The Star Fleet Battles Tournament Report Version 1.0

Robert W. Schirmer
Battlegroup Baltimore
Ellicott City, Maryland USA

5 September 2011

1 Introduction

The Star Fleet Battles Tournament Report is intended primarily for dedicated fans of the Star Fleet Battles (SFB) tournament game. It addresses two questions of perennial interest: how good is each tournament player, and how well are the tournament cruisers (TCs) balanced against one another? To answer these questions, the report presents a statistical analysis of the results of approximately 5000 sanctioned tournament matches. The match results are drawn from Star Fleet Battles Online (SFBOL), Captain's Log (CL), and tournament after-action reports on the ADB bulletin board. The main objectives of the analysis are (i) to tabulate and present data on tournament match results, broken down by player and ship; (ii) to quantify the relative skill levels of the players; and (iii) to quantify the balance between the various TCs.

2 Overview and Results

This Section briefly summarizes the method used for the analysis, and then moves straight to presenting the results. The full details of the calculations, which are not likely to be of general interest, are presented in Section 3.

2.1 Overview of Method

All match results in the database contain complete information: the winning player's name, the winner's TC, the losing player's name, and the loser's TC. Tabulation of game counts, win-loss records, and win percentages is thus a straightforward exercise in counting.

The quantification of player skill and ship balance is based on the assignment of numerical ratings. Each player is assigned a numerical rating that reflects their skill, the larger the number the better the player. The average player rating is arbitrarily set at 2000. Similarly, each TC pair is assigned a rating difference that reflects their relative capability, as determined by their ship systems displays (SSDs) and the rules of SFB. This rating difference corresponds to what players usually call the rock-paper-scissors (RPS) advantage of one of the ships in the pair over the other. A balanced pair of TCs has a rating difference of +0. Non-zero values indicate some level of advantage for one of the TCs in the pair.

The probability of winning or losing a match is then modeled as depending only on the difference in the ratings of the players and ships involved. Table 1 shows some rating differences and the corresponding win percentages.[1] As an illustrative example, assume Player α is rated $R_\alpha = 2000$ and Player β is rated $R_\beta = 2400$. Assume also that Player α selects Ship γ, Player β selects Ship δ, and that for this pair of ships the rating of γ vs. δ is $S_{\gamma\delta} = +200$ (that is, Ship γ is advantaged over δ by +200 rating points). Then for this match, the chance of Player α winning in Ship γ is 38%, as determined by the rating difference $(R_\alpha - R_\beta + S_{\gamma\delta}) = (2000 - 2400 + 200) = -200$.[2]

[1] The function that maps rating differences to win percentages is discussed later. Its exact form is not particularly important.

[2] From the standpoint of Player β, the chance of winning is determined by $(R_\beta - R_\alpha + S_{\delta\gamma}) = (2400 - 2000 - 200) = +200$, which yields the expected 62% chance of victory. Note that the definition of the rating difference for a ship pair requires that $S_{\gamma\delta} = -S_{\delta\gamma}$ for any two Ships γ and δ.

Rating Difference	Probability of Winning
+1000	0.92
+800	0.88
+600	0.82
+400	0.73
+200	0.62
0	0.50
-200	0.38
-400	0.27
-600	0.18
-800	0.12
-1000	0.08

Table 1: Sample rating differences and win probabilities.

The approach taken here is to calculate a maximum probability estimate (MPE) for the ratings of all the players and TC pairs, given the available match results. That is, the analysis finds the ratings for all of the players and ship pairs that best explain the actual, known match outcomes, given the probabilistic connection between ratings and match outcomes illustrated in Table 1. To put it yet another way, the analysis finds the ratings that maximize the probability of obtaining the match results in the database.

Two variants of the general approach described above are also implemented. The variant results are intended for comparison with those of the main analysis. In the first variant, an MPE for the player ratings is calculated under the assumption that all of the TCs are balanced with one another (i. e. +0 rating differences are assumed for all TC pairs). In the second variant, an MPE for player and ship ratings is calculated using only Ace level players and matches between two Ace level players. Ace level players are taken to be those with ratings of 2300 or more, as determined by the results of the first variant calculation.

ADB modified some of the TCs over the time frame during which the match data for this report was collected. No attempt has been made to break the data down by individual TC subvariants; for example, matches involving the WBS with and without drone upgrade points are all grouped together and treated equally under the WBS category. This obviously introduces a source of error into the analysis. However, most of the TC changes have been minor, and it seems likely that other sources of error are far more important than the design variations. The exception to this is the AND. Almost all of the AND data in this report is for the previous, unbalanced, multi-Origins winning version, and should not be applied to the current, much weakened AND design, for which there is very little data at all.

2.2 Overview of Results

The results are presented in Appendices A, B, and C. Appendix A contains the results of the general analysis for all players and ships. Appendix B contains the results for players only, given the assumption that all TCs are balanced. Appendix C contains the results for Ace level players and Ace vs. Ace matches. Again, an Ace is a player rated 2300 or above in Appendix B. The format of the tables in all Appendices is described in the following paragraphs.

The results include ratings for each player and TC pair, plus various pieces of statistical data, for example win-loss records, etc. Table 2 gives the three-letter identifiers for each of the sanctioned TCs.

Appendix Tables 1 through 7 show the results of the analysis for all SFBOL players. Each entry shows the player's rank, SFBOL callsign, total number of matches played, win-loss record, measured win percentage, MPE predicted win-loss record, MPE predicted win percentage, rating from the MPE calculation, and the corresponding uncertainty in the MPE rating (one standard deviation). Player rankings go from best to worst, based on rating. The measured win percentage is just the number of wins divided by the total number of games played. The MPE predicted win-loss record uses the MPE rating differences for a player's matches to calculate the player's expected (average) win-loss record across those matches. Thus it incorporates both player and ship ratings. The MPE predicted win percentage is just the MPE predicted wins divided by total games.

Appendix Table 8 shows summary results for all of the sanctioned TCs. Each entry shows the ship's ranking, three-letter identifier, frequency of appearance, number of match slots occupied (listed as the number of games played), win-loss record, measured frequency weighted win percentage, MPE predicted win-loss record, MPE fre-

Identifier	TC Name
AND	Andromedan Krait
FED	Federation
GRN	Gorn
HYD	Hydran Lord Marshal
ISC	ISC
KLI	Klingon D7CT
LDR	LDR Red Jaguar
LYR	Lyran White Tiger
ORI	Orion
RFH	Romulan Firehawk
RKE	Romulan King Eagle
RKR	Romulan KR
SEL	Seltorian
THA	Archeo-Tholian
THN	Neo-Tholian
WAX	Wyn Auxiliary Battlecruiser
WBS	Wyn Great Black Shark
ZIN	Kzinti

Table 2: Tournament cruiser 3-letter identifiers.

quency weighted win percentage with equal players, and the uncertainty in the MPE frequency weighted win percentage with equal players (one standard deviation).

Frequency of appearance is given by the number of match slots the TC appears in divided by the total number of match slots. Each match has two slots for TCs. Ship rankings go from best to worst, based on the MPE frequency weighted win percentage. The measured frequency weighted win percentage is the average of the measured win percentages against each individual TC type, weighted by the frequency of appearance of that TC type. The MPE predicted win-loss record uses the MPE rating differences for a ship's matches to calculate the ship's expected (average) win-loss record across those matches. Thus, it incorporates both player and ship ratings. The MPE frequency weighted win percentage with equal players is the average of the MPE win percentages against each individual TC type, weighted by the frequency of appearance of that TC type.

The MPE win percentage for a given TC pair is calculated from the estimated rating for the TC pair, assuming equal players. Thus, the MPE frequency weighted win percentage shows the percentage of games the TC is expected to win given equally skilled players, and given opposing TC types distributed per their observed frequencies of appearance. It is therefore a measure of the overall RPS strength of each TC.

Appendix Tables 9 through 26 contain detailed information for each TC. Each table breaks out the data for each opposing TC type. The columns give the opposing TC's three-letter identifier, the number of matches played against it, the win-loss record, the measured win percentage, the MPE predicted win-loss record, the MPE predicted win percentage, the MPE rating for the TC pair, the uncertainty in the MPE rating (one standard deviation), the win percentage for equal players corresponding to the MPE rating, and the uncertainty in the win percentage for equal players corresponding to the uncertainty in the MPE rating.

Appendix Table 27 is an RPS grid in ADB bulletin board format. Each entry shows the expected wins per 10 games for the row ship type vs. the column ship type. All entries are based on the MPE rating for the ship pair, and assume equal player skill.

3 Detailed Methodology

Section 3 presents a detailed description of the method used for the player and ship rating MPE, as well as the underlying assumptions and shortcomings of the analysis.

3.1 Motivation

The simplest approach to evaluating player skill and TC balance is to calculate win percentages for each player and for each TC pairing. This method provides a reasonable first estimate, but has significant shortcomings. For example, it is entirely possible for an average player to have a high win percentage that derives from numerous victories against rookies, interspersed with a few losses to experts. Win percentage does little to distinguish the average from the expert in such a case. Similarly, a strong player facing a weak player is likely to win regardless of the TCs being used, so such matches say little about TC balance. The problems with using win percentages can be mitigated by tracking not just wins and losses, but the quality of the opposition. That is essentially the approach taken here: player and ship strengths are modeled with a minimal set of numerical ratings, and the most probable assignment of ratings is derived from the available match results. The resulting player ratings factor in not only win percentages, but the quality of the opposing players, and the quality of the ships used. The resulting ship ratings directly reflect the RPS effect, that is, how balanced a pair of ships are when played by equally skilled players.

3.2 Player Strength Model

The strength of each player is represented by a single, numerical rating score that measures the player's skill level. The larger the number, the more skilled the player. The probability of one player defeating another is determined by the difference in their ratings, given balanced TCs. For historical reasons, 2000 is selected as the average *a priori* player rating; because only rating differences matter for the probability calculations, the selection of 2000 is arbitrary and has no impact on the results of the analysis.

The rating score represents the player's proficiency with the strategies, tactics, techniques, and procedures (STTP) one must grasp in order to be good at tournament SFB. Examples include, but are not limited to, knowing how and when to use the plasma ballet, reserve power, speed changes, weasels, Mizia attacks, and cloaking/subhunting.

Using a single rating for each player obviously neglects several effects. Two of the more significant neglected factors merit some discussion. First, a player can have different levels of skill and experience with each TC. It would therefore be more realistic to describe each player with eighteen ratings, one per TC, rather than a single, overall rating. Nevertheless, an argument can be made that a single rating score provides a reasonably good measure of a player's skill. The argument is that a great part of the STTP, the core, is either used or frequently encountered by every TC. A player strong in the core STTP will perform well, and a player weak in the core STTP will not, regardless of TC choice. The greater part of a player's ability is tied to proficiency with the core STTP, and this can be captured by one rating number. In this view, specialization in particular TCs is a secondary effect that provides a moderate correction to the core skill level when that particular TC is used. For ships where specialization is especially important (perhaps the THA or AND), the use of a single rating may obscure RPS effects.

Second, a player can improve over time, effectively moving from a lower to a higher rating.[3] Again, a single rating score cannot fully capture such an evolution. A player's rating at any point in time reflects his complete record. If a player improves steadily, his rating will lag below that appropriate for his current skill level, due to the influence of his early record, acquired when he was less skillful. However, skill at SFB does not increase indefinitely. The tendency is for a rapid increase as one moves from the beginner level to proficiency, followed by a leveling off. Therefore, over time, a player's rating should catch up to the correct number, with an ever more insignificant undervaluing due to the early record.

3.3 Ship Strength Model

The strength of each TC is represented by a set of eighteen numerical rating scores, one for each type of potential opposing TC. The rating score indicates the advantage or disadvantage the TC has in each matchup. The probability of one ship defeating another can be calculated from the rating difference for the pair, assuming equal players. A rating of +0 is selected as the average, *a priori* ship rating. A positive rating indicates advantage, the larger the better, and a negative rating represents disadvantage. Again, only rating differences matter, so these selections are arbitrary and have no impact on the analysis.

This ship strength model involves the minimum number of variables needed to address the question of TC balance. Two TCs are balanced if the pair has a rating difference of +0. The rating difference represents the relative capabilities of the ships as reflected by their SSDs and the rules of SFB. Note that these ratings subsume certain random elements in SFB. Also, note that there are only 153 independent ship ratings. Specifically, given 18 TCs, $[18^2 - 18]/2 = 153$, where each TC is obviously balanced against itself, and the rating difference S_{ij} in a match of i vs. j determines the rating difference S_{ji} in a match of j vs. i by $S_{ij} = -S_{ji}$.

[3] Obviously, the reverse is also true, for example when a player leaves the game for a long period of time and then returns only to find that they have become rusty.

3.4 Model Complexity

There is a general reason to keep the player and ship models as simple as possible: adding variables rapidly increases the number of match results needed in order to draw conclusions from the analysis. To put it another way, one can add numerous variables to the model, but then be unable to say much of anything about them due to an insufficient quantity of data. It is therefore desirable to use the simplest possible model that captures the major effects, which in this report is taken to be one rating score per player and 153 independent TC pair rating differences.

3.5 Probability Distributions

3.5.1 Joint Distribution

The joint probability distribution of interest is

$$P(M, S, R), \tag{1}$$

where M is the set of tournament match results, S is the set of ship ratings, and R is the set of player ratings. Distribution P relates the player and ship ratings to the probability of the match outcomes.

Given N_P players in the database, then $R = \{R_1, R_2, ..., R_{N_P}\}$ is the set of their ratings, where R_k is the rating of the k^{th} player. Similarly, given N_S sanctioned tournament ships in the database, then S is an $N_S \times N_S$ matrix of ratings for each ship pair,

$$S = \{S_{11}, S_{12}, ..., S_{1N_S}, S_{21}, S_{22}, ..., S_{2N_S}, ..., S_{N_S1}, ..., S_{N_SN_S}\}, \tag{2}$$

where S_{ij} is the rating of the i^{th} ship against the j^{th} ship. Here, $N_S = 18$. Note that $S_{ij} = -S_{ji}$, and $S_{ii} = 0$, for all i and j. Lastly, given N_M match results in the database, then $M = \{M_1, M_2, ..., M_{N_M}\}$, where M_l is the l^{th} match result. Each match result M_l is of the form: Player α in Ship γ defeats Player β in Ship δ, for some combination of players and ships. The two possible outcomes of a given match, thought of as a random trial, are of course victory or defeat for a particular player and ship.

Per the standard Bayesian formulation, P factors into its conditional distributions, giving

$$P(S, R|M) = \frac{P(M|S, R)P(S|R)P(R)}{P(M)}. \tag{3}$$

Ship ratings, as previously defined, represent the strength of the ship's SSD under the SFB rules. They are assumed to be independent of the player ratings, so $P(S|R) = P(S)$. The probability of a match outcome depends only on the ratings of the players and ships involved in that match. Therefore, $P(M|S, R)$ factors into a product of probabilities over individual match results, each conditioned on only the ratings of the players and ships involved. The player ratings in $P(R)$ are independent of one another, so $P(R)$ factors into products of probabilities over individual player ratings. Similarly, the ratings of all unique ship pairings in $P(S)$ are independent of one another, so $P(S)$ factors into a product of probabilities over unique ship pairings. Under these assumptions, Equation 3 becomes

$$P(S, R|M) = \frac{\prod_{l=1}^{N_M} P(M_l|S_{\gamma\delta}, R_\alpha, R_\beta) \prod_{i=2}^{N_S} \prod_{j=1}^{i-1} P(S_{ij}) \prod_{k=1}^{N_P} P(R_k)}{P(M)}, \tag{4}$$

where Player α in Ship γ plays Player β in Ship δ in match M_l. The conditional distributions on the right-hand side of Equation 4 must now be specified.

3.5.2 Match Result Conditional Distribution

The conditional distribution for a match result needs to satisfy several criteria, but otherwise its specific form is not particularly important. The criteria include the usual restriction on probability values,

$$0 \leq P(M_l|S_{\gamma\delta}, R_\alpha, R_\beta) \leq 1; \tag{5}$$

that the probability be a function only of the rating difference between the players and ships involved,

$$P(M_l|S_{\gamma\delta}, R_\alpha, R_\beta) = P(M_l|\Delta), \tag{6}$$

where

$$\Delta = R_\alpha - R_\beta + S_{\gamma\delta} \tag{7}$$

is the rating difference for the match; that the probability of Player α winning in Ship γ goes to 100% as the rating difference gets large and positive,

$$\lim_{\Delta\to\infty} P(M_l = win|\Delta) = 1; \tag{8}$$

that the probability of Player α winning in Ship γ goes to 0% as the rating difference gets large and negative,

$$\lim_{\Delta\to-\infty} P(M_l = win|\Delta) = 0; \tag{9}$$

and that the probability of Player α winning in Ship γ increases monotonically with the rating difference,

$$\frac{dP(M_l = win|\Delta)}{d\Delta} > 0. \tag{10}$$

In the above equations, $M_l = win$ means Player α in Ship γ defeats Player β in Ship δ.

For the purposes of this study, the probability distribution for a match result, given the ratings of the players and ships, is taken to be

$$P(M_l = win|S_{\gamma\delta}, R_\alpha, R_\beta) = \frac{1}{1 + e^{-(R_\alpha - R_\beta + S_{\gamma\delta})/\sigma}}, \tag{11}$$

and the scale is set to $\sigma = 400$. Normalization requires that

$$P(M_l = loss|S_{\gamma\delta}, R_\alpha, R_\beta) = 1 - P(M_l = win|S_{\gamma\delta}, R_\alpha, R_\beta), \tag{12}$$

where $M_l = loss$ means Player α in Ship γ loses to Player β in Ship δ. The distributions of Equations 11-12 satisfy the criteria of Equations 5-10. For example, it can be seen that the probability of victory in a given match depends only on difference between the ratings of the players and ships involved. If the rating difference is zero, then the probability of winning is 50%; if the rating difference is +400, then the advantaged player/ship combination has a 73% chance of winning. See Figure 1.

3.5.3 Player Strength Prior Distribution

The *a priori*, or prior, player strength distribution specifies how likely a rating value is thought to be in the absence of any match data. In this analysis, the prior is selected on the basis of three ideas: (1) that the average prior rating should be 2000, for historical reasons; (2) that a win probability range of 10-90% likely captures most of the spread of player skill in SFB; and (3) that the prior should be symmetrical about 2000. Note that the prior is significant in determining a player's most likely rating only when the player has few or no match results. For a player with a large number of matches, the prior is much less significant. The *a priori* distribution for each player's rating is taken to be

$$P(R_k) = \frac{1}{\sigma(1 + e^{-(R_k - 2000)/\sigma})(1 + e^{-(2000 - R_k)/\sigma})} \tag{13}$$

for all k. In Equation 13, $1/\sigma$ is a normalization constant whose value ensures that

$$\sum_{R_k} P(R_k) = 1, \tag{14}$$

where the sum is over all possible values of R_k. Figure 2 shows the prior distribution. There is an 85% chance that a player's rating falls between 1000 and 3000. Note also that a 3000 rated player has a 92% chance of defeating a 2000 rated player. Therefore, the prior distribution is not particularly restrictive, i. e. it allows for a significant spread of player skill levels. The prior may also be thought of as the distribution resulting from (1) starting with the assumption that a player's rating can have any value whatsoever, with equal likelihood; then (2) assuming each player has one win and one loss against a 2000-rated player using balanced TCs; and finally (3) calculating the likelihood of the player having rating R_k given assumption (1) and match results (2). The result is Equation 13.[4]

[4] This interpretation also gives the prior as being equivalent to two match results. For players with many more than two matches, the prior is relatively insignificant in determining the final, most probable rating.

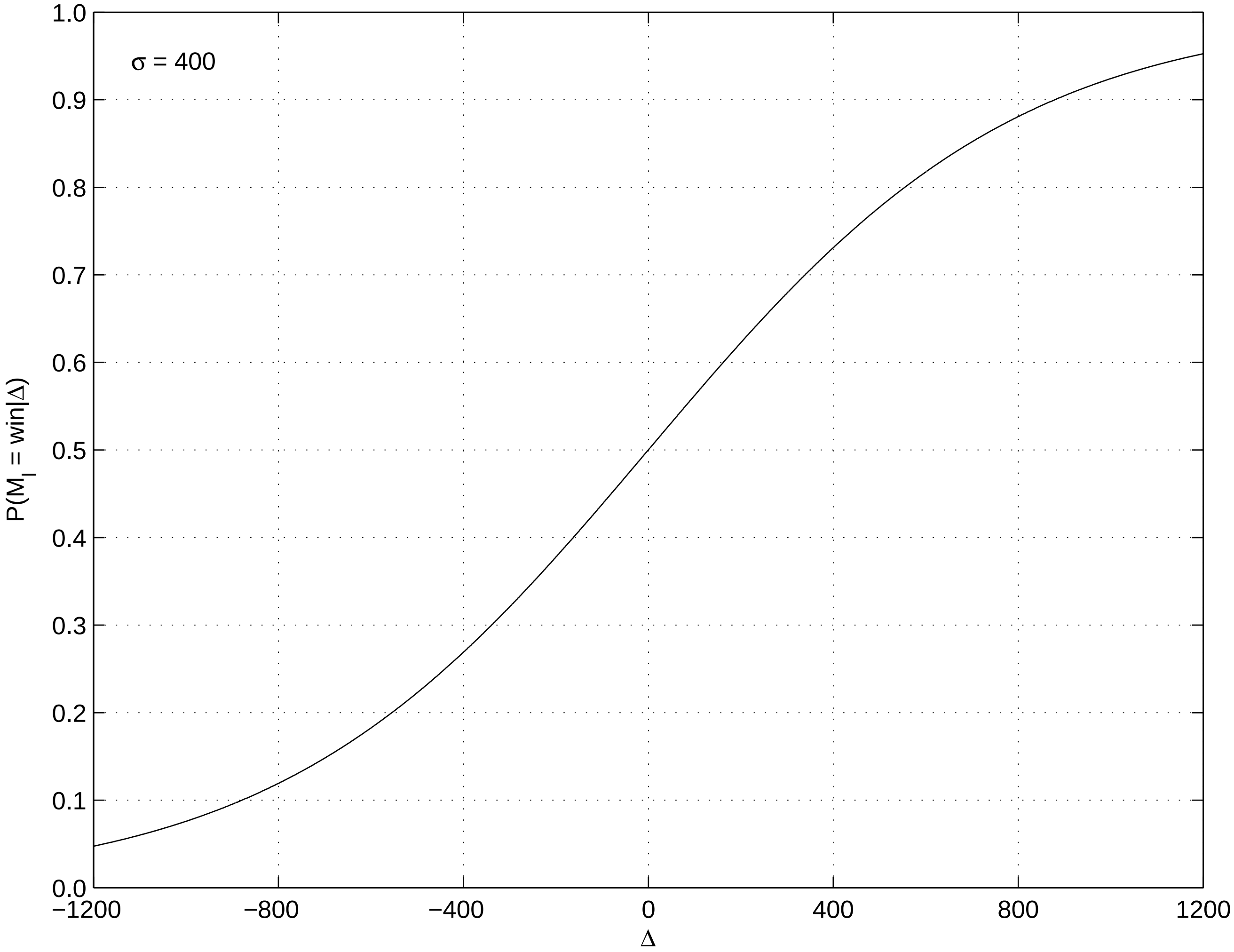

Figure 1: Plot showing probability of Player α in Ship γ winning the match vs. rating difference Δ. Values of Δ shown here range over $\pm 3\sigma$.

3.5.4 Ship Strength Prior Distribution

The ideas underlying the selection of a ship strength prior are identical to those underlying the prior of Section 3.5.3, save that the average prior rating is zero for a TC. The *a priori* distribution for ship strength is taken to be

$$P(S_{ij}) = \frac{1}{\sigma(1+e^{-S_{ij}/\sigma})(1+e^{S_{ij}/\sigma})} \tag{15}$$

for all i, j. In Equation 15, $1/\sigma$ is a normalization constant whose value ensures that

$$\sum_{S_{ij}} P(S_{ij}) = 1, \tag{16}$$

where the sum is over all possible values of S_{ij}. Figure 3 shows the prior distribution. There is an 85% chance that a ship's rating falls between -1000 and 1000. As with the player strength prior described above, the ship strength prior may be thought of as the distribution resulting from (1) starting with the assumption that a ship's rating can have any value, with equal likelihood; then (2) assuming each ship has one win and one loss against each opposing ship, with equal players; and finally (3) calculating the likelihood of the ship having rating S_{ij} given assumption (1) and match results (2). Additional justification for using the prior of Equation 15 comes from the fact that, even absent match results, inspection of the SSDs indicates that the tournament cruisers should be nearly balanced with one another due to their similar shielding, internals, power curves, numbers of heavy weapons, etc. Actually, given

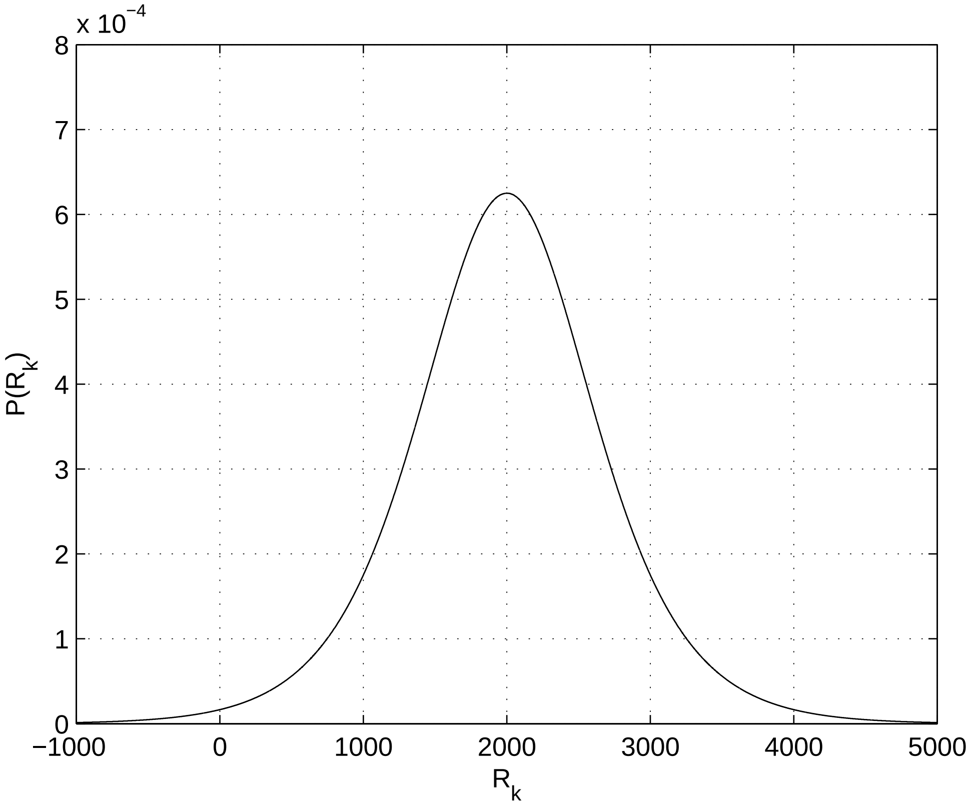

Figure 2: Plot shows the prior probability distribution for a player's rating, $P(R_k)$, as a function of rating R_k, in one rating point bins. There is an 85% chance that the player's rating falls between 1000 and 3000 in this distribution.

the effort made by ADB to balance the TCs by design, a narrower *a priori* distribution of S_{ij} than that used here could be justified.

3.6 Analysis

The objective here is to find the maximum probability estimate (MPE) for the ship and player ratings, given the match data and the probability model described in Section 3.5. In other words, given the results of all of the matches in the database, what are the most probable ratings of all the ships and players?

In mathematical terms, finding the MPE for the ratings is equivalent to maximizing the left-hand side of Equation 4, $P(S, R|M)$, which is exactly the probability of a set of player and ship ratings, given the match data. It turns out to be more convenient to work with the natural logarithm of the probability $L = \log P(S, R|M)$, which per Equation 4 is given by

$$L = \sum_{l=1}^{N_M} \log P(M_l|S_{\gamma\delta}, R_\alpha, R_\beta) + \sum_{i=2}^{N_S}\sum_{j=1}^{i-1} \log P(S_{ij}) + \sum_{k=1}^{N_P} \log P(R_k) - \log P(M). \tag{17}$$

Maximizing L is equivalent to maximizing $P(S, R|M)$. The maximization is done over all possible assignments of ratings R and S, for a fixed set of match results M. Note also that, because $P(M)$ is independent of S and R, it is just a normalization constant defined by the requirement that

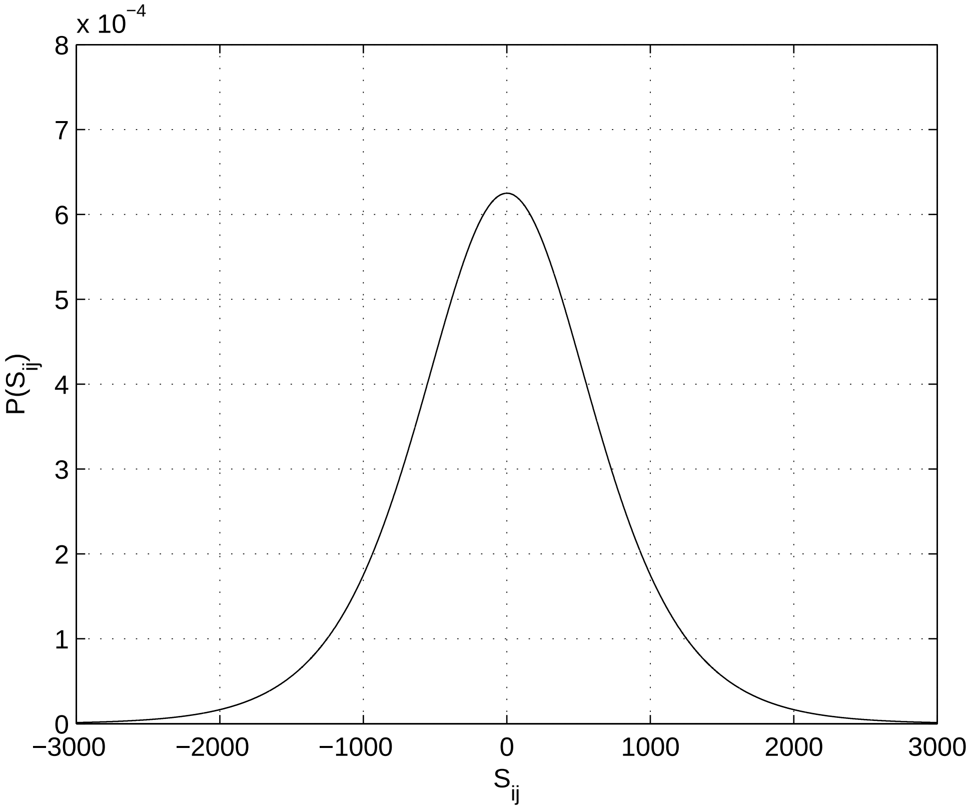

Figure 3: Plot shows the prior probability distribution for a ship's rating, $P(S_{ij})$, as a function of the rating S_{ij}, in one rating point bins. There is an 85% chance that the ship's rating falls between -1000 and 1000 in this distribution.

$$\sum_{S,R} P(S, R|M) = 1, \tag{18}$$

where the sum is over all possible rating assignments S and R. It therefore plays no role in the maximization process.

The task of determining the MPE for the ratings thus reduces to finding $max(L)$ as a function of S and R, where the match results M are known and fixed. This is done numerically using a gradient search method. That is, we find ratings $\bar{S}$ and $\bar{R}$ such that $\nabla L(\bar{S}, \bar{R}) = 0$, where $\bar{S}$ and $\bar{R}$ are then the most probable rating assignments. The uncertainties in the rating assignments (variances) may be derived from $\nabla\nabla L(\bar{S}, \bar{R})$, which is the matrix of mixed second partial derivatives of L with respect to the ratings. Specifically, the covariance matrix of the conditional distribution of the ratings given the match results is $\sigma^2_{ij} = -[\nabla\nabla L(\bar{S}, \bar{R})]^{-1}_{ij}$. The report tables quote the rating $\bar{R}_k$ for Player k and $\bar{S}_{ij}$ for Ship i vs. Ship j. The RPS value of the matchup Ship i vs. Ship j is defined by the chance of Ship i defeating Ship j in a match involving equal players. This is determined per Equation 11 by

$$P(M_l = win|\bar{S}_{ij}) = \frac{1}{1 + e^{-\bar{S}_{ij}/\sigma}}, \tag{19}$$

where $M_l = win$ means Ship i defeats j.