$

where Y_{ij} is the score of the i^{th} project given by the j^{th} judge, $\mu + \alpha_i$ measures the quality of the i^{th} student project as the fixed effect, $\beta_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$ is the random effect for each judge, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$ is a Gaussian error term. With a total of 65 observed scores from 33 posters and 18 judges, we get the maximum likelihood (ML) estimates of σ_B as 5.61 that is the variation between judges. To test if the variation between judges is significant, we compare the mixed effects model in equation (2) and the reduced model in equation (1), and find the random effect term is significant with $p < 0.001$ using the likelihood ratio test. Table 2 lists the top ten high school student projects and their ratings using the original linear model in equation (1) and the mixed effects model in equation (2), where the two lists agree with each other on the choices of top two, top four, top six and top eight projects. For the full rating list of 33 student projects, the correlation coefficients with 95% confidence intervals (CI) between these two lists are: (1) Pearson's r, 0.942 with a t CI $[0.886, 0.971]$, (2) Kendall's τ, 0.803 with a bootstrap CI $[0.680, 0.897]$, (3) Spearman's ρ, 0.931 with a jackknife CI $[0.865, 0.993]$.

Table 2: Award-winning high school student projects under the original judging system and using the mixed effects modeling

Original system model in (1)	Project ID	31	30	17	14	27	5	10	11	26	13
	Rating	64.5	62.5	57.5	56.5	54.0	53.5	53.0	52.5	49.0	48.5
Mixed effects model in (2)	Project ID	30	31	14	17	5	27	11	10	9	7
	Rating	63.1	63.0	57.5	56.9	55.6	55.3	54.8	53.0	51.6	49.3

3 Judging System: Rating and Ranking

Regression analysis is not the only method to improve the judging system described in Section 1. Rating and ranking methods used by data analysts in sports and online business can be applied to the project showcase judging system. In this section, we make use of six different types of methods, including (1) Google's PageRank algorithm (Brin and Page, 1998) and (2) the Hyperlink-Induced Topic Search (HITS) algorithm (Kleinberg, 1999) for ranking web pages, (3) the Perron-Frobenius-Keener (PFK) method (Keener, 1993), (4) the Massey rating method (Massey, 1997) and (5) the Colley matrix method (Colley, 2001) for ranking football teams, and (6) the Elo rating system (Elo, 1978) for rating chess players. To borrow the methods of sports or web pages ranking, we usually need to convert the project showcase data set from a judge-by-project setting to another setting of pairwise matchups between projects (Langville and Meyer, 2012). Specifically, assuming the j^{th} judge evaluated a total of m_j projects, these m_j scores from the same judge are converted into a total of $\binom{m_j}{2}$ matchups. For example, Judge 1 evaluated a total of three projects, which are Project 15, 16 and 22, and the corresponding scores from this judge were 41, 51 and 37. We can next imagine three games among these three projects: Project 15 lost to Project 16 a score of 41-51 but defeated Project 22 41-37, and the score of the game between Project 16 and 22 is 51-37.

With these imagined matchups among projects, we apply the ranking and rating methods listed above to the project showcase data set. (1) Google's PageRank lets projects vote with the number of points given up in the imagined games, and uses these votes to form a Markov graph (Langville and Meyer, 2006). A random surfer takes a random walk on this graph based on the transition probabilities. Projects' relative ratings are calculated as the stationary vector of the corresponding Markov chain, with a damping factor 0.85 to ensure the irreducibility of the stochastic matrix. The final scores are proportional to the relative ratings, where the mean score matches the direct sample average 46.4, and the standard deviation of the final scores matches the sample standard deviation of the raw scores 9.8. Therefore, it is possible that some scores are greater than the perfect score 70. (2) Kleinberg's HITS algorithm assigns a high hub rating to a discriminating judge, and assigns a high authority rating to a good project. A good project gets high ratings from discriminating judges, and a good judge give scores which match the quality of the projects (Liu, 2011; Langville and Meyer, 2012). Similarly, we match the mean authority scores of projects from HITS algorithm with the direct average score, and also the and the standard deviations. (3) the PFK method assumes the rating of a football team or a student project is proportional to its strength, and the unique eigenvector based on the Perron-Frobenius

theorem serves as a measure of relative strength. (4) Massey's method calculates a rating vector using the least square method to fit the margins of victory for the imagined matchups between projects. (5) Colley's method uses Laplace's rule of succession to calculate the winning percentages as the relative ratings for each project (Frey, 2007). (6) Elo's system treats projects as chess players, where the ratings are updated based on the matchups, and the reward to the project that wins the matchup depends on the difference between the ratings of the projects in this matchup. The parameters we used here come from the United States Chess Federation with a fixed K-factor 32 (Glickman and Doan, 2017).

Table 3: Top ten high school student projects using various rating and ranking methods

PageRank	Project ID	27	5	7	14	31	28	11	10	2	26
	Rating	73.1	63.3	60.1	59.1	56.3	55.5	54.8	53.2	50.7	50.0
HITS	Project ID	27	14	7	30	18	5	28	2	10	31
	Rating	81.6	68.1	61.0	57.8	53.4	52.7	50.5	50.1	50.1	49.1
PFK	Project ID	30	31	14	9	11	5	27	17	18	7
	Rating	66.2	62.8	60.4	59.4	59.1	58.2	58.0	55.5	55.1	53.1
Massey	Project ID	30	31	9	11	14	5	27	17	18	10
	Rating	69.2	67.0	60.3	59.5	59.0	57.6	56.1	55.2	53.4	53.2
Colley	Project ID	31	14	30	11	5	17	27	10	18	26
	Rating	63.7	63.7	61.3	61.2	59.2	57.9	57.5	52.3	51.7	49.7
Elo	Project ID	14	31	30	11	17	27	5	18	10	9
	Rating	63.9	63.2	62.8	59.6	59.0	57.8	57.3	55.1	52.3	52.3

Table 3 lists the top ten projects produced by the six methods. Considering the two lists from Table 2 using linear models, only four projects are included in all eight lists, which are Project 5, 14, 27 and 31. For the top-ten lists, Google's PageRank provides a quite different ranking list compared with other methods. For example, Project 30 is listed in the first place by the mixed effects model, PFK and Massey's method, in the second place by the original linear model, in the third place by Colley's and Elo's methods, and in the fourth place by the HITS algorithm. However, the PageRank lists Project 30 in the twentieth place and even not in the first half of the ranking list. For comparisons of the whole lists, we compute the correlation coefficients for the eight ranking lists in Table 4, where Pearson's r's are included in the upper triangular part and Kendall's τ's are presented in the lower triangular part. The mixed effects model (ME) provides the most similar ranking list of the list from the original system (Orig). The PageRank (PR) and HITS algorithms largely disagree with the other six methods, and these two webpage-ranking techniques also provide quite different ranking lists on the high school project showcase data set.

In the next section, we compare the eight rating and ranking methods in Section 2 and 3 via simulation studies. Similar approach can be used to compare the mixed effects model with many other ranking methods (Stern, 1991; Frey, 2005; Zhou and Lange, 2009; Langville and Meyer, 2012; Alvo and Yu, 2014).

Table 4: Correlation coefficients for pairwise comparisons among eight ranking methods: Pearson's r (upper triangular part) and Kendall's τ (lower triangular part)

Correlation	Orig	ME	PR	HITS	PFK	Massey	Colley	Elo
Orig	1	0.942	0.528	0.538	0.857	0.850	0.840	0.795
ME	0.803	1	0.568	0.576	0.967	0.962	0.907	0.887
PR	0.371	0.371	1	0.756	0.571	0.532	0.597	0.580
HITS	0.459	0.496	0.527	1	0.561	0.497	0.521	0.517
PFK	0.669	0.818	0.371	0.428	1	0.981	0.895	0.899
Massey	0.608	0.758	0.386	0.367	0.894	1	0.907	0.921
Colley	0.654	0.727	0.409	0.375	0.735	0.795	1	0.973
Elo	0.577	0.667	0.402	0.322	0.682	0.750	0.856	1

4 Judging System: Comparisons

We perform comparisons using simulated data sets with known underlying true project scores. The prespecified parameters in this simulation study are similar with the maximum likelihood estimate values under the mixed effects model in (2) using the high school project showcase data set described in Section 1. We assume a total of $a = 30$ student projects and $b = 15$ judges. The underlying true score of the i^{th} project is $\mu + \alpha_i$, which follows a uniform distribution between 40 and 60, independently for each $i = 1, \ldots, 30$. The systematic bias of the judge β_j follows a normal distribution with mean 0 and standard deviation $\sigma_B = 6$ independently, where $j = 1, \ldots, 15$. The random errors ϵ_{ij} are independently normally distributed with mean 0 and standard deviation $\sigma_\epsilon = 4$. The observed i^{th} student project score from the j^{th} judge y_{ij} is generated by adding $\mu + \alpha_i$, β_j and ϵ_{ij}. This is a complete data set including a total of $ab = 450$ observed scores. We put an average $\overline{y}_{i\cdot}$ across all judges for each project as its rating, and we get a ranking list of student projects based on the complete information provided by the simulated data set. Next, we randomly select a sample of 60 scores from the complete data set where each student project has been scored by $n = 2$ judges, and each judge has evaluated $m = 4$ projects. With this partial data set with 60 observations, we apply the simple average and mixed effects models described in Section 2 and six additional algorithms in Section 3, a total of eight ranking methods, and rank the student projects based on the calculated ratings from each algorithm. Finally, we compare the ranking list with the ground truth of the ranking result based on the known true scores $\mu + \alpha_i$. Two types of measures are applied to calculate the similarity between the ranking list computed and the ground truth: one is the correlation coefficients between full ranking lists (Critchlow, 1992), the other considers the top ten lists. We repeat this process 10^5 times and find the summary statistics and empirical distribution, as presented in Table 5 and Figure 1. Each time we generate a new complete data set and from which we draw one sample.

Denote the number of the true top ten student projects in the computed top ten list by N_{10}. We list the expected value of the number of the true top ten in a given top ten list $\mathbb{E}[N_{10}]$ and the probability that the computed top ten list contains at least eight true top ten $\Pr(N_{10} \geq 8)$ in Table 5. In addition, we include three correlation coefficients between the

ranking list calculated by a given method and the true ranking list in Table 5. The method of calculating the averages using the complete data set is included. All the other methods use the partial data set with 60 scores. The five measures of similarity are consistent, which show that the mixed effects model matches the ground truth the best, and PFK and Colley's method are the second and third. The mixed effects model is the only method that achieves a probability over 50% to find at least eight true top ten projects in its top ten list. The HITS algorithm performs less well under this setting than other ranking methods. We also include the empirical distributions of Kendall's τ with sample means and standard deviations using various ranking methods in Figure 1.

Table 5: Pearson's r, Kendall's τ, Spearman's ρ, $\mathbb{E}[N_{10}]$ and $\Pr(N_{10} \geq 8)$ for different judging systems, where the best three methods with the greatest values of each measure of the similarity are shown in bold

	Pearson's r	Kendall's τ	Spearman's ρ	$\mathbb{E}[N_{10}]$	$\Pr(N_{10} \geq 8)$
Complete Info	0.984	0.889	0.976	9.269	0.995
Simple Average	0.763	0.569	0.753	7.122	0.381
Mixed Effects	**0.820**	**0.627**	**0.810**	**7.523**	**0.535**
PageRank	0.751	0.557	0.741	6.999	0.334
HITS	0.248	0.151	0.213	4.246	0.003
PFK	**0.798**	**0.607**	**0.790**	**7.265**	**0.440**
Massey	0.765	0.571	0.753	7.104	0.392
Colley	**0.778**	**0.580**	**0.767**	**7.233**	**0.413**
Elo	0.759	0.566	0.752	7.096	0.370

5 Discussion

Mixed effects models play an important role in analyzing the non-independent data. Common examples of mixed effects modeling are generally from social research and trial studies (Raudenbush and Bryk, 2002). This article provides a different approach to apply the mixed effects models to build scoring systems from a new angle. Comparing with other existing ranking methods, it is natural to perform statistical inference and report confidence intervals and p-values when applying the mixed effects models.

Acknowledgments

The first author thanks Michelle L. Paulsen for providing the de-identified data for research and publication, and for equipping the students at Northwestern University with skills to tell the story behind their work through the RSG/RCTP program (Ready Set Go and Research Communication Training Program). The first author also thanks Ajit C. Tamhane for teaching him the design of experiments at Northwestern University. We thank Jesse

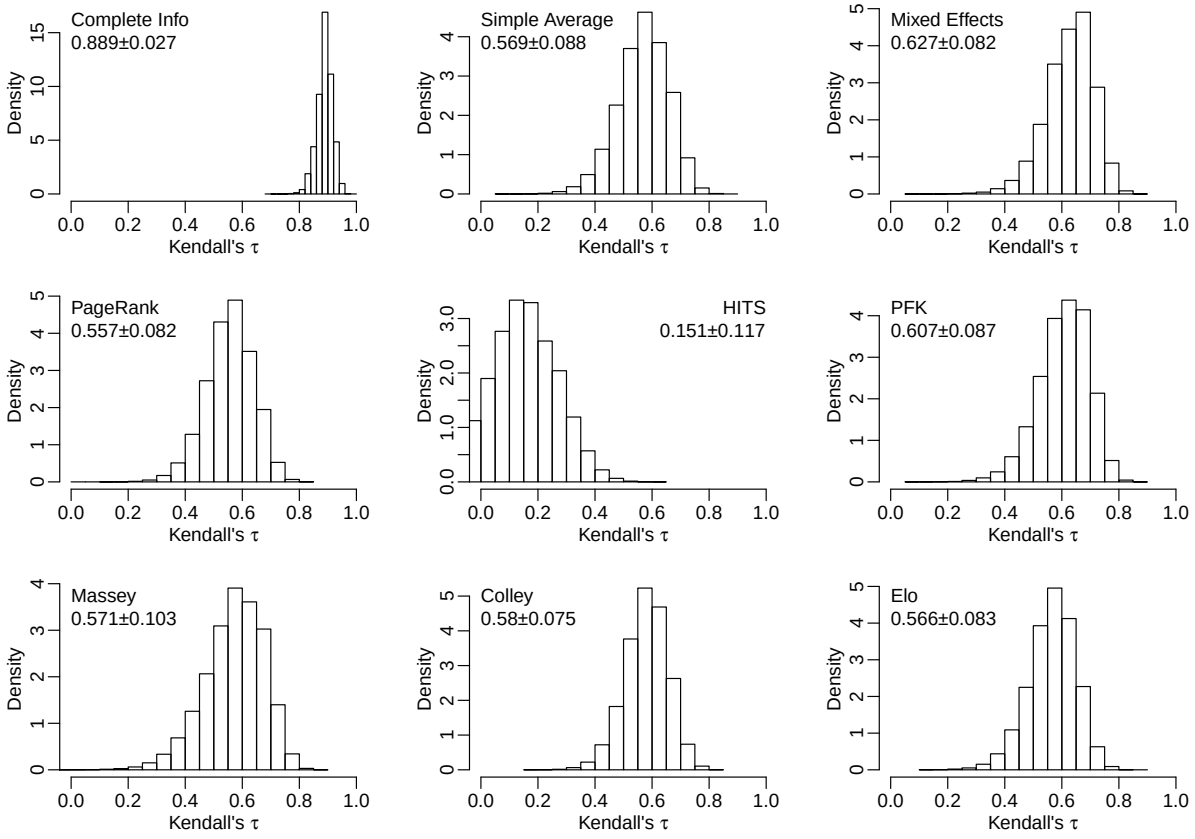

Figure 1: Empirical distributions of Kendall's τ between the true ranking and the ranking based on the complete data set (top left) or the ranking based on the partial data set using simple average (top middle), mixed effects models (top right), PageRank (middle left), HITS (center), PFK (middle right), Massey (bottom left), Colley (bottom middle) and Elo's method (bottom right)

Frey and Paul W. Bernhardt for discussing the design of MAT 8800-3 Independent Study: Statistical Learning Models in Spring 2020 at Villanova University.

Appendix

The appendix includes the High school project showcase data set, which is presented as a collection of judges. The student projects evaluated by the same judge are grouped in one line. For example, the first judge has evaluated three projects: Project 15 with score 41, Project 16 with score 51, and Project 22 with score 37.

Judge	Evaluated Projects	Judge	Evaluated Projects
1	P15: 41, P16: 51, P22: 37	10	P13: 53, P23: 50, P26: 54
2	P25: 48, P29: 52, P32: 45, P33: 52	11	P3: 37, P5: 47, P7: 45, P25: 37, P26: 44
3	P7: 49, P14: 62, P18: 50, P27: 51	12	P6: 59, P17: 60, P23: 46
4	P5: 60, P12: 38, P29: 40, P31: 66	13	P9: 36, P18: 37
5	P4: 37, P9: 56, P11: 58, P24: 33	14	P2: 41, P10: 59, P16: 43, P20: 40, P28: 59
6	P8: 32, P10: 47, P17: 55	15	P6: 33, P15: 45, P19: 32, P20: 26, P27: 48
7	P27: 63, P30: 64	16	P2: 43, P13: 44, P14: 51, P21: 39
8	P1: 39, P4: 47, P31: 63	17	P22: 29, P24: 38, P28: 34, P30: 61
9	P11: 47, P12: 40, P32: 37, P33: 32	18	P8: 49, P19: 54, P21: 51

References

Alvo, M. and Yu, P. L. (2014). *Statistical methods for ranking data*. Frontiers in probability and the statistical sciences. Springer-Verlag, New York.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30,** 107–117. Proceedings of the Seventh International World Wide Web Conference.

Colley, W. N. (2001). Colley's bias free college football ranking method: the Colley matrix explained. unpublished manuscript, available at `www.colleyrankings.com/matrate.pdf`.

Critchlow, D. E. (1992). On rank statistics: an approach via metrics on the permutation group. *Journal of Statistical Planning and Inference* **32,** 325–346.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing Company, New York.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburg and London.

Frey, J. (2005). A ranking method based on minimizing the number of in-sample errors. *The American Statistician* **59,** 207–216.

Frey, J. (2007). Bayesian inference on a proportion believed to be a simple fraction. *The American Statistician* **61,** 201–206.

Glickman, M. E. and Doan, T. (2017). The US chess rating system. unpublished manuscript, available at `http://www.glicko.net/ratings/rating.system.pdf`.

Keener, J. P. (1993). The Perron–Frobenius theorem and the ranking of football teams. *SIAM Review* **35,** 80–93.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* **46,** 604—-632.

Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

Langville, A. N. and Meyer, C. D. (2012). *Who's #1?: The Science of Rating and Ranking*. Princeton University Press.

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg, 2nd edition.

Massey, K. (1997). Statistical models applied to the rating of sports teams. Bachelor's Thesis, Bluefield College.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Advanced qualitative techniques in the social sciences. Sage Publications, Inc., Thousand Oaks, California, 2nd edition.

Stern, H. S. (1991). On the probability of winning a football game. *The American Statistician* **45,** 179–183.

Tamhane, A. C. (2009). *Statistical Analysis of Designed Experiments: Theory and Applications*. John Wiley & Sons, Hoboken, New Jersey.

Zhou, H. and Lange, K. (2009). Rating movies and rating the raters who rate them. *The American Statistician* **63,** 297–307.