Data-Driven Player Ratings for Recreational Ultimate Frisbee Leagues

Shane Bussmann* and Geoa Geer+
* Understory Weather
+ Boston Ultimate Disc Alliance

Abstract: Close to 1.4 million people in the United States play the sport of ultimate frisbee at least 13 times annually. Many of these core participants play in recreational leagues that attempt to build balanced teams by assigning players according to skill level. Historically, skill level has been determined by each player's self-assessment. Such ratings tend to be biased and create significantly imbalanced teams. In this contribution, we use seven years of recreational ultimate frisbee league data from one of the largest ultimate frisbee organizations in the world, the Boston Ultimate Disc Alliance. We show that a rating system based on self-assessment regularly results in highly imbalanced teams. We introduce an alternative method for rating players based on their playing history, which builds a more accurate measure of their ability and a better forecast of their team's success. We provide guidelines for organizers to follow when incorporating this data-driven rating system into their recreational ultimate frisbee leagues.

1. Introduction

Almost 5 million people in the United States participate in the sport of Ultimate Frisbee. Of these, 1.3 million are core participants—meaning that they play at least 13 times each year. This is comparable to the sports of lacrosse (0.9 million) and rugby (0.4 million) combined [1]. Moreover, USA Ultimate—the premier organization for ultimate frisbee in the United States—has seen a 60% growth in youth participation over the past five years [2]. While this segment of the population is focused on highly competitive play, we expect it to serve as a positive indicator for future growth of the sport at all levels.

Ultimate is unique in its emphasis on fair play, which is codified directly in the rulebook "Spirit of the Game". From Section 1, Introduction, item B:

> Spirit of the Game. Ultimate relies upon a spirit of sportsmanship that places the responsibility for fair play on the player. Highly competitive play is encouraged, but never at the expense of mutual respect among competitors, adherence to the agreed upon rules, or the basic joy of play.

This commitment to fair play helps to explain why many core participants play in recreational leagues in which teams are assembled randomly according to skill level such that every team has an equal chance at winning—in theory. These leagues typically have 10 - 20 teams with 15 - 20 players on each team. They run throughout the year, with colder climates using indoor facilities during winter.

Another manifestation of Ultimate's "Spirit of the Game" is the emphasis on individual assessment when determining skill level of the participants in a recreational league. A robust measure of skill level is critical to creating balanced teams. Historically, teams in recreational leagues have been created by ranking players according to their self-assessed skill level and then assigning them to teams according to a snake-like draft system in which the order reverses with each round. The team that drafts first in round 1 drafts last in round 2, then first in round 3, etc.

One of the goals of this paper is to investigate the performance of self-assessment ratings relative to what is expected for teams of equal skill level. Our results show the poor performance of self-assessment ratings and have motivated us to undertake a second goal: identify a more effective rating system that builds teams based on the data of previous playing experience—we call this a "club rating".

We make use of seven years of data from the Boston Ultimate Disc Alliance (BUDA) to achieve our goals. BUDA is a non-profit organization run by volunteers to give people in the Boston area the opportunity to play Ultimate. BUDA organizes nearly 100% of the leagues in the greater Boston area. BUDA has collected data on who has played on what team and how that team performed during the season. This is the data that is used to compute club ratings in Section 4. BUDA also has stored self-assessment ratings and captain ratings over the years. The excellent BUDA database is what makes the research in this paper possible.

In this paper, we focus on data during the spring season. There are five seasons of BUDA recreational league play in a single year: one each in spring, summer, and fall, and two that take place indoors during the long New England winter. Spring is an ideal testbed for this analysis because the player pool exhibits the widest range of skill level compared to the other seasons. In summer and fall, the best players need practice time with their club teams, so they skip recreational league. In winter, the cost to play is higher and the distance to the facility is greater. Even more important, the playing field is a converted hockey rink—it is much smaller than a normal field for Ultimate and has walls along the edge that must be avoided when playing. For these reasons, only the most committed players continue to play in winter time. This skews the average skill level of winter leagues much higher than the outdoor leagues. We expect that the results of this paper will apply to the non-spring recreational leagues, but testing exactly how much so is a topic for future work.

BUDA has collected data since 2001. In this paper, we focus on the most recent seven years of data, from 2010 through 2016. Prior to 2010, the fraction of the player population in each league that had a captain rating was too low to provide a significant benefit compared to self-assessment ratings. Similarly, the data needed to compute a club rating for each player was too sparse prior to 2010. By focusing on data after 2010, we ensure that the captain ratings and club ratings are a fair representation of their ability to forecast future performance.

The outline of our paper is as follows. In Section 2, we quantify the performance of the existing BUDA rating system relative to expectations for equally skilled teams. Section 3 shows that most of the poor performance of the existing system is due to its reliance on self-assessment ratings. This is achieved via a comparison of the predictive power of self-assessment ratings and captain-assessment ratings regarding team performance. Section 4 describes the methodology we use to compute club-assessment ratings and shows that club-assessment ratings are more predictive of team performance than either self-assessment ratings or captain-assessment ratings. In Section 5, we present clear guidelines for league organizers to follow in order to make teams in recreational ultimate frisbee leagues as balanced as possible. Finally, in Section 6, we summarize our findings and highlight goals of future work.

2. Performance of Existing BUDA Rating System

To quantify the performance of the existing player rating system used by BUDA, we use points per game differential. There are many justifiable ways to rate team performance [3]. We choose points per game differential for two reasons. One is that these BUDA recreational leagues only have seven games per season, so using won-loss records leaves the sample size very small. There are typically 15-20 points scored per game, so using points per game differential increases the sample size significantly. The second reason is more philosophical in nature. For the purposes of a recreational league, it is our belief that a team that loses every game by only one point has a better experience than a team that wins one game out of seven but loses on average by four points.

Figure 1 shows the observed distribution (blue histogram) of points per game differential from seven years of BUDA spring recreational league data. The full dataset comprises 171 teams. Of these, 16 teams have points per game differential above +5 while 14 teams register below -5.

To understand how these numbers compare to expectations for a balanced league, we simulate games between equally skilled teams and monitor the simulated point differentials. BUDA spring league games last 70 minutes. Teams combine to score 18 points per game, on average. This implies a point-scoring rate of 0.26 points/minute, or 0.13 points/minute for a single team. BUDA does not collect data on when points are scored, so we assume that point-scoring in Ultimate follows a Poisson distribution, similar to soccer [4].

The simulation includes 171 teams, each of which has 7 games against an equally skilled team—i.e., both teams score points according to a Poisson distribution with an expectation value of 0.13 points/minute. The resulting distribution of points per game differential is represented by the grey histogram in the left panel of Figure 1. In the simulation, there are exactly zero teams with points per game differential above +5 or below -5. The simulation also results in fewer teams with points per game differential above +3 or below -3 compared to the observed BUDA distribution. In contrast, there are many more simulated teams with points per game differential close to zero compared to the observed BUDA distribution. This conclusively demonstrates that teams in BUDA recreational leagues are significantly imbalanced.

To give a concrete understanding of the skill disparity needed to lose by an average of 5 points per game, we include the right panel of Figure 1. This uses the same simulations referenced earlier in this section to show the percent chance that a given team has of winning a game versus a team against which it is evenly matched (blue curve) and against which it is expected to lose by 5 points per game (green curve). We simulate unbalanced teams by adjusting the scoring rate of the weaker team down by 3.5 points over 70 minutes (for a point scoring rate of 0.08 points/minute) and the stronger team up by 1.5 points over 70 minutes (for a point scoring rate of 0.15 points/minute).

Consider first the case of equally skilled teams. At the beginning of the game, the evenly matched teams both have a 44.7% chance of winning (chance of a tie is 10.6%). As the remaining time in the game decreases, the chance of a tie increases gradually. With 10 minutes remaining, the chance of a tie begins to increase dramatically, but each team remains equally likely to win.

Next, the unbalanced teams. The team that expects to lose by five starts the game with only an 8.5% chance at winning. By halftime, the weaker team expects to be down by 2.5 points, so the chance of winning has decayed to only 3.4%. With 10 minutes remaining, the chance of winning is negligible. This emphasizes that it is very significant to lose games by an average of 5 points per game.

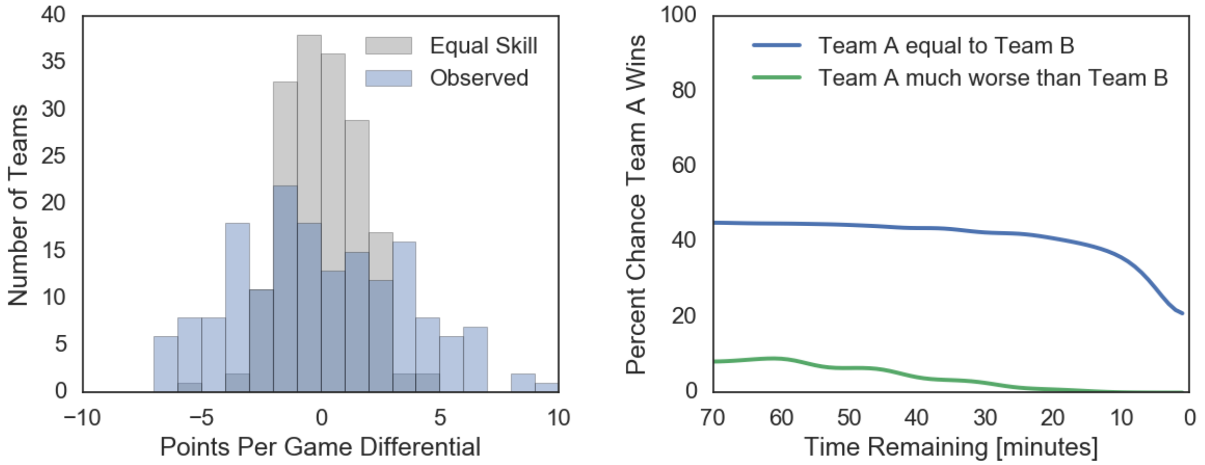

Figure 1: Left*: Distribution of points per game differential for 171 teams from seven years of BUDA spring season recreational ultimate frisbee leagues. Blue curve shows the observed distribution. Grey curve shows the expected distribution for teams of equal skill level. The observed distribution is much broader than expected, indicating significant imbalance in team skill level.* Right*: Percent chance of Team A beating Team B as a function of time remaining in the game, if the teams are of equal skill (blue line) and if Team A loses its games to Team B by an average of 5 goals per game (green). In the unbalanced scenario, the weaker team begins with an 8.5% chance of winning and on average starts halftime with a 3.4% chance of winning. This demonstrates that a team that loses by 5 goals per game is significantly inferior to the average team in each league.*

3. Self-Assessment Ratings and Captain-Assessment Ratings

BUDA's recreational league teams are assembled using player ratings. The rating for a given player is derived from that player's self-assessment of skill level (self-assessment rating) as well as from a captain's assessment of skill level (captain-assessment rating). In BUDA recreational leagues, all players have a self-assessment rating, but not all have captain's assessment ratings. In this section, we discuss each of these ratings in turn.

Self-assessment ratings are generated from survey answers that players provide when first signing up for a BUDA recreational league. Survey questions address each player's ability to throw, catch, and run. A question related to each player's experience level is included, but it is given little weight in the final self-assessment rating. Most of the weight is given to questions on the topic of throwing. Self-assessment ratings are scaled to range from 10 (lowest) to 100 (highest).

One of the limitations of these ratings is that players often forget to update their survey responses over time. Many of the self-assessment ratings in the BUDA database are several years old and do not reflect present injuries or significantly increased skill.

Figure 2 shows points per game differential as a function of self-assessment rating. Each dot on this plot represents one team out of the 171 spring recreational league teams in the past seven years of BUDA. We use points per game differential as a proxy for team performance for the same reasons outlined at the beginning of Section 2. The self-assessment rating for each team is the average of the individual self-assessment ratings for the players on that team.

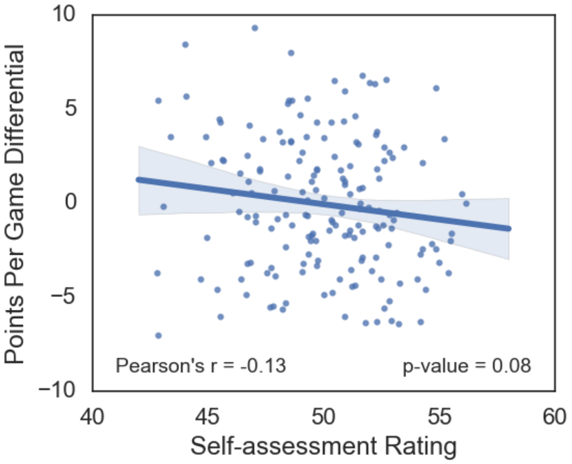

Figure 2: Correlation between points per game differential and self-assessment rating observed in seven years of BUDA recreational ultimate frisbee league data. The Pearson's r correlation coefficient is printed in the lower left corner of the panel. Self-assessment ratings show a negative correlation with actual in-game performance ($r = -0.13$), although the p-value of 0.08 indicates as much as an 8% chance that this correlation could be a result of random noise. The absence of a strong, positive correlation in this plot indicates that players are not very good at rating themselves, to the extent that self-assessment ratings are not a useful predictor of team performance.

The most striking result from this plot is that teams with high self-assessment ratings tend to do worse than teams with low self-assessment ratings. This statement can be quantified using the Pearson's r correlation coefficient. The value of $r = -0.13$ (printed in the lower left corner of the panel) indicates a weak negative correlation between self-assessment rating and team performance. The p-value of 0.08 (printed in the lower right corner) indicates a roughly 8% chance that the observed correlation could happen randomly. Whether or not the negative correlation is random, it is clear that there is no positive correlation between these variables. Self-assessment ratings have limited utility in terms of predicting team performance.

BUDA has long been aware of the inconsistency of self-assessment ratings. In response, BUDA has implemented captain-assessment ratings. For each player on their team, captains are asked to answer the same survey questions used in the self-assessment rating. Captain ratings should be more reliable than self-assessment ratings because captains are rating their players in a relative sense (how good is player X compared to the average player on my team?), whereas individuals are rating themselves in an absolute sense (how good am I overall?). In addition, captains tend to have more experience with league play and therefore have a better baseline from which to judge players. This is especially true in comparison to newer players that have little or no league play experience.

Figure 3 shows how points per game differential relates to captain rating. As before, a single dot on this diagram is one team from the past seven years of BUDA spring recreational league. The captain-assessment rating for a given team is the average of the captain-assessment ratings for the players on that team. When a captain-assessment rating is not available for a given player, that player's self-assessment rating is used instead (more on this at the end of this section).

There is a strong, positive correlation between captain-assessment rating and team performance—if the captain rating indicates a team will do well, they do well (and vice versa). The Pearson's correlation coefficient is $r = 0.31$, and the p-value of < 0.01 indicates a very low probability of this

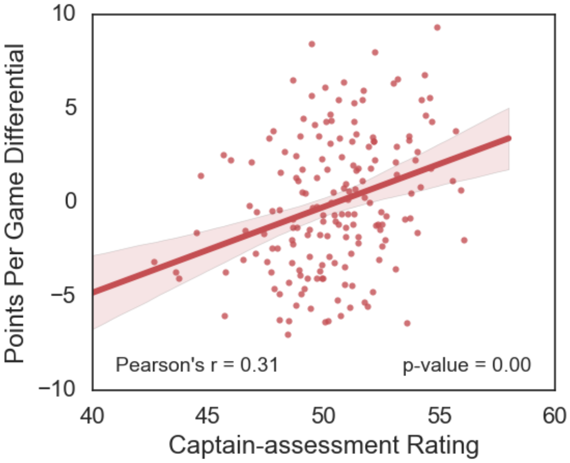

Figure 3: Correlation between points per game differential and captain-assessment rating observed in seven years of BUDA recreational ultimate frisbee league data. The Pearson's r correlation coefficient is printed in the lower left corner of the panel. Unlike self-assessment ratings, captain-assessment ratings are positively correlated with per-game plus/minus ($r = 0.31$). This shows that ratings from captains are more accurate than ratings based on self-assessment.

result occurring by chance. There remains a broad range of team performance for a given captain rating, but some of this is to be expected given the inherent uncertainty in points per game differential, even between two teams of equal skill level (as shown in Section 2).

The rating system currently in place in BUDA recreational leagues is the arithmetic mean of the self-assessment rating and the captain-assessment rating. For example, if a player has a self-assessment rating of 40 and a captain-assessment rating of 70, their BUDA rating will be 55. Figure 4 shows the relationship between team performance and current BUDA rating. There is a weak positive correlation ($r = 0.10$), but the p-value of 0.21 indicates a roughly 1 in 5 chance that this result could happen randomly. In other words, the existing BUDA rating is not predictive of team performance. This result shows that imbalanced teams in recreational leagues are due to imperfections in the rating system rather than the draft process itself.

As a final note to this section, it is worthwhile to comment further on captain-assessment ratings. BUDA asks captains to rate the players on their teams, but the role of the captain is a volunteer position. From the captain's perspective, there is no monetary reward for going the extra mile and rating all the players on one's team——typically around 16 players. Consequently, many captains abdicate their responsibility of rating the players on their team. As a result, many players in the BUDA database have no captain rating, despite playing for years in BUDA. On an average roster of 16 players, 4 players will not have a captain rating. In some cases, as many as 8 players lack a captain-assessment rating. This limits the utility of captain-assessment ratings and motivates the need for an alternative, data-driven rating system that does not rely on continual human intervention to function properly.

4. Club-Assessment Ratings: A Data-Driven Rating System

As an alternative to direct human assessment ratings (whether self-assessment or captain-assessment), we advocate instead for a data-driven rating system that considers each player's history within BUDA club leagues. Unlike recreational leagues in which players are randomly assigned to teams according to skill level, players in club leagues form their own teams and select a

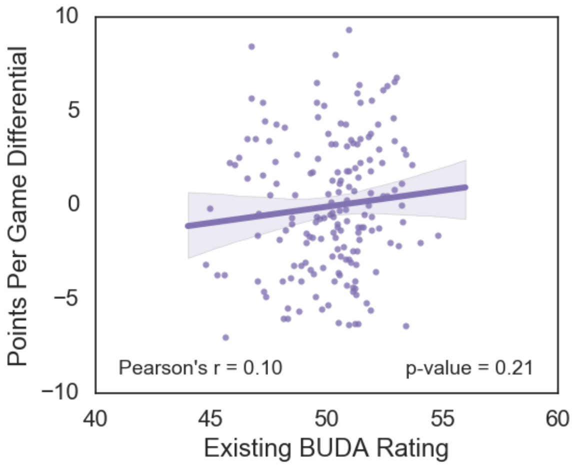

Figure 4: Correlation between points per game differential and rating according to the existing BUDA system (see text for details) observed in seven years of BUDA recreational ultimate frisbee league data. The Pearson's r correlation coefficient is printed in the lower left corner of the panel. There is evidence for mildly positive correlation ($r = 0.10$), but the p-value indicates about a one in five chance that the observed correlation results from random noise. This shows that including individual self-assessment on equal footing with captain-assessment ratings considerably reduces the predictive power of the ratings. It provides a clear explanation for why the existing BUDA rating system regularly results in imbalanced teams.

division of competition which they believe to be appropriate to their team-wide skill level. There are four divisions ranging from casual (Division 4) to highly competitive (Division 1).

Each club team uses a vetting process to determine whether or not a player can join their team. This is often as simple as a friend's recommendation, but it can also be as involved as a multi-day tryout. The key point is that to join a team, a player must have exhibited evidence that they will be competent at the skill level of the division in which that team competes.

Based on our experience of playing in different BUDA club league divisions, we begin by assigning the following ratings to each division in the style of the Elo system[1]: Division 1: 1800, Division 2: 1400, Division 3: 1100, and Division 4: 900. The idea is that a Division 1 team is expected to beat a Division 2 team 90% of the time, a Division 2 team expects to win 85% of the time against a Division 3 team, and a Division 3 team expects to beat a Division 4 team 75% of the time.

The next step is to acknowledge that within each division, there is a wide range of team-wide skill level. Our experience with BUDA suggests that a weak Division 1 team is comparable to a strong division 2 team. Thus, we calculate team ratings as the following:

$$Rating_{team} = Rating_{division} + 60\ times\ PPGD$$

where $PPGD$ is the points per game differential for that team. Under this definition, a Division 1 team that loses on average by 3.3 points per game has an equal rating to a Division 2 team that wins on average by 3.3 points per game.

[1] https://en.wikipedia.org/wiki/Elo_rating_system

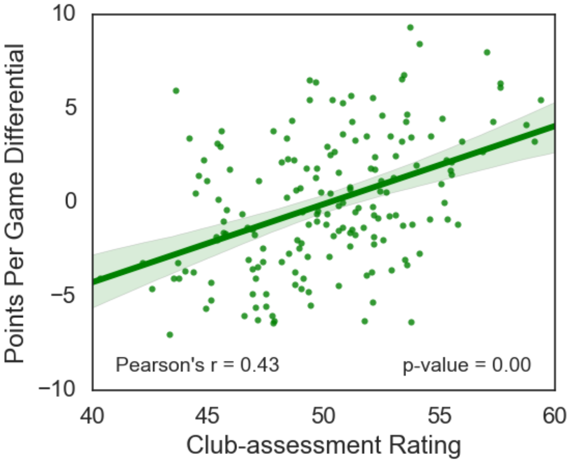

Figure 5: Correlation between points per game differential and club-assessment observed in seven years of BUDA recreational ultimate frisbee league data. The Pearson's r correlation coefficient of $r = 0.43$ is significantly higher than the correlation coefficient based on captain-assessment ratings ($r = 0.31$) and much higher than that based on self-assessment ratings ($r = -0.13$). This shows that club-assessment rating is the best predictor of team performance and should be used whenever possible to assign ratings.

This methodology yields a team rating for every club team in the BUDA database. The club-assessment rating for a given player is the arithmetic mean of all of their team ratings leading up to that point in time. Here is a concrete example. Suppose a player has experience on an average Division 1 team ($Rating_{division} = 1800, PPGD = 0$), a strong Division 2 team ($Rating_{division} = 1400, PPGD = 3$), and an average Division 2 team ($Rating_{division} = 1400, PPGD = 0$). In this scenario, their club-assessment rating will be $Rating_{team} = \frac{1}{3}(1800 + 1580 + 1400) = 1593$.

The last step that is unique to assigning a club-assessment rating to a player is the conversion to the same 0 to 100 scale used by BUDA for self-assessment ratings and captain-assessment ratings. The simple philosophy guiding this process is our experience with BUDA, which suggests that an average Division 1 player will have a club-assessment rating of 80, while an average Division 2 player will have a club-assessment rating of 60. An average Division 3 player gets a club-assessment rating of 45 and an average Division 4 player gets a club-assessment rating of 30. We use interpolation to fill in the gaps for ratings between these values.

Figure 5 shows the relationship between club-assessment rating and team performance (as measured by points per game differential). Club-assessment rating correlates more strongly with team performance than captain-assessment rating ($r = 0.43$ for club-assessment rating, $r = 0.31$ for captain-assessment rating). This demonstrates that club-assessment rating is a better predictor of team performance than captain-assessment rating—and a much better predictor than self-assessment rating. If a player has a high club-assessment rating, that is a very strong indicator that they are a good player (and vice versa).

One of the limitations of the club-assessment rating method for building recreational league teams is the presence of players in recreational leagues that have no club experience. On average, a roster of 16 players can expect to have 3 players without club experience. Although this is better than the average of 4 players that typically do not have a captain rating—see Section 3—it must still be taken into consideration. Some of these are players new to ultimate or new to the Boston area. Individual self-assessment ratings are the only means by which to rate these players. Other players have significant experience in recreational leagues, but zero experience in club leagues. For this latter subset of players, we use a captain rating if available.

We wish to emphasize that the club-assessment rating method outlined in this section requires no manual input by humans to function, unlike captain ratings and individual self-assessment ratings. The primary effort lies in recording who plays on what team and how those teams do. BUDA has shown that collecting and storing the data is not only feasible, but often an automatic part of running the league. This paper shows how to use the data to make real improvements in rating systems.

Finally, it is instructive to consider the practical implications of the findings in this section. In particular, how would the roster of the worst team need to be changed to yield an average team? The difference in club-assessment rating between the worst team (about 43) and an average team (about 51) is 8 rating points per player. On a team with 16 players, this corresponds to 128 total rating points. The best players have club-assessment ratings in the range 80 - 90, while average players have ratings in the range 45 - 55, and the worst players have club-assessment ratings between 10 - 20. To achieve a swing of 128 total rating points requires significant roster changes—even replacing a very weak player (rating of 15) with a very strong player (rating of 85) is insufficient because it brings about a change of only 70 points. Repeating this process (i.e., taking the two worst players on the team and replacing them with superstars) is barely sufficient to turn a terrible team into an above average team. Indeed, balancing the teams created by the existing BUDA rating system requires significant roster adjustments.

5. Recommendations for League Organizers

The purpose of this section is to aid league organizers who wish to use the information presented in this paper to improve parity in their recreational leagues. Our work reveals that a player's experience in club leagues, where there is a vetting process to determine if a player can participate at their preferred division of competition, is more predictive than their own self-assessment as well as the assessments of any captains they have had in the past. For this reason, our first recommendation is that leagues use club ratings when they are available.

Captain-assessment ratings should be used when club ratings are not present. This is a key result for regions that do not have the participation numbers to run club leagues like BUDA. For smaller organizations that are frustrated by the frequent occurrence of imbalanced teams, we highly recommend that captains regularly rate the players on their recreational league teams. Captain-assessment ratings are significantly more predictive of true skill level and future team performance than self-assessment ratings. League organizers might even consider providing a financial reward to captains that complete player assessments at the end of each season, perhaps a refund of the registration fee or discount for future league participation.

It is inevitable that some fraction of the ultimate frisbee population in a given league will not have a club-assessment rating or a captain-assessment rating. For these players, we recommend the use of self-assessment ratings.

It is important to emphasize that maximizing the value of club-assessment ratings requires that leagues continuously track and store roster data and game scores. Furthermore, club-assessment ratings will have the most utility if there are stratified divisions within the club league, such that a player who participates in one division is clearly better or worse (on average) than a player who participates in another division. The appropriate base rating for each division should be tailored to each league. We recommend using the Elo system as we have done here, and then translating the resulting Elo rating into the rating system in use by the league in question. The translation process should also be tailored to each league. To aid league organizers, the code we have used to build the club-assessment ratings for BUDA is publicly available [5].

6. Summary and Future Work

We have established that individual self-assessment ratings should be used as a last resort. We have also shown that club-assessments ratings are more predictive of true skill level than captain-assessment ratings, which are themselves much better than self-assessment ratings. We expect that the guidelines provided in Section 5 will yield teams with much more balance than is currently typical in BUDA recreational leagues. Yet, there remains the potential to improve player ratings even further given the data at hand. Here are the most significant potential paths of research from our perspective.

- BUDA tracks individual game scores for all league games. This is important because in a typical BUDA recreational league, the season is 7 games long. A given team will play only 64% of the possible competition. Under these conditions, it should be possible to use individual game scores to produce more accurate measures of team performance [3]. On the other hand, variation in individual game scores is likely to be influenced significantly by random factors such as attendance and luck. An important topic for future research is to determine the utility of individual game scores.

- Individual game scores can also be used to measure team performance in club leagues. This could improve the predictive ability of club-assessment ratings for players in recreational leagues. The same benefits and caveats to using individual game scores apply here as well.

- It is possible to use recreational league playing history to develop a rec-assessment rating that applies the club-assessment methodology to recreational league data. The challenge with this approach is that these teams are designed to be comparable in skill level, so the signal-to-noise is lower. The best approach may be to consider which seasons a player has played in, since some seasons are indicative of stronger skill (winter indoor leagues) whereas others indicate weaker skill (summer or fall leagues).

- It may be the case that differences between club-assessment ratings or captain-assessment ratings and self-assessment ratings hold predictive power of team performance. For example, if someone rates themselves as a 75 (a very strong player), but their club-assessment rating is 50 (an average player), this could indicate someone who has an inflated opinion of their skills. This is likely to translate on the field to turnovers resulting from poor decision-making. Including rating differentials as an additional component of player assessment may improve predictive power.

- Machine learning competitions have shown very clearly that an ensemble of distinct model predictions (or ratings, in this case) outperform individual model predictions[2]. Future work should identify the most promising approaches to ensembling.

- Finally, this paper has focused on building the most accurate player rating system possible. However, generating balanced teams requires assigning players to each team. Future work should seek to optimize the draft process.

[2] http://mlwave.com/kaggle-ensembling-guide/

References

[1] SFIA (2016) 2016 Sports, Fitness, and Leisure Activities Topline Participation Report. Sports and Fitness Industry Association

[2] USAU (2016) USA Ultimate Annual Report. USA Ultimate http://www.usaultimate.org/about/usaultimate/annual_report.aspx

[3] Langville, A. N., & Meyer, C. D. (2012). Who's #1?: The science of rating and ranking. Princeton: Princeton University Press.

[4] Heuer A, Müller C, Rubner O (2010) Soccer: Is scoring goals a predictable Poissonian process?. EPL (Europhysics Letters) 89:38007. doi: 10.1209/0295-5075/89/38007

[5] Bussmann S (2016) BUDA Ratings. Github https://github.com/sbussmann/buda-ratings