Elo and Glicko Standardised Rating Systems

by Nicolás Wiedersheim Sendagorta

Introduction

"Manchester United is a great team" argues my friend. "But it will never be better than Manchester City". People rate everything, from films and fashion to parks and politicians. It's usually an effortless task. A pool of judges quantify somethings value, and an average is determined. However, what if it's value is unknown. In this scenario, economists turn to auction theory or statistical rating systems. "Skill" is one of these cases. The Elo and Glicko rating systems tend to be the go-to when quantifying skill in Boolean games.

Elo Rating System

Figure 1: Arpad Elo

As I started playing competitive chess, I came across the Elo rating system. It is a method of calculating a player's relative skill in zero-sum games. Quantifying ability is helpful to determine how much better some players are, as well as for skill-based matchmaking. It was devised in 1960 by Arpad Elo, a professor of Physics and chess master. The International Chess Federation would soon adopt this rating system to divide the skill of its players. The algorithm's success in categorising people on ability would start to gain traction in other win-lose games; Association football, NBA, and "*esports*" being just some of the many examples. *Esports* are highly competitive videogames.

Figure 2: Top Elo Ratings for 2014 FIFA World Cup

Brazil	2113
Spain	2086
Germany	2046
Argentina	1989
Netherlands	1959
England	1914
Portugal	1902
Colombia	1897
Uruguay	1895
Chile	1895

The Elo rating system consists of a formula that can determine your quantified skill based on the outcome of your games. The "skill" trend follows a gaussian bell curve. Most of the community lies around 1600, while fewer lie on the extremes. Players scores will slightly fluctuate. Better players, however, will score higher performance ratings more consistently. In the graph below, the blue indicates the worst performing player while the red is the better performing player. Their ratings correspond to the mean of these curves.

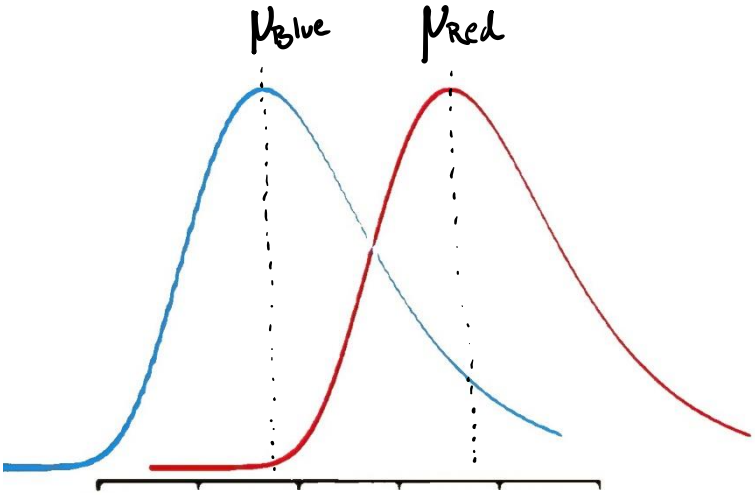


Figure 3: Lower and Higher Performing Players in Elo System

The formula to determine the expected success rate of a player is

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}$$

To use this, one has to consider the logistic function. It is the normal distribution of success in the player's difference of score rating.

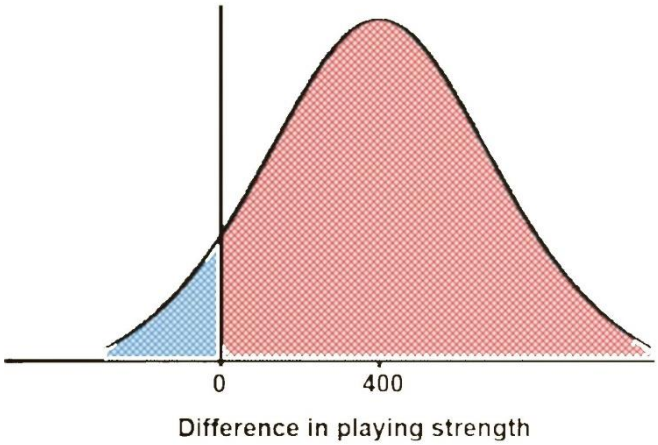


Figure 4: Logistic Function of Difference in Elo Rating

Here, the score difference between the players is 400. For every 400 point difference, the higher-rated player is ten times more likely to win the game. This means that a 2400 rated player is 100 times more likely to beat a 1600 rated player.

This winning probability can be expressed with the formula

$$P(p_1 > p_2|s_1,\ s_2) = \frac{\Phi(R_a - R_b)}{\sqrt{2}\beta}$$

Where Φ is the cumulative density of a zero-mean Gaussian with unit-variance. β^2 is the variance.

The linear approximation is

$$P(A\ Wins) = 10^{\frac{R_A - R_B}{400}} * P(B\ Wins)$$

If the probability output is 1, then a win is certain. On the other hand, a probability of 0 is an inevitable loss. By defining 1 as a win and 0 as a loss, the probability can be used as the expected score. However, what happens if a player beats the odds of losing? If the player does better then expected, then the rating will increase. The more surprising the win is, the more the rating will increase. The more unexpected the defeat, the more points are subtracted from the rating. The "rating update" formula for several games in a period is defined as:

$$R_{new} = R_{old} + \frac{K}{2}\left(W - (L + \frac{(\sum_i D_i)}{C}\right)$$

or

$$R_{new} = R_{old} + K(S - S_e)$$

For individual games. W is wins, L is losses, $D_i$is the rating difference, C is a constant of 400, and K is the K-factor, the total potential increase or decrease. S_f is the final score(0,0.5,1) while $S_e$is the expected score.

Consider player A and player B, both of 1600 rating. If player A wins, his new rating will be:

$$R_{new} = 1600 + 32(1 - 0.5) = 1616$$

The k-factor being 32 under current chess federation regulations.

If a new player joins without any game history, his ranking will be pre-set at 1500. It does not start at 0, as it might take too long for someone to reach their actual rank if that were to be the case.

To apply a rating algorithm, a collection of games within a "rating period" is treated. The rating periods can last as much as several months. The finalised updated rating with minimum standard deviation would be at the end of these rating periods. This is one of the reasons for competitive videogames to have ranking "seasons" as periods to determine the skill of its players.

The ranking system, however, is only highly accurate for 1 vs 1 games. It is harder to apply to multi-teamed games such as Chinese checkers or monopoly.

Glicko rating system

The Glicko rating system was made by Mark Glickman as an improvement on the Elo rating system. It addresses a problem relating to the reliability of a players rating. Consider two players, both rated 1700, played a tournament together. Player A beats player B. According to the Elo system, one player would gain 16 points, and the other would lose 16 points. However, suppose that player A has recently returned to chess after many years, while player B plays every day. In this scenario, player A's rating is not reliable, while that of player B is much more trustworthy. To improve this, player A's rating should technically increase by a significant amount (more than 16 points), and player B's rating should decrease by less than 16 points. Therefore, Glicko extends on the Elo rating by considering the rating and its standard deviation. A high Rating Deviation (RD) indicates a high rating uncertainty, and a low RD suggests that it is more confident due to frequent playing. As more rated games are played, the lower the RD gets; additional information is obtained about the player's actual rating. However, as time passes without playing games, the more the RD increases.

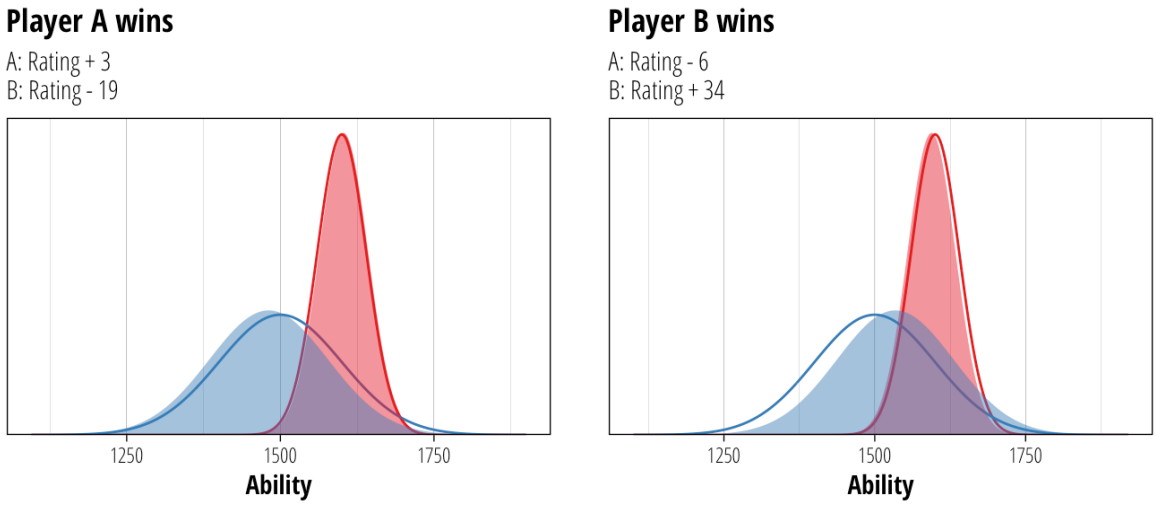


Figure 5: Rating update in Glicko System

If it's the players first game, much like in the Elo system, the rating is set to 1500 and the deviation at 350. Otherwise, the deviation is determined from the formula:

$$Rating\ Deviation, RD = \min \sqrt{RD_{old}^2 + c^2}$$

Where c is a constant that controls the increase of uncertainty of RD between rating periods. (when someone takes a rest from playing).

To update the rating after a rating period one can apply the formulas:

$$r' = r + \frac{q}{\frac{1}{RD^2} + \frac{1}{d^2}} \sum_{j=1}^{m} g(RD_j)(s_j - E(s|r, r_j, RD_j))$$

$$RD' = \sqrt{\left(\frac{1}{RD^2} + \frac{1}{d^2}\right)^{-1}}$$

Here, r and RD is the pre-period rating and deviation, while r' and RD' is the post-period rating and deviation.

Where $q = 0.0057565$ and

$$g(RD) = \frac{1}{\sqrt{1 + \frac{3q^2(RD^2)}{\pi^2}}}$$

$$E(s|r, r_j, RD_j) = \frac{1}{1 + 10^{-\frac{g(RD_j)(r - r_j)}{400}}}$$

$$d^2 = q^2 \sum_{j=1}^{m} \left(\left(g(RD_j)\right)^2 E(s|r, r_j, RD_j) * \left(s_j - E(s|r, r_j, RD_j)\right) \right)$$

To set an example, a player of 1500 rating goes against 1400, 1550 and 1700 rated players. Their RD is 30, 100 and 300 respectively. Our player wins against the 1400 rated player but loses the other 2.

j	r_j	RD_j	$g(RD_j)$	$E(s\|r, r_j, RD_j)$	**outcome**
1	1400	30	0.9955	0.639	1
2	1550	100	0.9531	0.432	0
3	1700	300	0.7242	0.303	0

$$d^2 = (0.0057565)^2[(0.9955)^2\,(0.639)(1 - 0.639) + (0.9531)^2\,(0.432)(1 - 0.432) + (0.7242)^2(0.303)(1 - 0.303)]^{-1} - 1 = 53670.85 = 231.67^2$$

$$r' = 1500 + \frac{0.0057565}{\frac{1}{200^2} + \frac{1}{231.67^2}} + [0.9955(1 - 0.639) + 0.9531(0 - 0.432) + 0.7242(0 - 0.303)]$$
$$= 1500 + 131.9(-0.272) = 1500 - 36 = 1464$$

and therefore

$$RD' = \sqrt{\left(\frac{1}{200^2} + \frac{1}{231.67^2}\right)^{-1}} = \sqrt{22918.9} = 151.4$$

So our player's new rating would be 1464 with 151.4 RD

Much like in the Elo system, the glicko system also allows one to find the expected outcome of a game by applying the equation

$$E = \frac{1}{1 + 10^{-\frac{g\sqrt{RD_i^2 + RD_j^2}(r_i - r_j)}{400}}}$$

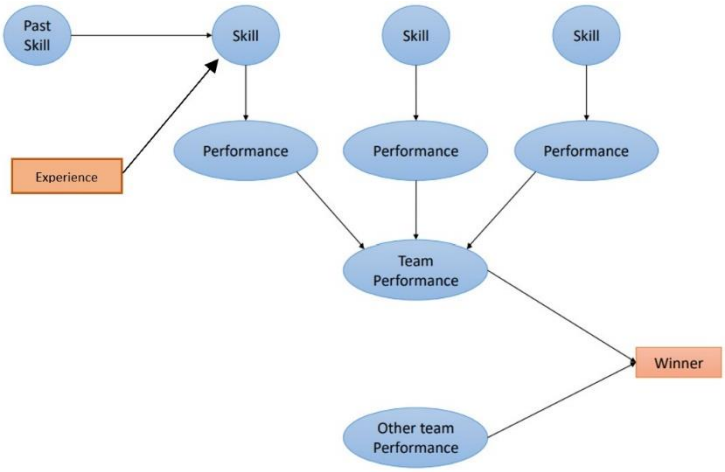


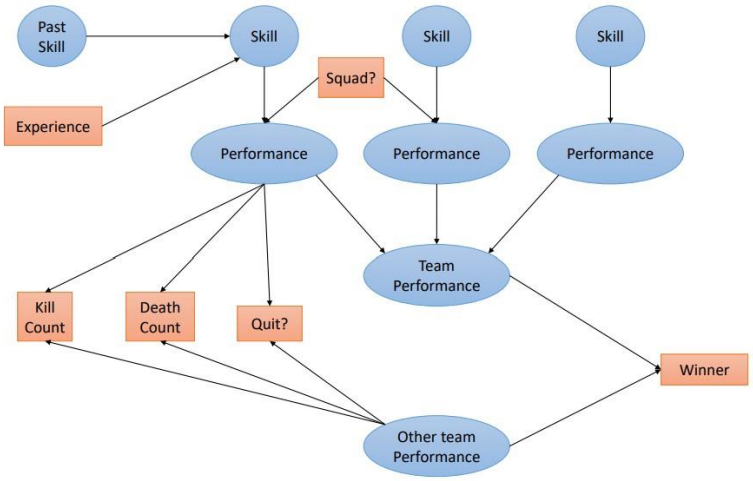
Figure 6: Glicko Rating System

The expected performance of individual players can then be averaged to find the total expected performance in team based games such as Basketball or Football. This is sometimes used as a metric by gamblers to forecast the outcome of games.

It must be stated that both ELO and Glicko are ratings based on other people's ratings. If everyone got worst at chess throughout time, the rating ladder would be easier to climb. Therefore, it can only rate players at a given time but cannot compare players from different eras. This also applies to many videogames. However, it was suggested in 2017 that a new system should compare rating by movement accuracy when compared to that of a powerful computer engine (Alliot, 2017). This new evaluation method considers the result of the game and the movements played within it. Another potential problem with these rating systems is that they can discourage game activity to protect rating. This specific problem happened in a "Magic the Gathering" tournament. Professional Players were attempting to limit their game activity to prevent a potential decrease. To solve this problem, chess grandmaster John Nunn suggested implementing a point bonus for those who kept on playing.

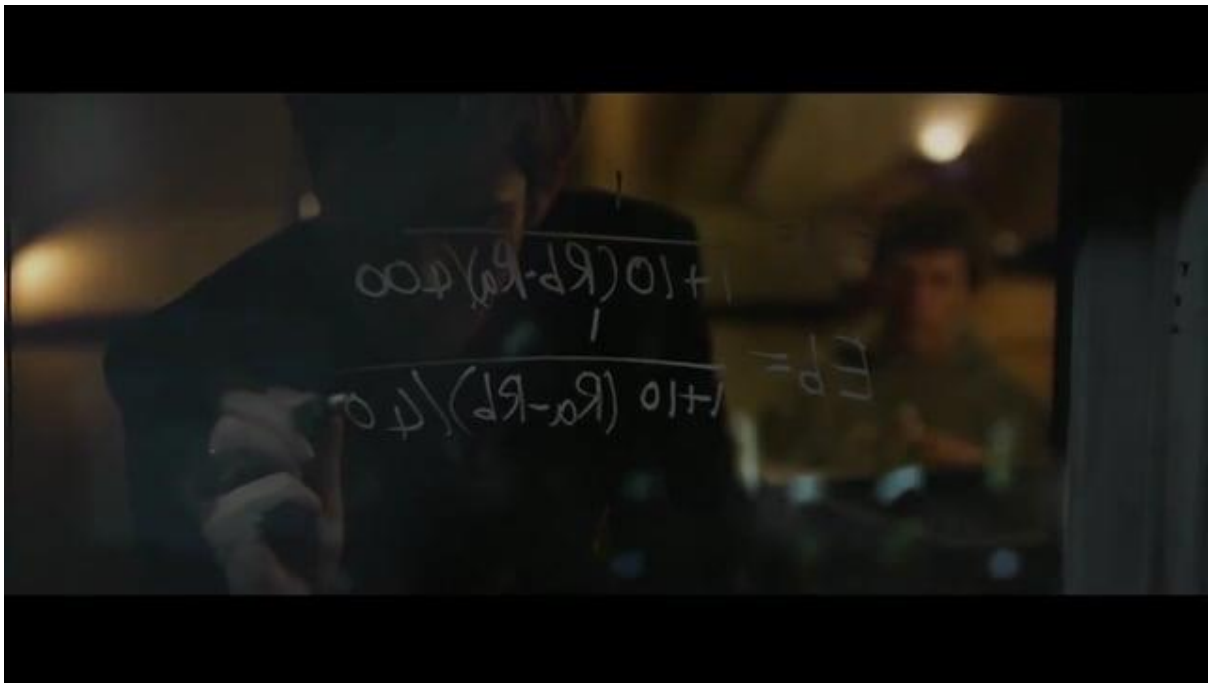
Some other ranking systems that stem from these are “TrueSkill 2”, which Microsoft researched to improve ranking based on in-game activity rather than just outcome and experience. “Trueskill 2” is mainly used for videogames

Figure 7: TrueSkill 2



“Matchmaking to date” can be used quite literally when talking about the algorithms applications beyond games. The Elo would determine the “desirability score” in the dating application “Tinder”; used to pair people of similar attractiveness together (Carr, 2016). This was also done by Mark Zuckerberg when making the website “Facemash”, predecessor of Facebook (Fincher, 2017). These are only some of the many sectors in which the Elo and Glicko run in the background. The formulas discussed give an insight as to how these algorithms operate to compare groups abilities. Comparisons which can be detrimental in multimillion dollar sports associations and businesses alike.

Figure 8: Scene from "The Social Network" mentioning "facemash" and Elo Rating



Bibliography

Alliot, J.-M. (2017). Who is the Master? *ICGA Journal*, vol. 39, no. 1, pp. 3-43.

Carr, A. (2016). I Found Out My Secret Internal Tinder Rating And Now I Wish I Hadnt. *Fast Company*, 1.

FIFA. (2018). Revision of the FIFA / Coca-Cola World Ranking . *WayBack Machine*, 2.

Fincher, D. (Director). (2017). *The Social Network* [Motion Picture].

Glickman, D. M. (revised 2016). The Glicko system. 6.

Grimes, J. (2019, Feb 15). *The Elo rating system for chess and beyond*. Retrieved from Youtube: https://www.youtube.com/watch?v=AsYfbmp0To0&ab_channel=singingbanana

Lars Magnus Hvattum, H. A. (2010). Using ELO ratings for match result prediction in association. *ijforecast*.

Lyons, K. (2014). What are the World Football Elo Ratings? *The Conversation*, 2.

Pokemon. (n.d.). Play! Pokemon Glossary. *Pokemon Tournament Glossary*.

system, T. 2. (2018, 3 8). *TrueSkill 2: An improved Bayesian skill rating system.* Retrieved from www.microsoft.com: https://www.microsoft.com/en-us/research/publication/trueskill-2-improved-bayesian-skill-rating-system/

Elo, Arpad E “Logistic Probability as a Rating Basis". *The Rating of Chessplayers, Past&Present*. Bronx NY 10453: ISHI Press International. ISBN 978-0-923891-27-5.