Paired Comparisons with Ties: Modeling Game Outcomes in Chess

Daniel Shawul Rémi Coulom

September 19, 2013

Abstract

Bayesian rating of chess players requires a statistical model of the probabilities of a win, a draw, and a loss as a function of the rating difference between opponents. Some models are used in popular rating systems, but they were chosen rather arbitrarily, and it was not clear which fits the data best. In this paper, the goodness of fit of the Glenn-David (TrueSkill), Rao-Kupper (BayesElo), and Davidson models were measured for various databases of games between computers. Results demonstrate that the Davidson model fits the data best. The Davidson model features a draw distribution with longer tails, and, unlike the other models, makes two draws equivalent to one win and one loss. The Davidson model had not been used in any popular rating system, and the results presented in this paper will lead to a new improved version of BayesElo.

1 Introduction

Rating systems have greatly contributed to the popularity of chess and other games. The first chess rating system used in tournaments to produce numerical ratings was the Ingo rating system developed in 1948 by Anton Hoesslinger (Glickman, 1995). Different versions of this system were used in the following years, however they all lacked solid statistical foundations. Elo (1978) is usually credited with developing the first modern rating system that has sound statistical basis. However the basis for the Elo rating system, so called paired comparison in statistics, was first described much earlier by Zermelo (1929).

The Thurstone - Mosteller paired comparison mode (Thurstone, 1927),that assumes the performance of a player is normally distributed, is used in the Elo rating system. While Elo acknowledged each player may have different standard deviation (σ) of his Elo rating, he assumed the contrary and used a fixed value

of 200 elo points as the uncertainity margin. Therfore given performance of players (wins, losses, draws) in a tournament, the difference in ratings between two players can be calculated assuming ratings are normally distributed with sigma of $200\sqrt{2}$. A computationally simpler model known as Bradley-Terry paired comparison model Bradley and Terry (1952a) assumes players tend to over-perform, and therefore display a strength distribution skewed to the right. A generalized extreme value distribution (GEV type-I) that has long tails to the right is used for the model. Thus the difference in rating between two players will follow a logistic distribution, which is very close to Elo's gaussian assumption for all practical purposes. Henery (1992a) argues neither model is accurate because chess is usually won by a combination of accumulation of small advantages and brilliant moves.

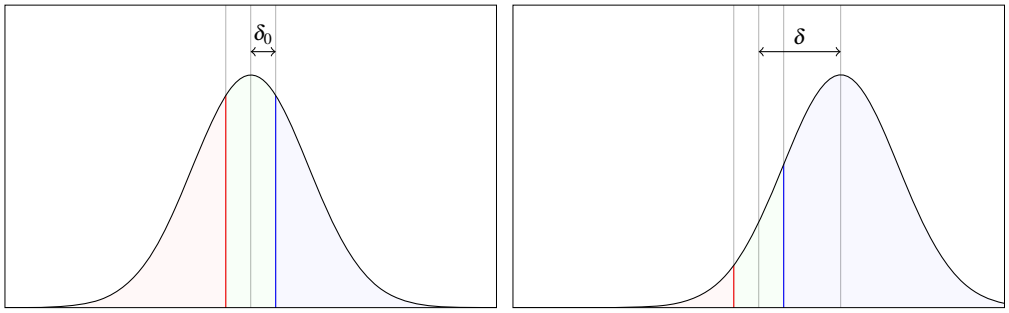
Hal (1992) examined a class of linear paired comparison models based on gamma random variables with different value of shape parameter k. The limiting values of k,i.e. 1 and ∞, give the Thurstone-Mosteller and Bradley-Terry models respectively. He found that the selected gamma model has a minimal effect on ratings obtained for samples of size encountered in practice. Thus he concluded that all linear models are essentially equivalent. However the conclusion is incomplete since the examination did not include linear models that are not of convolution type.

The paired comparison models discussed above ignore the effect of ties, home advantage, and other factors that are not relevant to chess Elo rating such as the effect of order and covariates. Thus paired comparison models need to be modified to include these effects as required. While the effect of home advantage is usually either ignored or handled the same way in many models, effect of ties has led to development of various models. This paper compares three different draw models used in different rating systems. The question posed by Stern "Are all linear models equivalent?" is interesting with the added effect of ties to ratings.

2 Models for Paired Comparisons with Ties

In the law of comparative judgment Thurstone (1927), all differences are assumed to be perceptible by the judge, thus no ties can occur. However ties do occur in games when the difference in performance ratings of two players fall below a certain threshold. This is the basis for all draw models investigated in this study namely the Glenn and David (1960) (GD),Rao and Kupper (1967) (RK) and Davidson (1970) (DV) models.

All of the draw modes investigated in this work are used in well known rating systems for computer games therefore this study is of a practical interest. The RK model is the basis of BayesElo (Coulom, 2005), a freeware tool popular in the chess



(a) $\delta = 0$ (b) $\delta > 0$

Figure 1: Principle of the Glenn-David model: the performance of a player in a game is assumed to be a random variable with a normal distribution. The difference between the performances of two opponents, plotted on these figures, is also normally distributed. A draw occurs when the performances of the opponents are within δ_0 of each other. The areas of the three regions represent the probabilities of a loss, a draw, and a win.

programming community. The GD model is used in TrueSkill (Herbrich et al., 2006), a rating system developed at Microsoft, and used in their Xbox game servers. The Davidson model is used in Edo ratings (Edwards, 2004). The paired comparison model used in RK and DV draw models is Bradley Terry, wheres the GD model uses the Thurstone-Mosteller model. Because the models for the strength distribution of a player are different,it is expected that calculated ratings will be different. Since the true rating of players is not known, the accuracy of the different models can not be judged by compaing against it. Thus the performance of the models are evaluated by how good it fits its own model constructed from a different data set. Given a set of results for each participant , part of the result can be used for training, and the rest to test prediction ability of the model. Accuracy of ratings, while very important, may not be the deciding factor for use in practical rating tools. Ease of use and rating computation time can sometimes be governing factors. For example, the Bradley-Terry models lend themselves to very fast computation using Minorization-Maximization methods, which is not applicable in the case Thurstone-Mosteller models. Also when fast computation is of ultimate importance,as is the case when ratings of thousands of players are continually updated e.g TrueSkill for Xbox, incremental approaches are preferred.

2.1 The Glenn-David Model

Glenn and David (1960)

$$\begin{aligned} P(W|\delta) &= \Phi(+\delta - \delta_0) \ ; \\ P(L|\delta) &= \Phi(-\delta - \delta_0) \ ; \\ P(D|\delta) &= 1 - P(W|\delta) - P(L|\delta) \ . \end{aligned}$$

Thurstone (1927), Fig. 1, Henery (1992b), Batchelder and Bershad (1979)

2.2 The Rao-Kupper Model

The Rao-Kupper model (1967) is similar to the Glenn-David model, except that Φ is replaced by the logistic function:

$$f(x) = \frac{1}{1+e^{-x}} \ .$$

Outcome probabilities become

$$\begin{aligned} P(W|\delta) &= f(+\delta - \delta_0) \ ; \\ P(L|\delta) &= f(-\delta - \delta_0) \ ; \\ P(D|\delta) &= 1 - P(W|\delta) - P(L|\delta) = \left(e^{2\delta_0} - 1\right) P(W|\delta) P(L|\delta) \ . \end{aligned}$$

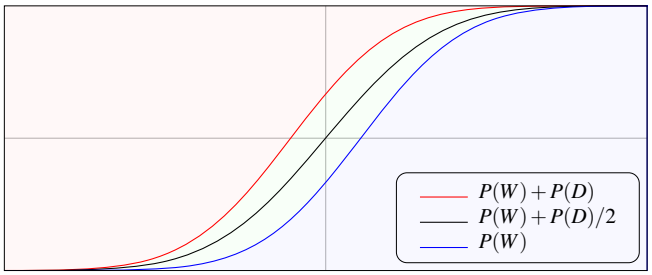
With the Rao-Kupper model, one win and one loss are equivalent to one draw. When $\delta_0 = 0$, the Rao-Kupper model becomes the Bradley-Terry model (1952b).

2.3 The Davidson Model

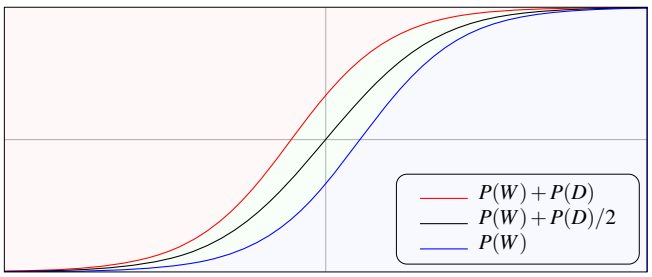
Davidson (1970) proposed another variation of the Bradley-Terry model. Unlike the Rao-Kupper model, the Davidson model assumes that one win and one loss are equivalent to two draws (instead of one):

$$\begin{aligned} d(\delta) &= \nu \sqrt{f(+\delta) f(-\delta)} \ ; \\ P(W|\delta) &= f(+\delta)/(1+d(\delta)) \ ; \\ P(L|\delta) &= f(-\delta)/(1+d(\delta)) \ ; \\ P(D|\delta) &= d(\delta)/(1+d(\delta)) = \nu \sqrt{P(W|\delta) P(L|\delta)} \ . \end{aligned}$$

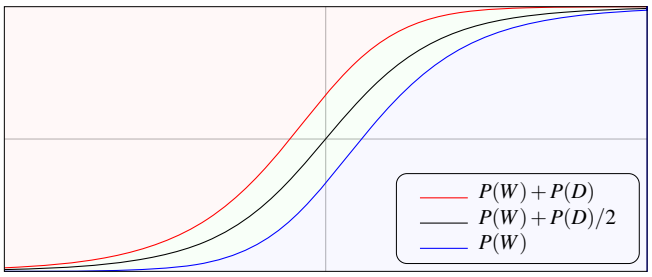
ν is a parameter of the model that indicates the probability of draws. $\nu = 0$ is equivalent to the Bradley-Terry model.



(a) Glenn-David

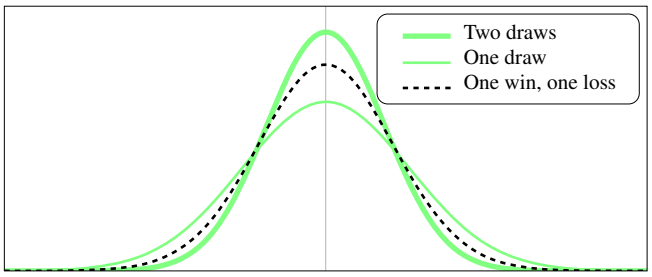


(b) Rao-Kupper

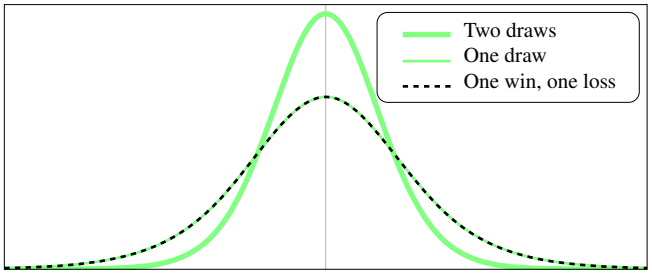


(c) Davidson

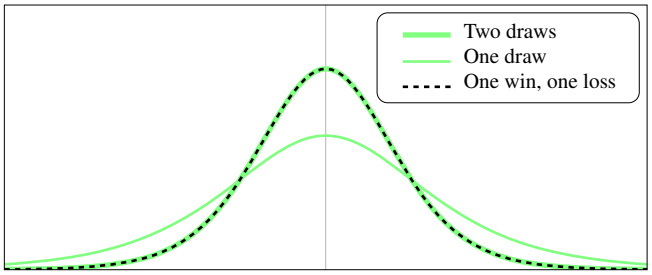
Figure 2: Outcome probabilities as a function of rating difference δ. Parameters of the models were chosen so that $P(W|\delta = 0) = P(D|\delta = 0) = P(L|\delta = 0) = 1/3$. Horizontal axes were scaled so that $P(W) + P(D)/2$ has the same derivative at $\delta = 0$ for all models.



(a) Glenn-David



(b) Rao-Kupper



(c) Davidson

Figure 3: Posterior rating probability densities with a uniform prior. Parameters and scales are like in Figure 2.

2.4 Individual Draw Percentages

The models discussed so far assume same values of parameters of draw and home advantage for all the participants. In real games of chess or soccer, the draw percentage may vary from player to player, or even be different in games played at home and away. In such cases a draw threshold i can be associated with each player (Joe, 1990). Furthermore two different values per player may be kept to account for home ground difference (Kuk, 1995). Kuk does the same thing for home advantage parameter. In the case of human ratings, temporal variations of these parameters are common due to ageing, learning etc. Once per player draw parameters are determined, draw threshold for a game between player i and j may be calculated as a sum $\sigma_{ij} = \sigma_i + \sigma_j$.

Joe used the winning percentage $p(win) + p(draw)/2$ in the linear model. Kuk argues this is not appealing as it hides the meaning of the strength parameters. With Joe's model larger differences of strength may be a result of higher draw rates. Therefore to allow for large number of draws, Kuk modeled $p(win)$ and $p(draw)$ separately. Joe used the Davidson draw model to study ranking of chess players and found that the draw model does not fit the data well. The reason for this is described as the lack of separate draw parameters for each player in the Davidson model.

While the use of separate draw and home advantage parameters allows more freedom, it can increase computational costs. Simpler alternatives with fewer parameters will be investigated in this work. One can assume linear or otherwise variation of these parameters with strength or time to reduce modeling complexity.

3 Model Selection

All the data used in our experiment comes from games between computer programs. This has an advantage in that large number of games is available from existing rating lists that will make the statistical study more reliable. For example the chess data collected from computer chess rating lists, CEGT and CCRL, at different time controls total about 2 million.

Tests for model selection are carried out using cross-validation on the collection of games. The K-fold cross validation method is used with partitions of 2, 4 and 10. In this method the k-1 partitions are used as a training data and then the predictive power of the model is tested on that one partition. We believe that this is more appropriate than testing goodness of fit on the same data that the model is trained with. Finally the k separate tests are averaged to produce a single estimation of the test parameter.

Random partitioning of data set of wins, draws and losses into k equal parts

is a problem of sampling without replacement (hyper geometric process). When the number of games between two players is less than the number of partitions k, sampling with replacement is used so that each partition has equal number of games in it.

The likelihood ratio test is used to compare the goodness of fit of two models. Bayeselo uses maximum likelihood method to determine Elo ratings, thus the log-likelihood is readily available. In the case of one of the draw models, Glenn-David, the computation takes very long time because fast minorization-maximization methods cannot be used. Instead a rather slow conjugate gradient with line search has to be used to maximize the likelihood. This has an important implication on the practical usability of the model since calculating ratings of a thousand or so players could take up to an hour.

4 Results

The first result is from comparison of each draw model with the actual observed frequency of draws. A bin of 5 elo is used to collect frequency data where the value for each bin is represented as a dot on the plots. The second plot depicts the well-known logistic curve that relates winning ratio with elo difference of players. The plots clearly show that the data fits the corresponding model very well. However this does not tell the whole story therefore a cross-validation test is carried out to measure prediction performance.

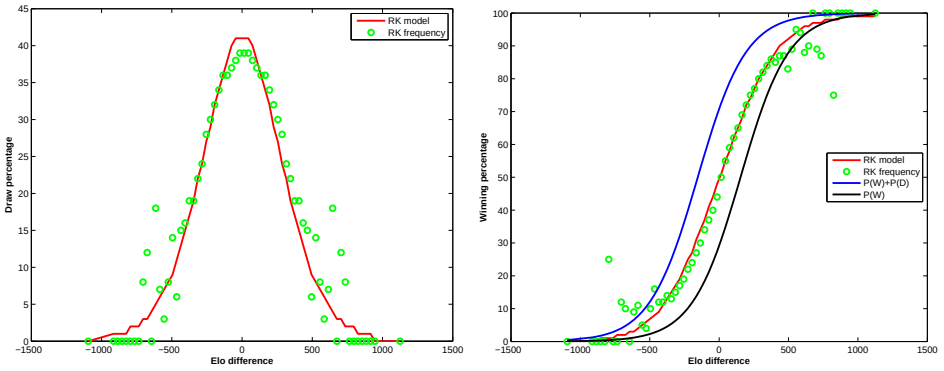


Figure 4: Results of Rao Kupper model.

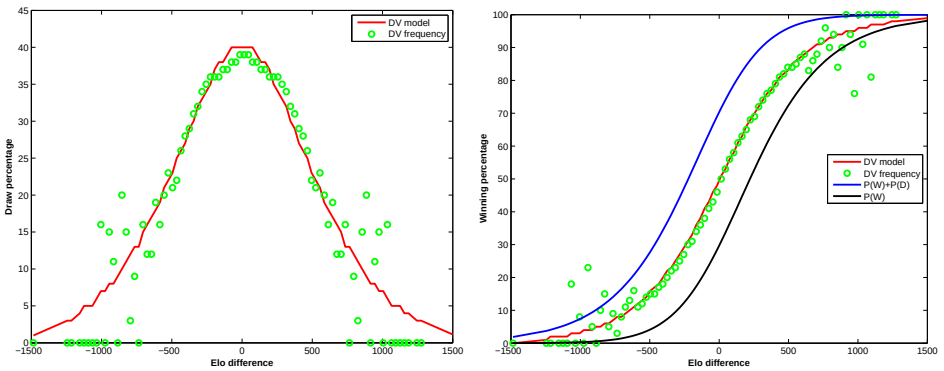


Figure 5: Results of Davidson model.

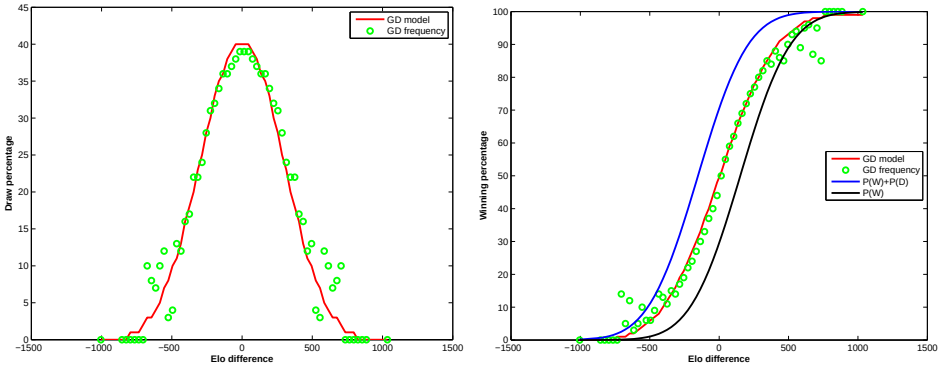


Figure 6: Results of Glenn David model.

Table 1: CCRL40/40

cross-2			**cross-4**			**cross-10**		
RK	**GD**	**DV**	**RK**	**GD**	**DV**	**RK**	**GD**	**DV**
-421526	-421285	-421182	-211181	-211068	-211024	-103052	-103023	-102997
-428867	-428602	-428471	-211292	-211181	-211137	-103281	-103249	-103235
			-211498	-211392	-211324	-103001	-102969	-102954
			-254640	-254500	-254435	-103061	-103048	-103027
						-103206	-103184	-103166
						-103231	-103215	-103211
						-103121	-103086	-103068
						-103294	-103285	-103267
						-102963	-102946	-102939
						-178792	-178713	-178655
-425197	-424944	-424827	-222153	-222035	-221980	-110700	-110672	-110652
740	**234**		**345.5**	**110.5**		**96.6**	**39.8**	

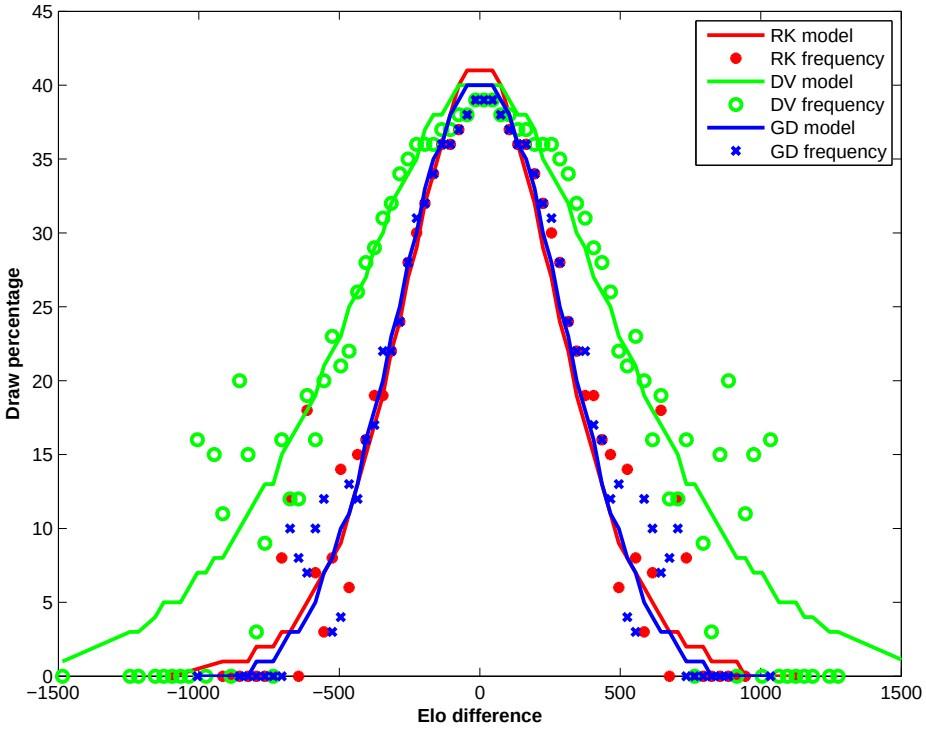


Figure 7: Summary of draw percentage predictions.

Table 2: CEGT-blitz

cross-2			cross-4			cross-10		
RK	**GD**	**DV**	**RK**	**GD**	**DV**	**RK**	**GD**	**DV**
-981281	-980854	-980705	-487156	-486982	-486897	-198654	-198576	-198552
-985422	-985067	-984944	-487332	-487133	-487059	-198549	-198466	-198422
			-487085	-486888	-486826	-198914	-198835	-198820
			-514526	-514346	-514278	-198750	-198699	-198684
						-198857	-198793	-198762
						-199205	-199136	-199092
						-199041	-198984	-198947
						-198567	-198497	-198458
						-199190	-199140	-199123
						-234711	-234640	-234592
-983352	-982961	-982825	-494025	-493837	-493765	-202444	-202377	-202345
1054	**272**		**519.5**	**144.5**		**197.2**	**62.8**	

Table 3: CCRL-blitz

cross-2			cross-4			cross-10		
RK	**GD**	**DV**	**RK**	**GD**	**DV**	**RK**	**GD**	**DV**
-916160	-915672	-915487	-448620	-448403	-448322	-173063	-172973	-172941
-925864	-925376	-925153	-448253	-448026	-447928	-173569	-173470	-173423
			-448581	-448342	-448248	-173541	-173451	-173420
			-497778	-497494	-497370	-173338	-173249	-173208
						-173070	-172989	-172944
						-173590	-173521	-173484
						-173590	-173513	-173492
						-173949	-173851	-173822
						-173273	-173179	-173145
						-289983	-289825	-289751
-921012	-920524	-920320	-460808	-460566	-460467	-185097	-185002	-184963
1384	**408**		**682**	**198.5**		**267.2**	**78.2**	

5 Conclusion

Davidson fits computer chess rating data better than the other two modes. This result is in contradiction with the finding of Joe that the model does not fit human ratings well. However the reason for that behavior was the lack of separate draw threshold parameters for each player. Here in our experiment all the models suffer from the same problem, so it should not be much of a surprise that Davidson came out as the best. Dangauthier et al. (2007), Coulom (2008)

References

Batchelder, W. H. and Bershad, N. J. (1979). The statistical analysis of a Thurstonian model for rating chess players. *Journal of Mathematical Psychology*, 19(1):39–60.

Bradley, R. A. and Terry, M. E. (1952a). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika Trust*, 39(3):324–345.

Bradley, R. A. and Terry, M. E. (1952b). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39(3–4):324–345.

Coulom, R. (2005). Bayeselo. http://remi.coulom.free.fr/Bayesian-Elo/.

Coulom, R. (2008). Whole-history rating: A Bayesian rating system for players of time-varying strength. In van den Herik, H. J., Xu, X., and Ma, Z., editors, *Proceedings of the 6th International Conference on Computer and Games*, volume 5131 of *Lecture Notes in Computer Science*, pages 113–124, Beijing, China. Springer.

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill through time: Revisiting the history of chess. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 337–344, Vancouver, Canada. MIT Press.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

Edwards, R. (2004). Edo historical chess ratings. http://members.shaw.ca/edo1/.

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

Glenn, W. A. and David, H. A. (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics*, 16(1):86–109.

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, (3):59–102.

Hal, S. (1992). Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103–117.

Henery, R. J. (1992a). An extension to the thurstone-mosteller model for chess. *Journal of the Royal Statistical Society*, 41(5):559–567.

Henery, R. J. (1992b). An extension to the Thurstone-Mosteller model for chess. *The Statistician*, (41):559–567.

Herbrich, R., Minka, T., and Graepel, T. (2006). TrueSkill$^{\text{TM}}$: A Bayesian skill rating system. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 569–576, Vancouver, British Columbia, Canada. MIT Press.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406.

Joe, H. (1990). Extended use of paired comparion models, with applications to chess rankings. *Journal of the Royal Statistical Society*, 39(1):85–93.

Kuk, A. Y. (1995). Extended use of paired comparion models, with applications to chess rankings. *Journal of the Royal Statistical Society*, 44(4):523–528.

Rao, P. V. and Kupper, L. L. (1967). Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62:194–204.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460.

A MM Formula for the Rao-Kupper and Davidson Models

Hunter (2004)

Data: w_{ij}, l_{ij}, d_{ij} are respectively wins, losses and draws of i against j, i playing as White.

Model parameters: γ_i is the strength of player i. θ_w is the advantage of playing as White. θ_d is the draw parameter.

Model (i is White):

A.1 Rao Kupper Model

Outcome probabilities:

$$P(i \text{ beats } j) = \frac{\theta_w \gamma_i}{\theta_w \gamma_i + \theta_d \gamma_j}$$
$$P(j \text{ beats } i) = \frac{\gamma_j}{\theta_w \theta_d \gamma_i + \gamma_j}$$
$$P(i \text{ ties } j) = (\theta_d^2 - 1) P(i \text{ beats } j) P(j \text{ beats } i)$$

Update rules:

$$\gamma_i \leftarrow \frac{\sum_j w_{ij} + d_{ij} + l_{ji} + d_{ji}}{\sum_j \frac{(d_{ij} + w_{ij})\theta_w}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(d_{ij} + l_{ij})\theta_d\theta_w}{\theta_d\theta_w\gamma_i + \gamma_j} + \frac{(d_{ji} + w_{ji})\theta_d}{\theta_w\gamma_j + \theta_d\gamma_i} + \frac{d_{ji} + l_{ji}}{\theta_d\theta_w\gamma_j + \gamma_i}}$$

$$\theta_w \leftarrow \frac{\sum_{ij} w_{ij} + d_{ij}}{\sum_{ij} \frac{(w_{ij} + d_{ij})\gamma_i}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(l_{ij} + d_{ij})\theta_d\gamma_i}{\theta_d\theta_w\gamma_i + \gamma_j}}$$

$$\theta_d \leftarrow \alpha + \sqrt{\alpha^2 + 1}, \text{ with } \alpha = \frac{\sum_{ij} d_{ij}}{\sum_{ij} \frac{(w_{ij} + d_{ij})\gamma_j}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(l_{ij} + d_{ij})\theta_w\gamma_i}{\theta_d\theta_w\gamma_i + \gamma_j}}$$

A.2 Davidson Model

Outcome probabilities:

$$P(i \text{ beats } j) = \frac{\theta_w\gamma_i}{\theta_w\gamma_i + \gamma_j + \theta_d\sqrt{\theta_w\gamma_i\gamma_j}}$$

$$P(j \text{ beats } i) = \frac{\gamma_j}{\theta_w\gamma_i + \gamma_j + \theta_d\sqrt{\theta_w\gamma_i\gamma_j}}$$

$$P(i \text{ ties } j) = \theta_d\sqrt{P(i \text{ beats } j)P(j \text{ beats } i)}$$

Update rules:

$$\gamma_i \leftarrow \frac{\sum_j w_{ij} + \frac{d_{ij}}{2} + l_{ji} + \frac{d_{ji}}{2}}{\sum_j \left(\theta_w + \theta_d \sqrt{\frac{\theta_w \gamma_j}{\gamma_i}}\right) \frac{w_{ij} + d_{ij} + l_{ij}}{\theta_w \gamma_i + \gamma_j + \theta_d \sqrt{\theta_w \gamma_i \gamma_j}} + \left(1 + \theta_d \sqrt{\frac{\theta_w \gamma_j}{\gamma_i}}\right) \frac{w_{ji} + d_{ji} + l_{ji}}{\theta_w \gamma_j + \gamma_i + \theta_d \sqrt{\theta_w \gamma_i \gamma_j}}}$$

$$\theta_w \leftarrow \left(\frac{-b + \sqrt{b^2 + 16ac}}{4a}\right)^2, \text{ with}$$

$$a = \sum_{ij} \frac{(w_{ij} + d_{ij} + l_{ij})\gamma_i}{\theta_w \gamma_i + \gamma_j + \theta_d \sqrt{\theta_w \gamma_i \gamma_j}},$$

$$b = \sum_{ij} \frac{(w_{ij} + d_{ij} + l_{ij})\theta_d \sqrt{\gamma_i \gamma_j}}{\theta_w \gamma_i + \gamma_j + \theta_d \sqrt{\theta_w \gamma_i \gamma_j}}, \text{ and}$$

$$c = \sum_{ij} w_{ij} + \frac{d_{ij}}{2}$$

$$\theta_d \leftarrow \frac{\sum_{ij} d_{ij}}{\sum_{ij} \frac{(w_{ij} + d_{ij} + l_{ij})\sqrt{\theta_w \gamma_i \gamma_j}}{\theta_w \gamma_i + \gamma_j + \theta_d \sqrt{\theta_w \gamma_i \gamma_j}}}$$