Ratings Central: Accurate, Automated, Bayesian Table Tennis Ratings for Clubs, Leagues, Tournaments, and Organizations

David J. Marcus

Joint Statistical Meetings
July 30–August 4, 2011

History

- Since the 1970s, USA Table Tennis has had a rating system.
  - The rating system has a superficial similarity to the Elo system (i.e., the chess system).
- Players of all levels (from champion to five-year old) compete in USATT-sanctioned tournaments and get ratings.
- As a member of the USATT Ratings Committee in 1997, I was asked if I could develop a better rating system.
  - Problems with the ratings included excessive volatility and players protecting their ratings by not playing.
- By the time I had developed the new system in 1999, supportive USATT Board members had been replaced by people who would oppose anything we suggested.
  - USATT did not adopt the new system.
- We launched Ratings Central on our own in 2004.

Ratings Central Components

- Two Windows desktop apps for submitting events
  - Zermelo: Manages all aspects of running a table tennis tournament.
  - Cantor: A much simpler app whose sole purpose is to submit events to Ratings Central.
- The ratings processor app that runs on my PC at home.
- The website (www.ratingscentral.com) where match results and ratings are displayed.

System Flow

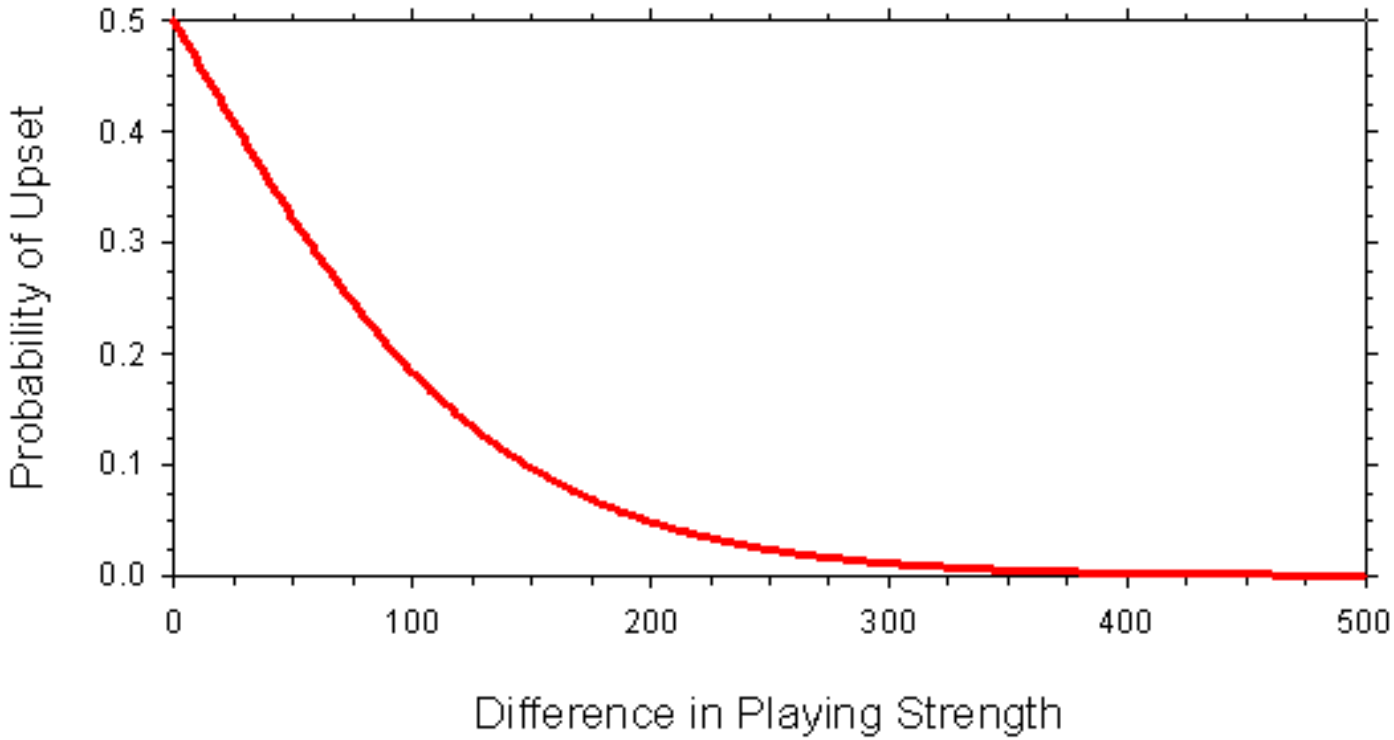
- Zermelo/Cantor submit events via email using a simple text format.
- The ratings processor retrieves the email, processes the event, and uploads the new ratings to the website.
  - This typically takes a few minutes from the time the director submits the event.
- Corrections are handled by resubmitting the event.
  - The system automatically reprocesses all affected events.
- Several event directors bypass Zermelo/Cantor and submit directly into the system using their own software.

Who is Using

- Currently, the system contains
  - 8,042 events
  - 37,142 players
  - 755,958 matches
  - 464 clubs
  - 127 event directors
- Many clubs in the U.S. submit leagues or tournaments that they run.
- Ratings Central provides official ratings/rankings for
  - Austrian Table Tennis Association, www.oettv.org
  - Lower Austrian Table Tennis Association, www.noettv.org (state in Austria)
  - Salzburg Table Tennis Association, www.sttv.co.at (state in Austria)
  - Table Tennis Queensland, www.tabletennisqld.org (state in Australia)
  - National Collegiate Table Tennis Association, www.nctta.org

The Bayesian Model

- Each player has a playing strength, i.e., a number that quantifies how strong the player is.
- Define the *probability-of-upset function* $\pi(x) := 1/(1 + e^{x/67})$:



- The probability that a player with playing strength s will upset a player with playing strength $t > s$ is $\pi(t - s)$.

The Bayesian Model (cont.)

- A player's playing strength is not known, so model it as a random variable (the *player's law*) with a normal prior.
- The temporal update models a player's playing strength changing with time:
  - Add a zero-mean normal random walk to the player's law with a standard deviation of 70 rating points per year.

Intractability of Direct Calculation of Posterior

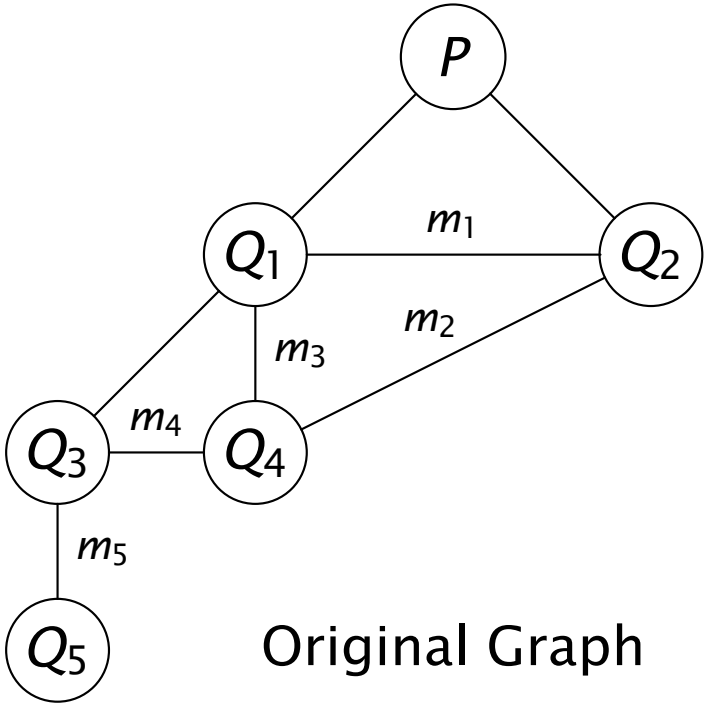
- Let N be the number of players in an event. Let L_j be the initial (i.e., start of event) law for player j.
- Let M be the number of matches in the event. Let $p(i)$ be the number of the player who wins the ith match. Let $q(i)$ be the number of the loser.
- Define

$$U(x_1) := \int_{\mathbf{R}^{N-1}} \prod_{i=1}^{M} \pi(x_{q(i)} - x_{p(i)})\, dL_2(x_2) \cdots dL_N(x_N) L_1(x_1).$$

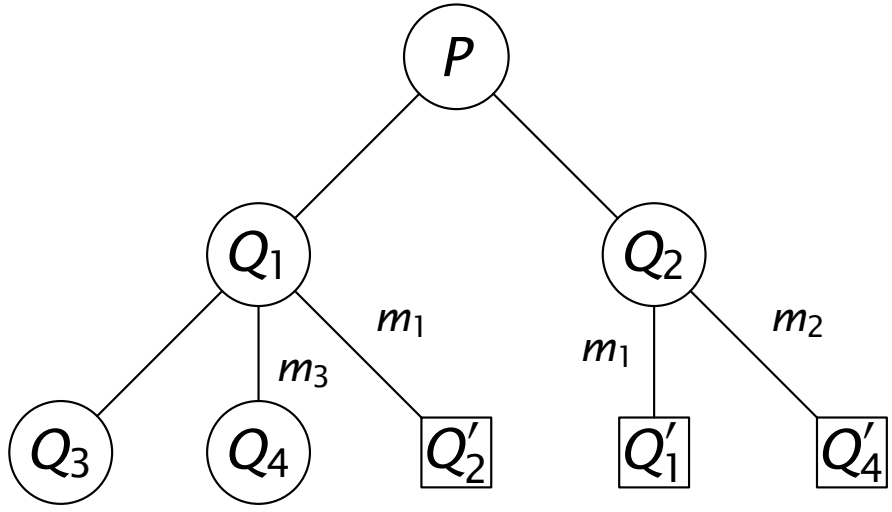
- The posterior law for player 1 is $U / \int_{\mathbf{R}} dU$.
- For the values of N (up to 1000) and M (up to 3000) that we have, it is not feasible to calculate this directly.

Tournament Surgery

- For an event, consider the graph where each player is a node and each match is an edge connecting two players.
- For each player P, construct a modified graph (as explained on the following slides) and use the modified graph to calculate P's posterior law.



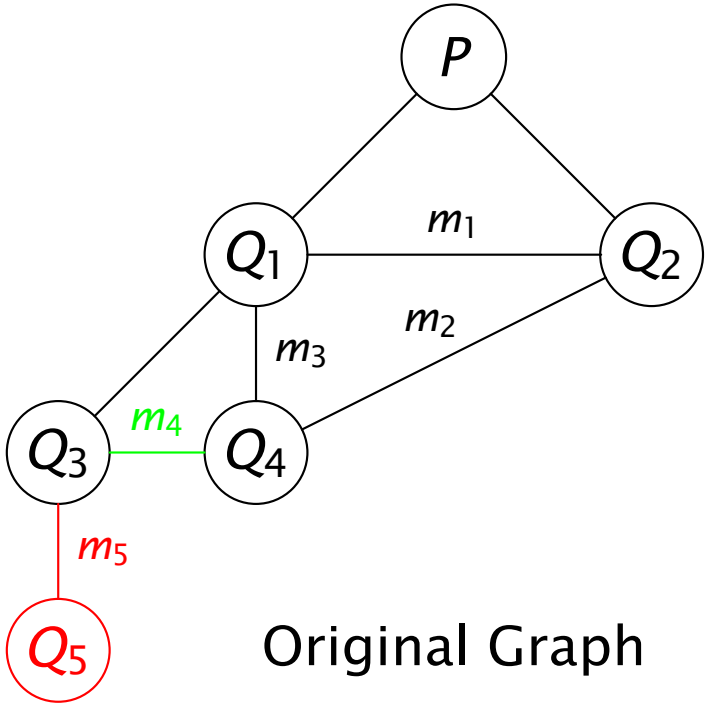
Original Graph



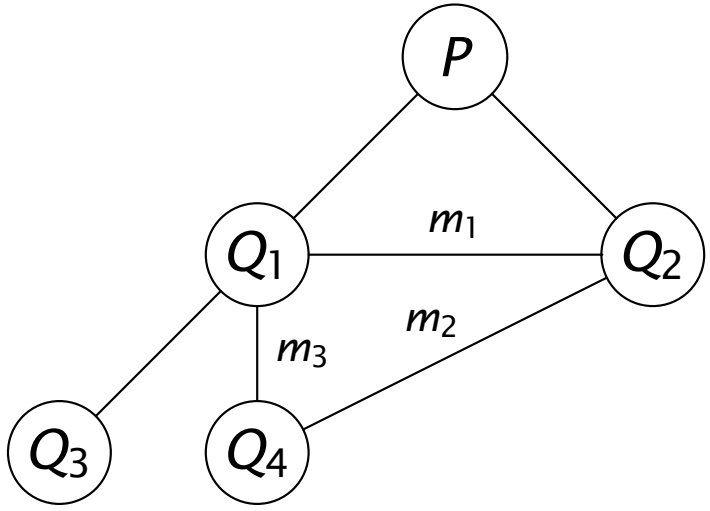
Modified Graph

Tournament Surgery

Steps 1–2



Original Graph

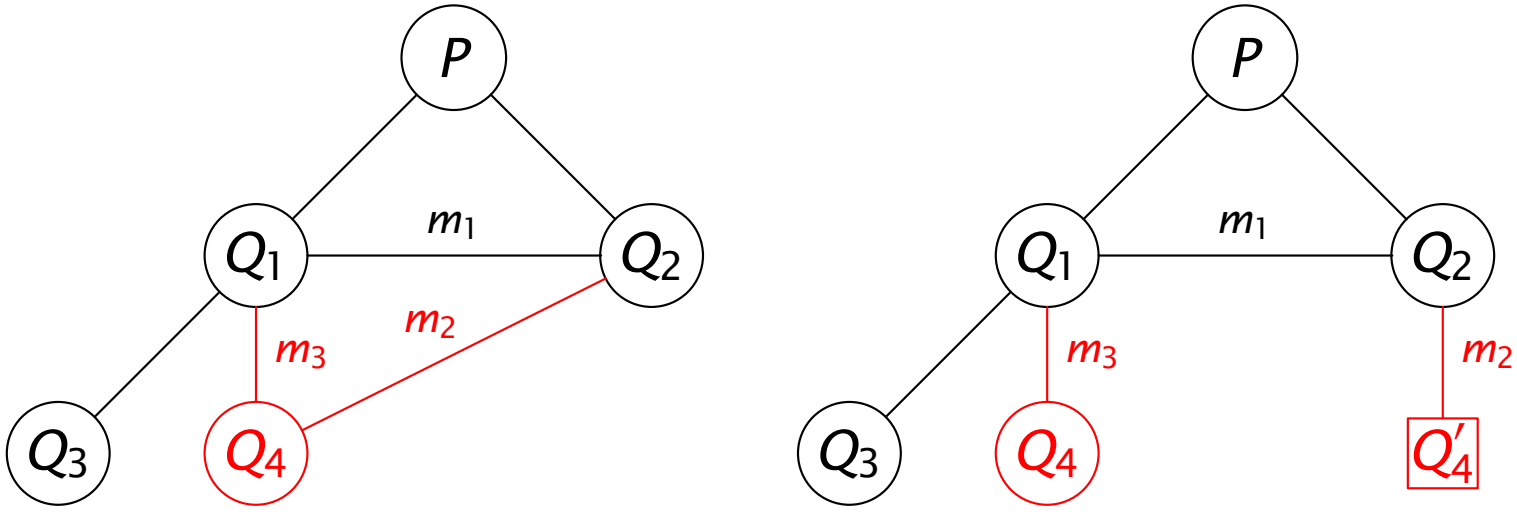


After Step 2

- Discard all edges that extend down from the second level of opponents (e.g., match m_5).
- Discard all edges that connect two nodes at the second level (e.g., m_4).

Tournament Surgery

Step 3

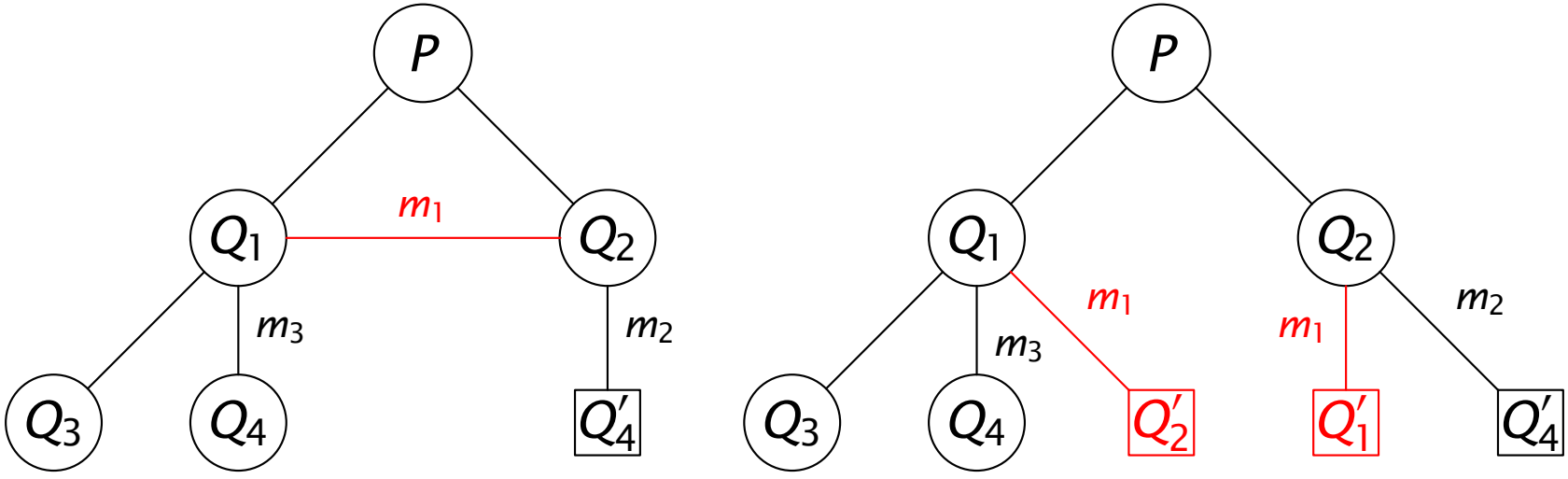


Before Step 3 After Step 3

- If a node at the second level (e.g., Q_4) connects to two or more nodes at the first level, add twins of the node (e.g., Q_4') and connect each of the nodes at the first level to exactly one of the twins.
- A *twin* is a new player who has the same initial law as the original player.

Tournament Surgery

Final Step



Before Final Step

Final Graph

- For each edge connecting two nodes at the first level of opponents (e.g., m_1), cut the edge and insert two new nodes on the newly created ends.
- The new nodes (e.g., Q'_2, Q'_1) are twins of the original nodes.

Algorithm Intuition

- When I go to a tournament and play somebody, I'm usually interested in how good my opponent is (especially if I just lost to them).
- I can look up their rating at the start of the tournament, but their rating may be out of date or they may be playing better or worse than their rating would indicate.
- So, I go to the posted draw sheets and find the other matches my opponent has played in the current tournament and see how they've done in those matches.
- These are the same matches that the rating system looks at for each player.

Priors

Elicitation

- Event directors must set a prior (i.e., mean and standard deviation) for the event.
- They may also set a prior for individual players, but are not required to.
- For a given unrated player, the system uses the player prior if it is set, otherwise, it uses the event prior.
- The full instructions given to the event directors for setting priors are at www.ratingscentral.com/UnratedPlayers.php.

Priors

Player Priors

- Excerpts from the instructions for player priors:
  - The prior standard deviation for a player measures how sure you are that you know that player's playing strength.
  - Here are some very rough guidelines: If you know an unrated player extremely well (e.g., they play at your club every week), then you might use a prior standard deviation of 50–75. . . .

Priors

Event Priors

- Excerpts from the instructions for event priors:
  - It is best to interpret the event prior mean and standard deviation as describing the range of unrated players at your event.
  - For example, if you think the unrated players range from 800 to 1400, then you would use the average of these two values (i.e., 1100) as the mean and the difference of these two values divided by four (i.e., 150) as the standard deviation.
  - . . . you should interpret the range as being plus or minus two standard deviations, not three.
- Initially, I tried to interpret the range as being plus or minus three standard deviations.
  - This always produced standard deviations that were too small.

Confusions and Misconceptions

My Rating is Lower

- In 1999, when we originally announced the proposed new system to the USATT membership, we posted on the Web all the tournament results in the U.S. for 5½ years (15,549 players, 330,079 matches) with both the USATT ratings and the ratings calculated by the new system.
- Generally, if a player's rating in the new system was higher than in the USATT system, the player liked the new system. If it was lower, they did not like it.
- Getting tired of illogical complaints, we decided to raise all the ratings in the new system by 100 points.
  - This significantly decreased the number of complaints.
  - Ratings are relative, so the change made no real difference.

Confusions and Misconceptions

Standard Deviation Measures Consistency

- Quite a few people think the standard deviation measures how consistent the player is.
- It doesn't. There is nothing in the model that measures consistency.
- People aren't familiar with using probability to model uncertainty.

Confusions and Misconceptions

Standard Deviation Should Increase for Unexpected Results

- Sometimes people suggest that the standard deviation should increase if a player's results are inconsistent with their rating.
- Suppose a 2000 player loses to five 1500 players and beats five 2500 players.
- These are certainly unexpected results. But, the model says every player has a playing strength, so the system must pick a rating for the player.
- 2000 is the most reasonable rating for such a player. Any other rating would make the results even more unlikely.
- The standard deviation always decreases when a player plays a match.

Problems

Austrian Junior Tournaments: Background

- In Austria, they submit tournaments where the best juniors from all of Austria compete.
  - These occur 4–8 times a year.
  - Only the best juniors in each category are eligible; these players usually improve rapidly.
- The format is round robin groups of 10–15 players.
  - 10% to 15% of the players in a group are from Lower Austria and so play in many leagues and tournaments that are submitted to the system.
  - For the rest of the players, the only events they play in that are currently submitted are these relatively infrequent tournaments.

Problems

Austrian Junior Tournaments: Problem and Solution

- Going into such an event, the Lower Austria players are accurately rated, while many of the other players have significantly improved since the last event they played in that was submitted to the system.
- The result is that Lower Austria players almost always lose points in these events.
- The solution we are using is to increase the temporal update to 200 points per year (instead of the usual 70) when calculating the initial ratings for the players in these events.
  - The result is the system gives less weight to the initial ratings of the non-Lower-Austria players.

Problems

Jumps

- In rare cases, players change playing strength very rapidly, e.g., 250 points in a month. Effectively, they wake up one day and they are much better.
- The system won't react this rapidly, so the player's rating lags their level for a while.
- A possible solution is to modify the temporal update to include a small probability of a jump of say ± 250 points.
  - The probability of the jump could be something like $1 - e^{aD}$ where $a < 0$ is a parameter to be determined and D is the number of days since the player's last event.
- I haven't tried this.

Reference

Marcus, D. J. (2001),
New Table-Tennis Rating System.
Journal of the Royal Statistical Society: Series D (The Statistician), 50: 191–208. doi: 10.1111/1467-9884.00271
www.ratingscentral.com/Doc/NewTTRS.pdf