The Glicko system

Professor Mark E. Glickman
Boston University

Arguably one of the greatest fascinations of tournament chess players and competitors of other games is the measurement of playing strength. The Elo rating system, developed by Arpad Elo in the early 1960's, was the first chess rating system that had probabilistic underpinnings, and was then adopted by many chess federations, and eventually by organizations for other games (e.g., Scrabble, table tennis, etc.). While Elo's system is a great improvement over earlier systems, it too has its problems. In 1995, I created the Glicko rating system in response to a particular deficiency in the Elo system which I describe below. My system was derived by considering a statistical model for chess game outcomes, and then making mathematical approximations that would enable simple computation. The Elo system, coincidentally, turns out to be a special case of my system. The mathematical details of the derivation can be found in a technical paper called "Parameter estimation in large paired comparison experiments" which is published in the refereed statistics journal *Applied Statistics* (48, pp. 377–394), but can also be downloaded from **http://math.bu.edu/people/mg/research.html**. The Glicko system is currently implemented on the free internet chess server (FICS), and variations of the Glicko system have been adapted for several commercial internet gaming organizations such as ChronX, Case's Ladder, and the Gothic Chess Association.

The problem with the Elo system that the Glicko system addresses has to do with the reliability of a player's rating. Suppose two players, both rated 1700, played a tournament game with the first player defeating the second. Under the US Chess Federation's version of the Elo system, the first player would gain 16 rating points and the second player would lose 16 points. But suppose that the first player had just returned to tournament play after many years, while the second player plays every weekend. In this situation, the first player's rating of 1700 is not a very reliable measure of his strength, while the second player's rating of 1700 is much more trustworthy. My intuition tells me that (1) the first player's rating should increase by a large amount (more than 16 points) because his rating of 1700 is not believable in the first place, and that defeating a player with a fairly precise rating of 1700 is reasonable evidence that his strength is probably much higher than 1700, and (2) the second player's rating should decrease by a small amount (less than 16 points) because his rating is already precisely measured to be near 1700, and that he loses to a player whose rating cannot be trusted, so that very little information about his own playing strength has been learned.

While most situations are not so extreme, I felt it would be useful to incorporate into a rating system a measure of reliability of one's rating. The Glicko system therefore extends the Elo system by computing not only a rating, which can be thought of as a "best guess"

of one's playing strength, but also a "ratings deviation" (RD) or, in statistical terminology, a standard deviation, which measures the uncertainty in a rating (high RD's correspond to unreliable ratings). A high RD indicates that a player may not be competing frequently or that a player has only competed in a small number of tournament games. A low RD indicates that a player competes frequently.

In the Glicko system, a player's rating changes only from game outcomes, but his/her RD changes both from game outcomes and also from the passage of time when not playing. One feature of the system is that game outcomes always decrease a player's RD, and that time passing without competing in rated games always increases a player's RD. The reason is that the more games played, the more information is learned about a player's ability, so the more precise the rating becomes. As time passes, we become more uncertain about the player's strength, so this is reflected in the RD increasing.

It is interesting to note that, in the Glicko system, rating changes are not balanced as they usually are in the Elo system. If one player's rating increases by x, the opponent's rating does not usually decrease by x as in the Elo system. In fact, in the Glicko system, the amount by which the opponent's rating decreases is governed by both players' RD's.

Because a player in the Glicko system has both a rating and an RD, it is usually more informative to summarize a player's strength in the form of an interval (rather than merely report a rating). One way to do this is to report a 95% confidence interval. The lowest value in the interval is the player's rating minus twice the RD, and the highest value is the player's rating plus twice the RD. So, for example, if a player's rating is 1850 and the RD is 50, the interval would go from 1750 to 1950. We would then say that we're 95% confident that the player's actual strength is between 1750 and 1950. When a player has a low RD, the interval would be narrow, so that we would be 95% confident about a player's strength being in a small interval of values.

The formulas:

To apply the rating algorithm, we treat a collection of games within a "rating period" to have occurred simultaneously. A rating period could be as long as several months, or could be as short as one minute. In the former case, players would have ratings and RD's at the beginning of the rating period, game outcomes would be observed, and then updated ratings and RD's would be computed at the end of the rating period (which would then be used as the pre-period ratings and RD's for the subsequent rating period). In the latter case, ratings and RD's would be updated on a game-by-game basis (this is currently the system used by FICS). The Glicko system works best when the number of games in a rating period is moderate, say an average of 5-10 games per player in a rating period. The length of time for a rating period is at the discretion of the administrator.

Step 1. Determine a rating and RD for each player at the onset of the rating period.

(a) If the player is unrated, set the rating to 1500 and the RD to 350.

(b) Otherwise, use the player's most recent rating, and calculate the new RD from the old RD (RD_{old}) by the formula

$$\mathrm{RD} = \min(\sqrt{\mathrm{RD}_{old}^2 + c^2 t},\ 350)$$

where t is the number of rating periods since last competition (e.g., if the player competed in the most recent rating period, $t = 1$) and c is a constant that governs the increase in uncertainty over time. See below for a discussion of the choice of c. The formula above ensures that an RD at the beginning of a rating period is never larger than 350, the RD for an unrated player.

Step 2. Carry out the following updating calculations for each player separately:

Assume that the player's pre-period rating is r, and the ratings deviation is RD. Let the pre-period ratings of the m opponents be $r_1, r_2, \ldots, r_m$ and the ratings deviations be $\mathrm{RD}_1, \mathrm{RD}_2, \ldots, \mathrm{RD}_m$. Also let $s_1, \ldots, s_m$ be the outcome against each opponent, with an outcome being either 1, $\frac{1}{2}$, or 0 for a win, draw and loss. Note that multiple games against the same opponent are treated as games against multiple opponents with the same rating and RD.

Let r' and RD' denote the post-period rating and ratings deviation for the player. The updating formulas are given by

$$\begin{aligned} r' &= r + \frac{q}{1/\mathrm{RD}^2 + 1/d^2} \sum_{j=1}^{m} g(\mathrm{RD}_j)(s_j - \mathrm{E}(s|r, r_j, \mathrm{RD}_j)) \\ \mathrm{RD}' &= \sqrt{\left(\frac{1}{\mathrm{RD}^2} + \frac{1}{d^2}\right)^{-1}} \end{aligned}$$

where

$$\begin{aligned} q &= \frac{\ln 10}{400} = 0.0057565 \\ g(\mathrm{RD}) &= \frac{1}{\sqrt{1 + 3q^2(\mathrm{RD}^2)/\pi^2}} \\ \mathrm{E}(s|r, r_j, \mathrm{RD}_j) &= \frac{1}{1 + 10^{-g(\mathrm{RD}_j)(r - r_j)/400}} \\ d^2 &= \left(q^2 \sum_{j=1}^{m} (g(\mathrm{RD}_j))^2 \mathrm{E}(s|r, r_j, \mathrm{RD}_j)(1 - \mathrm{E}(s|r, r_j, \mathrm{RD}_j)) \right)^{-1}. \end{aligned}$$

These calculations are carried out for every player competing in the rating period.

Example calculation:

To demonstrate Step 2 of the calculations above, suppose a player rated 1500 competes against players rated 1400, 1550 and 1700, winning the first game and losing the next two. Assume the 1500-rated player's rating deviation is 200, and his opponents' are 30, 100 and 300, respectively.

We can calculate:

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

We can then compute

$$\begin{aligned} d^2 &= \Big((0.0057565)^2[(0.9955)^2(0.639)(1-0.639) \\ &\quad +(0.9531)^2(0.432)(1-0.432)+(0.7242)^2(0.303)(1-0.303)]\Big)^{-1} \\ &= 53670.85 = 231.67^2. \end{aligned}$$

We now have

$$\begin{aligned} r' &= 1500 + \frac{0.0057565}{\left(\frac{1}{200^2}+\frac{1}{231.67^2}\right)} \times \\ &\quad [0.9955(1-0.639) \\ &\quad +0.9531(0-0.432) \\ &\quad +0.7242(0-0.303)] \\ &= 1500 + 131.9(-0.272) = 1500 - 36 = \underline{1464} \end{aligned}$$

and

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

Implementation issues:

The value of c used in the Step 1b of the rating algorithm can be determined by data analysis, though this could be a computing-intensive process. Another approach is to determine how much time (in units of rating periods) would need to pass before a rating for a typical player becomes as uncertain as that of an unrated player. To demonstrate the calculation that would result from this approach, suppose a typical player has an RD of 50, rating periods last two months, and that it is assumed that 5 years (60 months) would need to pass before the typical player's rating becomes as unreliable as an unrated player's "rating." The time that must pass would be $t = 30$ rating periods (30 2-month periods). We want to solve for c such that

$$350 = \sqrt{50^2 + c^2(30)}.$$

In this case, $c = 63.2$ would be used.

One practical problem with the Glicko system is that when a player competes very frequently, his/her rating stops changing appreciably which reflects that the RD is very small. This may sometimes prevent a player's rating from changing substantially when the player is truly improving. I would therefore recommend that an RD never drop below a threshold value, such as 30, so that ratings can change appreciably even in a relatively short time.