Using Threat To Value & Contextualize On-Puck Actions

Avery Ellis and Matt Hurley

March 5th, 2021

1 Introduction

A player possessing the puck, no matter their position on the ice, has three possible actions available to them: shoot the puck, skate with the puck, or pass the puck to another player, successfully or not. Much of public analytics work to date has focused on the first choice, creating models that focus on shots and goals, and evaluating players and teams based on those results. However, there has been comparatively less research on passing and skating. Corey Sznadjer's microstats and Ryan Stimson's Passing Project have given some insight into neutral zone play and passes leading to shots, respectively. However, the Big Data Cup's comprehensive location data for *all* on-puck events provides an immense opportunity to build on that research and refine our understanding and valuation of passing and skating.

One approach to do so is an expected threat model, a model that assigns value to on-puck events based on the location-based value the event generates. Location-based expected threat models in hockey (Forstner, 2020) and soccer (Singh, 2019) have been implemented successfully and can address a central question that one would like to ask of the dataset: given only a 40-game season, how can a team properly evaluate a prospect's contributions, development, and potential? Our paper will explore one answer to that question by implementing and contextualizing an expected threat model on non-shooting on-puck actions; that is, passes (completed and failed), and skating with the puck.

2 Building the Threat Model

2.1 Model Methodology

Following the methods of Singh and Forstner, we divide the rink into h vertical zones and k horizontal zones; these were empirically chosen as $h = 17$ and $k = 40$, making each zone 5 feet by 5 feet. We then define the expected total threat, xTT, of a zone (x, y) on the ice by:

$$xTT(x,y) = \big(p(S_{(x,y)}) \cdot p(G_{(x,y)})\big) + \left(p(M_{(x,y)}) \cdot \sum_{(i,j)} T_{(x,y)\to(i,j)} \cdot xTT(i,j)\right) - \left(p(L_{(x,y)}) \cdot \sum_{(i,j)} T_{(x,y)\to(i,j)} \cdot xTT(i,j)^*\right).$$

Let's break down this three-part formula.

The first component, $p(S_{(x,y)}) \cdot p(G_{(x,y})$, describes the "shot value" of a zone; it is the probability $p(S_{(x,y)})$ of shooting from a zone (x, y), weighted by the probability $p(G_{(x,y)})$ of scoring from that zone.

The second term describes the "move value" of a zone, whether due to passing or skating. If a player is in a zone (x, y), we calculate the probability that they move to some other zone (i, j), and multiply this by the xTT generated at (i, j); summing this quantity over *all* zones on the rink and weighting by the overall probability $p(M_{(x,y)})$ of moving from (x, y) gives the expected value of moving.

Finally, the third term defines the "loss value" of a zone: when a player loses the puck (whether it is stolen at that position or given up by failing a pass), the puck is recovered by the other team at some position (i, j). The product of the probability that it is recovered there and the xTT of that zone's mirror across the center-ice faceoff dot, $(i, j)^*$, is summed over all zones on the ice to represent the threat gained by the opposing team; we then weight by the probability $p(L_{(x,y)})$ of turning over the puck at (x, y) to compute the negative value of a zone. This "loss value" was not part of Singh's original implementation of xT for soccer; to acknowledge these negative values, we call our model xTT, expected *total* threat.

In the equation describing xTT, the move term and the loss term are both dependent on the value of all other zones. Without any preliminary values for xTT, we recursively generate it, first defining the xTT of each zone to be 0. Thus, for the first iteration of our xTT model, our equation simplifies to the "shot value" term, a naive way to describe expected goals.

With those values generated, we can then run our model until sufficient convergence is achieved, using the prior iteration's xTT values to define the xTT of the current iteration. Thus, after, say, three iterations, the xTT of a given zone is equivalent to the expected net probability of scoring versus being scored on over the next three actions originating from that zone. Singh and Forstner both empirically found that their model converged after five iterations, and we found the same.

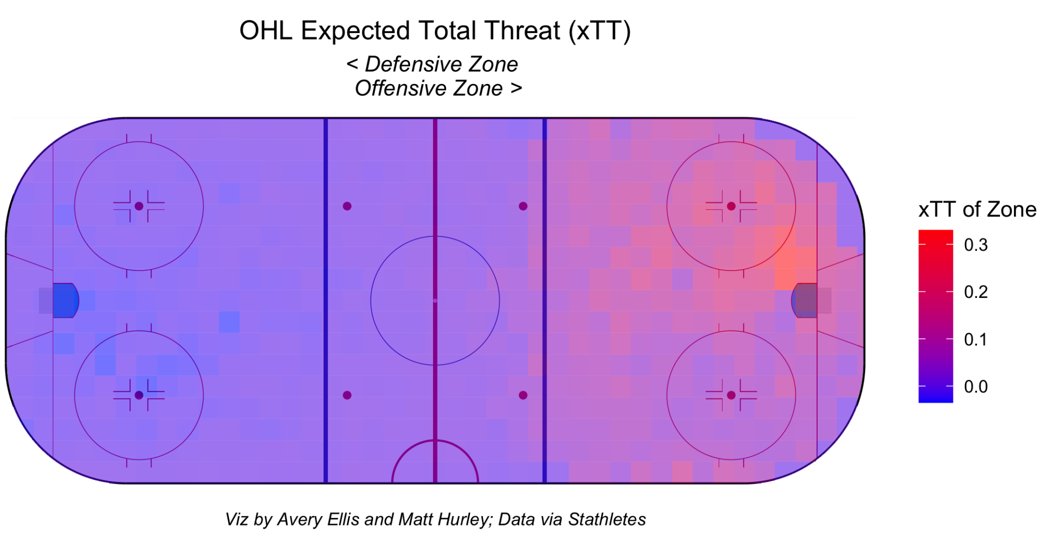

Figure 1: The xTT of every 5x5 zone

2.2 Model Issues

One of the features of an expected threat model is that it is a Markov model: future states of the model depend *only* on the present state and not on any past states. This memoryless feature is incredibly useful: it allows one to calculate an objective numerical value of each event, based on the information contained in a given dataset, without requiring further context. But that is also the downside—xTT treats each event (pass, shoot, skate) as entirely independent of one another. In a team sport like hockey this is of course not the case. To see where xTT stumbles, consider this sequence in the Chicago Blackhawks vs. Columbus Blue Jackets game from February 25th, 2021.

The play begins as Janmark (13) picks up the puck behind the Blackhawks' net. Janmark carries the puck from there up towards the blue line—an event that generates a positive xTT—but then notices a Columbus defender, and begins to circle back. With Kane (88) streaking into the neutral zone, Janmark looks for him, is unable to find an easy pass, and so sends the puck to de Haan (44) near the left side of the Blackhawks' net. De Haan attempts a pass to DeBrincat (12) near the Columbus blue line, but the pass is blocked by a Columbus defender, before being recovered by Kane, who walks into the offensive zone and snipes the puck past Korpisalo to open the game's scoring.

One way to analyze this possession would be to assign each individual the sum of the xTT generated by all of their individual actions. But this approach fails to fully account for the context of the possession. In the case of Janmark, the xTT of his pass is actually *negative*, moving from a very low (positive) threat to a negative threat, suggesting that, on average, the pass is putting the team in a more vulnerable position and more likely to lead to a goal against rather than a goal for. But in the context of this sequence, we can see that this was certainly not the case. Janmark merits credit for a sharp decision to set up de Haan's stretch pass, which led to Kane's carry and goal. So, how can we account for this complicated balance?

3 xTT Chain

3.1 Introduction

Our solution to this problem is what we call xTT Chain. We begin by defining a possession as any number of consecutive events where one team controls the puck. xTT Chain attempts to account for the context of a play by evenly weighting an individual's contributions (which may be negative) with that individual's share of the total xTT generated by that possession, provided the player's contributions advance a possession via a pass. Those players who are part of a possession but do not pass the puck (e.g., receive or recover the puck and then turn it over, receive the puck and then shoot) are assigned an xTT Chain that is simply equal to their personal contribution (discussed further below). Thus, even if an individual's actions do not have a positive xTT, they can still provide a valuable contribution to a possession where the overall threat increases; in the example sequence, Janmark would receive credit for the overall advancement of the possession, in addition to his carry's personal contribution.

3.2 Derivation

An passing player's xTT Chain from a possession involving N passing players is given by

$$\frac{1}{2}\left(\frac{\Delta xTT}{N} + xTT_P\right),$$

where xTT_P is an individual's personal contribution, calculated by summing the change in xTT over each individual event (passes, carries, puck recoveries, and turnovers). The personal xTT contributed by a carry or a pass is (as described above) the difference between start and end points. The personal xTT associated with an incomplete pass is the negative xTT of the zone where the puck was passed from, and the personal xTT associated with a recovery or takeaway is the xTT of the position where the puck was recovered. For a failed pass or turnover by team A, the xTT is the negative of the sum of team B's xTT where team B obtains the puck and team A's xTT at the passing/turnover position. Moreover, ΔxTT is the total change in xTT over the course of the possession; that is, the difference in xTT between start and end points of the possessing team.

With these definitions in place, we can compute the combined xTT of the N passing players. To do so, suppose that the jth player makes an incomplete pass, which the $j+1$ player recovers. Additionally, let xTT_i represent the xTT at the point where the ith player receives the puck (xTT_1 being the xTT zone where the first player starts with the puck). Then the sum of the individual xTT Chains is

$$\begin{aligned}
\Sigma xTT_{Chain} &= \sum_{i=1}^{N} \frac{1}{2}\left(\frac{\Delta xTT}{N} + xTT_{P_i}\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\frac{\Delta xTT}{N} + \frac{1}{2}\left[\sum_{i=1}^{j-1}(xTT_{i+1} - xTT_i) - xTT_j + xTT_{j+1} + \sum_{i=j+1}^{N}(xTT_{i+1} - xTT_i)\right] \\
&= \frac{\Delta xTT}{2} + \frac{1}{2}\sum_{i=1}^{N}(xTT_{i+1} - xTT_i) \\
&= \frac{\Delta xTT}{2} + \frac{xTT_{N+1} - xTT_1}{2} \\
&= \frac{\Delta xTT}{2} + \frac{\Delta xTT}{2} \\
&= \Delta xTT.
\end{aligned}$$

This computation extends to possessions with any number of incomplete passes and recoveries. The crux is that an incomplete-recovery pair is equivalent to a pass from the incomplete position (xTT_1) to a toss-up between the teams (xTT of 0) then followed by the recovery (xTT_2), and so we find that the change in xTT of that incompletion-recovery play is simply the change in xTT for an equivalent pass. Thus, we find that the sum of personal contributions is half of ΔxTT. As a result, the sum of the individual xTT Chains for a given possession is just the net xTT of the possession itself.

After calculating each player's *xTT* Chain for 5-on-5 possessions, we then divided their numbers by the number of games they appeared in to better compare their on-ice impact. We chose to normalize our values based on something that's not events or possessions; without ice time values, we decided to weight by games played. From there, we restricted our analysis to players who had played more than one game; since there are only seven teams with more than two games played in the dataset, we felt that doing so was the best way to balance a thorough analysis with a very limited sample.

3.3 Results, Conclusions & Insights

Below are the top 10 players from the scouting dataset by *xTT* Chain per game, subject to the minimum-game restriction:

	Team	Player	xTT Chain Per Game
1	Windsor Spitfires	Egor Afanasyev	0.5256
2	Flint Firebirds	Yevgeni Oksentyuk	0.4985
3	Kitchener Rangers	Donovan Sebrango	0.4773
4	London Knights	Liam Foudy	0.4501
5	Hamilton Bulldogs	Arthur Kaliyev	0.4015
6	London Knights	Kirill Steklov	0.3961
7	Flint Firebirds	Vladislav Kolyachonok	0.3952
8	Mississauga Steelheads	Keean Washkurak	0.3738
9	London Knights	Nathan Dunkley	0.3672
10	Flint Firebirds/Barrie Colts	Tyler Tucker	0.3642

Notably, eight of these players have been drafted by NHL teams, including five of those eight within the first three rounds. Clearly there is some correlation between NHL scouts' perception of players' offensive skill and our numerical analysis of these players' offensive value.

There are three major action points from our results. Firstly, while many of these prospects have already been drafted, there are a number of players in the dataset (including two of the top ten) who were undrafted or are currently draft-eligible. NHL teams have a limited amount of time to analyze prospects, and especially with a shortened season, combining traditional in-person or video-based scouting along with applications of Stathletes' comprehensive data in metrics like *xTT* Chain can help teams gain a fuller understanding of the numerical value of players' offensive contributions. With both video and numerical evidence to guide them, teams can find prospects whose offensive value may not be captured in traditional counting statistics.

Secondly, *xTT* Chain as a concept and a measure easily transfers to other hockey leagues. Forstner showed that an *xTT* model can be built to analyze the offensive value of players in the NHL, and teams looking to target offensive play-drivers in trades can utilize *xTT* Chain to find players whose offensive value is not necessarily captured in other shot-based metrics, or who are held back by their linemates.

Finally, the combination of personal contributions and team contributions to xTT Chain means that analysis of where a player's xTT comes from (whether via individualistic play or by working passing plays with their teammates to generate threat) can identify potential line combinations or team makeups to maximize offensive threat. Since xTT Chain measures the value of *all* on-puck non-shooting events regardless of location, it can easily be combined with shot-based metrics to paint a full picture of players' value added across all areas when they control the puck. Two potential areas of research include exploring the value added by specific tactics for breakouts, or building a xTT model to define the numeric values of power play or penalty kill structures.

4 Sources of Uncertainty & Next Steps

We wish to acknowledge the three largest sources of uncertainty that we ran into when building the xTT model and in our design of xTT Chain. This is not a comprehensive list, but understanding the inherent limitations of our work is crucial to the interpretation and reliability of our results, and to any future research on the contextualization of expected threat.

Firstly, and most importantly, is the small number of goals from the scouting dataset that account for the entirety of the first iteration. This impacts our xTT model greatly, as if a zone has no goals scored in it, its first-iteration xTT is significantly reduced (as seen in the asymmetry of our heatmap). We considered two things to fix this: first, utilizing goal location data from the NHL to calculate the values of these zones in the first iteration; second, utilizing logistic regression based on our dataset to find the probability of scoring from a zone based on the distance and angle to the goal. We chose not to incorporate the NHL data, believing that the differing skill level would mean less applicable results. We also wanted to utilize logistic regression, but did not have enough time to incorporate it for our final paper. However, if time permits in the future, we will update our model to include this.

Secondly, there were also a number of zones that simply had no events occur in them. We believe this to simply be a function of the small dataset, and because no events occur in them, it doesn't impact the model directly. However, if this model were applied to a larger dataset, our results and the xTT values of all zones will necessarily be adjusted as events occur in those unused zones. (The beauty of the xTT model is that nothing has to change to handle a different dataset—one can simply rerun the numbers with the new data.)

Finally, our last source of uncertainty comes from the normalization of our xTT Chain values. Without the data to find each player's time on ice, it is impossible to calculate their true normalized xTT Chain, both as a percentage of on-ice xTT Chain or per-60 minutes. Adjusting to per-game values is a step in the right direction, but to truly compare players' xTT Chain values to each other, they should ideally either be compared based on their share of on-ice xTT Chain or their per-60 values, as weighting per-game gives unequal benefits to players who accumulate more minutes in a first-line or top-pairing role.

5 Appendix

5.1 Acknowledgments

We would like to thank everyone at Stathletes for working diligently to make this dataset public and for creating this competition. Furthermore, we owe a prodigious amount of gratitude to Sam Forstner, for taking the time to answer so many of our questions as we built our model, and to Ian Anderson and Brian Macdonald who met with us through the HANIC office hours and helped guide our development of the xTT model. We also would like to recognize Karun Singh for his original implementation and thorough explanation of xT in soccer; our work would have been impossible without the foundation he built. Lastly, and most importantly, we would like to thank Chris van Benthuysen of the Latin School of Chicago, for all he does to encourage learning and the love of mathematics for all students. Mr. Van has been an invaluable resource for both of us in our development as students and mathematicians, and his guidance throughout the past six years in all matters has been invaluable. We wish him only the best honey mustard sauce and without his trademark van-gents neither of us would have found nearly the same level of love for math.

5.2 Code, Results, & Figures

All code and figures for this project can be found at `https://github.com/mhurley4/BigDataCup`.

5.3 References

Karun Singh's original implementation of xT for soccer can be found here.

A contextualization of xT and a critical study of its features and flaws in soccer that informed our creation of xTT Chain can be found here.

Sam Forstner's implementation of xT in hockey is courtesy of the ISOL-HAC (2020) conference, his presentation begins at the 42:21 mark of this video.