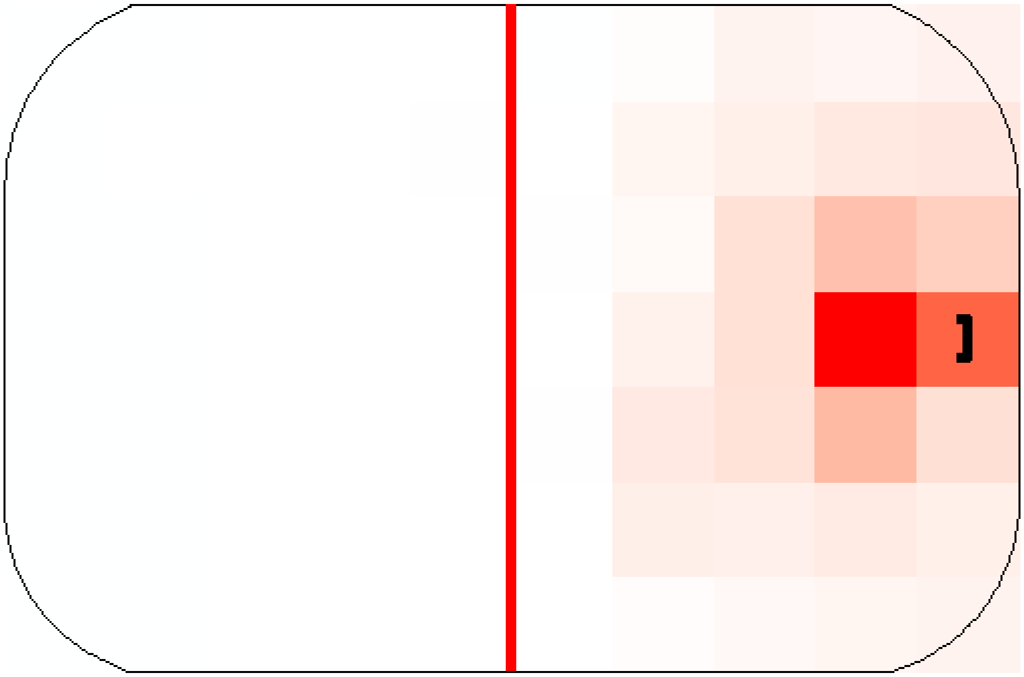
How Do We Get There: Quantifying Pass Types and their Value

Ben Howell, University of Texas at Austin

Big Data Cup 2021

March 04, 2021



Introduction

This project focuses on measuring the value of an action taken by a team with regards to how it affects (positively or negatively) their chance of scoring. Hockey is unique with the idea of *hockey assists*, rewarding up to two players for their contribution to a goal. Basketball and soccer assign assists, but only to one player who passed to the shooter. The chess moves to set up a good shot are always occurring and the pass right before a shot is not necessarily the most important action in that sequence.

In this project, I define *Expected Attack Value (xAV)* as the value of every pass, movement, or shot based on how an action sets up future actions. The data has been provided by Stathletes for the 2021 Big Data Cup, containing 2018 Women's Olympic Tournament, NCAA, and 2021 NWHL bubble data.

This paper is heavily inspired two concepts from other sports: *linear weights* in baseball and Expected Threat (xT) analysis in soccer by *Karun Singh*.

In baseball, there are 24 out and on-base states, each with a distinct Run Expectancy. Bases loaded with 0 outs has a higher Run Expectancy than no one on, 2 outs. Events are valued by how they change the run expectancy, leading to Weighted On-Base Average (wOBA) and valuing a home run more than a single.

This project's Expected Attack Value (xAV) is based on the idea that actions from different locations on the rink are worth more than others and attempts to assign every action (shot/pass/movement) a value.

The Karun Singh's soccer xT analysis provided a framework for my xAV formula and was crucial to taking this analysis from theoretical to a realized project.

Measuring the Value of an Action

To make things manageable, I split up the hockey rink into a 10 x 7 grid, representing 70 zones. I then calculated xAV by zone, which then allows us to measure the value of moving from your current zone (x, y) to your next (z, w).

The xAV of an action is defined in two parts: the chance of scoring a goal given a shot is taken (shot score) + movement (move score). The equation for calculating xAV of a zone is below.

$$xAV_{xy} = (s_{xy} * g_{xy}) + (m_{xy} * \sum_{x=1}^{10} \sum_{y=1}^{7} ((P_{xy->zw} * T_{xy->zw}) * xAV_{zw}))$$

- (x, y) represents current/starting zone and (z, w) represents value of next zone
- xAV_{xy}: Expected Attack Value of an action based on starting location (x, y)
- s_{xy}: Probability of taking a shot from starting location
- g_{xy}: Given a shot is taken, what is the probability of scoring a goal
- m_{xy}: Probability of moving from a starting location ($s_{xy} + m_{xy} = 1$)
- $P_{xy->zw}$: Probability of a successful pass/movement from (x, y) location to (z, w) location
- $T_{xy->zw}$: Probability of moving from starting location to next location
- xAV_{zw}: Expected Attack Value of next location

Calculating xAV is simple: if a shot is taken, the xAV of the zone is your result; if it's a movement, you take the subtract xAV_{zw} - xAV_{xy} to get the value added of a movement.

The shot score is essentially a simple xG model. Given the initial position of the player (noted by x, y), multiply the percent of actions from this zone that have been shots by the percentage of shots that result in goals from that zone. This returns a simple xG score for each zone. (An early version of this project built xAV off xG from an `xgboost` model. You'll see the xG results later, but that plan was scrapped because the xG model took into account variables such as number of opposing skaters and I wanted to keep xAV free of that influence.)

The shot score of a location is then added to the movement score. The movement score represents the overall value that can be added by moving to a zone from your current one. The probability of moving from your current zone is represented by m_{xy}; m_{xy} and $shot_{xy}$ should add to one, representing all possible actions from a zone.

The movement score represents the expected value that comes from moving to a specific zone. $T_{xy->zw}$ is the probability of moving from the current zone (x, y) to another specific zone (z, w). I multiplied that by success% of actions from the current zone to the next zone, whether through a completed pass or skating without losing possession. Once I had that likelihood of successfully ending up in $zone_{zw}$ from $zone_{xy}$, I multiplied in the xAV of the zone in question, giving us a xAV of moving from one zone to another.

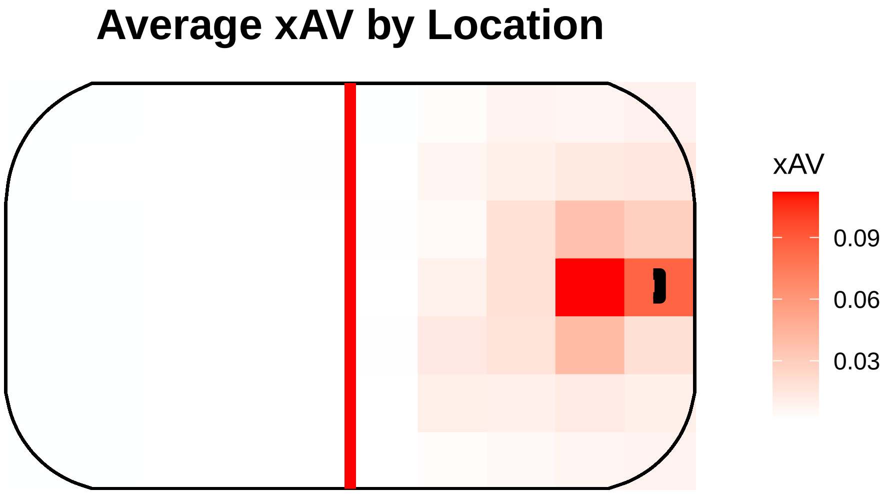
To finish, I sum all xAV of the potential zones that a player *could* move to from $zone_{xy}$. Multiplying back in m_{xy} is the final step before we add back in the shot score.

The movement score boils down to this: if a player chooses to move, which they do X% of the time from their starting point, players have moved to each zone (z, w) Y% of the time, with W% success, resulting in Z attack value.

Getting this calculation started was the hardest part. The values of s_{xy}, g_{xy}, m_{xy}, $P_{xy->zw}$, and $T_{xy->zw}$ are constants. In our Big Data Cup dataset, players moved from Zone A to Zone B X% of the time, that never changes. But, we don't have an initial xAV for the zones that we want to move to, which prevents us from running the formula, preventing us from getting xAV per zone, and so on.

We have to **evaluate this formula in iterations, until the values converge** (this concept was introduced to me through Singh's soccer xT work). The first time we run this model, it produces a simple xG score for each zone. Once we have that, we can re-run the formula, using that new xAV per zone that we just calculated as the new xAV_{zw} for the next iteration of the model. I ran this formula five times until the xAV per zone no longer changed with new iterations.

Below is a visualization of xAV by zone. Unsurprisingly, the zones closest to the goal and in the offensive zone have the highest xAV, meaning that actions moving into those zones are worth the most.



Ben Howell | @benhowell71 | benhowell71.com

Now that we've created the formula and framework for xAV, here's a look at some of the highest-rated players by cumulative xAV from the NWHL data and Olympic data we have. For comparison, I created an `xgboost` xG model and added that score to our table.

Given that the xAV and xG scores are cumulative, I've standardized them to 100 actions or shots. Many of the top $xG/100$ leaders are represented in the top 8 $xAV/100$ leaders in each league, but it's not a one-to-one comparison, which is a good thing! The discrepancy between $xAV/100$ and $xG/100$ indicates that they're measuring and valuing different things.

In our NWHL leaderboard, Boston and Toronto are the only teams represented. In our Olympic leaderboard, Canadian players take seven of the top eight $xAV/100$ spots. These teams have the most observations in their respective NWHL/Olympic dataset and, since I used 150 events as a cutoff, those teams are much more likely to have players hit that 150 event mark.

You can find an interactive plot of movement probabilites from zone to zone, as well as a full leaderboard of *xAV/100* and xG metrics **here**.

Table 1: NWHL xAV/100, xG/100 Comparison (min 150 Events)

Team	Player	Events	xAV/100	xAV Rank	xG/100	xG Rank
Boston	Taylor Wenczkowski	160	0.91	1	5.24	5
Boston	McKenna Brand	234	0.86	2	4.79	9
Boston	Samantha Davis	205	0.82	3	8.55	2
Toronto	Breanne Wilson-Bennett	163	0.73	4	9.06	1
Toronto	Mikyla Grant-Mentis	202	0.71	5	6.01	3
Boston	Tereza Vanisova	203	0.57	6	4.42	11
Boston	Christina Putigna	184	0.57	7	5.08	7
Toronto	Emma Woods	153	0.43	8	5.12	6

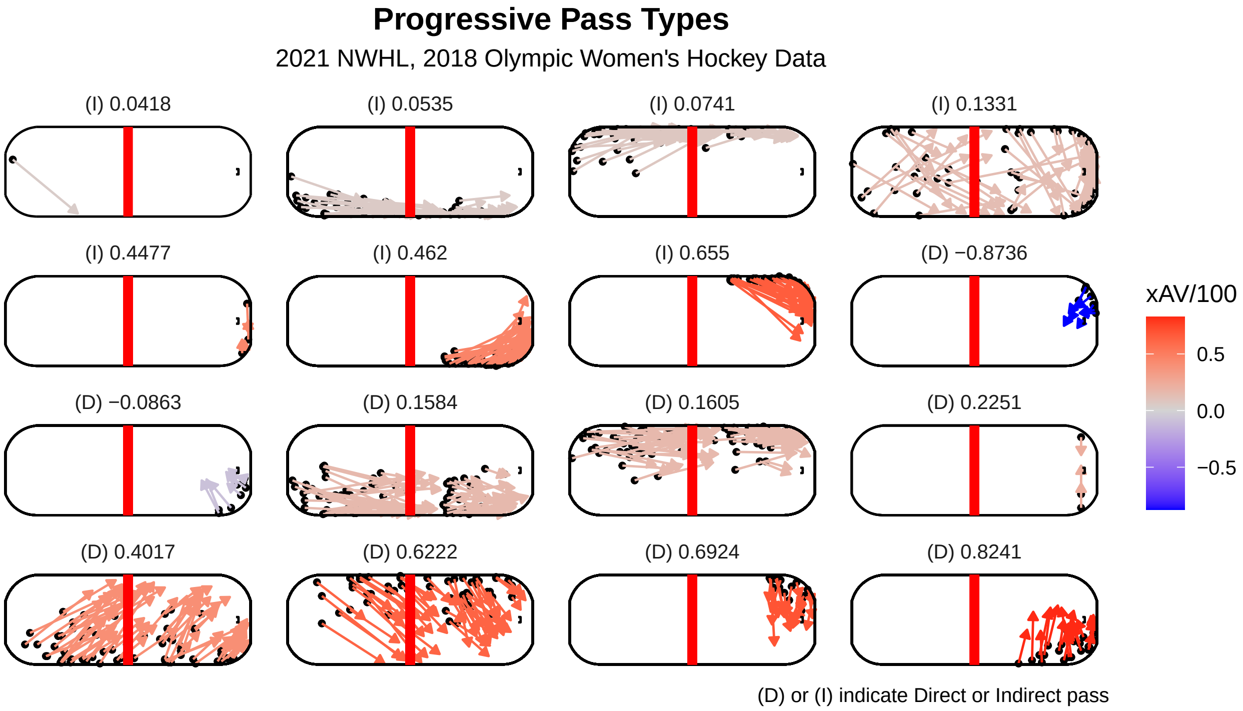
Table 2: Olympic xAV/100, xG/100 Comparison (min 150 Events)

Team	Player	Events	xAV/100	xAV Rank	xG/100	xG Rank
Canada	Meghan Agosta	161	0.66	1	6.51	5
Canada	Natalie Spooner	185	0.65	2	8.61	2
Canada	Sarah Nurse	177	0.65	3	4.98	14
USA	Hilary Knight	231	0.63	4	6.62	3
Canada	Rebecca Johnston	375	0.62	5	5.31	10
Canada	Marie-Philip Poulin	330	0.54	6	5.44	9
Canada	Brianne Jenner	276	0.54	7	6.12	7
Canada	Melodie Daoust	229	0.54	8	6.30	6

Common Pass Types

Now that I've defined xAV and shown specific player results, I was curious to see if there were any patterns in common pass types and how successful they were (using $xAV/100$).

I separated Direct and Indirect passes and ran a Gaussian Mixture Model (`Mclust` from the `mclust` package in R) to identify subsets of passes within the broader Direct/Indirect pass categories. Variables included in the clustering model include the start and end coordinates, the direction of the pass, and the distance the pass traveled from Point A to Point B.



The model returned nine direct pass types and eight indirect pass types. However, I further pared it down to 16 total subsets when I removed passes that were not progressive and didn't move the puck towards the goal (defined by moving at least 25% closer).

Above is a visualization of each progressive pass type (only 10% of each type are shown for a cleaner visualization). Passes that center the puck sport the highest xAV/100; it would be interesting to compare this with a men's hockey dataset, as the limitations on body checking may open up the middle of the rink for these attacks. Most of the top pass types were direct passes, which may be related to the previous point as it's hard to consistently send indirect passes to the center of the ice.

Next Steps

There are a few ways to push this analysis further, primarily dealing with accounting for the skater situation on the rink (power-play or even strength). This would likely appear in a true xG or Pass Success Probability model in place of the g_{xy} and $P_{xy->zw}$ I ended up using, which were derived solely from the recorded events in the Big Data Cup dataset.

Acknowledgements

I would like to thank Stathletes for hosting the Big Data Cup and making their Olympics data available; open competitions like this are a great way to drive innovation and working with women's hockey was an exciting experience. I'd like to thank the NHWL for making their data available as well.

Thank you to two of my friends, Sweta Ghose and Abhi Mandalam, for looking over this paper and providing feedback.

Appendix

- Interactive Movement Frequency Plot and full xAV leaderboard:
  https://benhowell71.shinyapps.io/BigDataCupApp/
- Code for this project is on GitHub:
  https://github.com/benhowell71/Big-Data-Cup
- Karun Singh's xT analysis for soccer:
  https://karun.in/blog/expected-threat.html
- Reading on Linear Weights in baseball from FanGraphs:
  https://library.fangraphs.com/principles/linear-weights/