NETWORK THEORY ANALYSIS OF SOCCER: TUNED PASSING INFLUENCE

A THESIS

Presented to

The Faculty of the Department of Economics and Business

The Colorado College

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Arts

By

Daniel Krueger

December 2018

NETWORK THEORY ANALYSIS OF SOCCER: TUNED PASSING INFLUENCE

Daniel Krueger

December 2018

Mathematical Economics

danielpkrueger1@gmail.com

danielpkrueger.blogspot.com

Abstract: This paper examines whether success – defined as a win (3 points), tie (1 point) or loss (0 points) - is affected by a passing rating which is compiled through network theory analysis of passing in the sport of soccer. The proposed measurement builds a weighted network that accounts for the total number of teammates passed to, and the number of passes to those unique players. The model is run on 2012 Major League Soccer (MLS) season using Opta F24 event data. The Multinomial Probit Model suggests that the rating has no significant correlation with success.

KEYWORDS: Network Theory, Soccer, Passing, Degree, Edge Strength, Degree Centrality, Weighted Networks, Opta F24, Opta, Soccer Analytics, Major League Soccer, Multinomial Probit Model

JEL CODES: Z21, Z29, C12, C15, C25, C31

ON MY HONOR, I HAVE NEITHER GIVEN NOR RECEIVE UNAUTHORIZED AID ON THIS THESIS

Daniel Krueger

TABLE OF CONTENTS

Acknowledgments

There are a number of people that made this research possible. First, I would like to thank my family whose dedication to education has inspired me to continue to learn. I am not who I am today without their patience and love. I would like to thank Dr. Howard Hamilton of Soccer Metrics Research for our conversations about my metric and applications of network theory analyses in general. I would like to thank Dr. James Curley of the University of Texas, Austin for providing the data, and Dr. Kirsten Hogenson of Colorado College for our conversations about the theory underlying my proposed metric. Dr. Tore Opsahl is to thank for the degree centrality metric itself. His groundbreaking work on weighted networks and his CRAN published tnet R package, made my research possible. Joseph Mulberry deserves my thanks for his help with building out the weighted passing network graphic. His create_opta_pitch() function saved hours of ggplot scripting. A huge thanks also go to my classmate Jacob Miller who is a co-publisher of the final R script. He is a friend and a future colleague. Finally, I would like to give my sincere gratitude to my thesis advisor Dr. Neal Rappaport for his continual support from the early preparation, all the way to the end with the write-up. Dr. Rappaport, you are an inspiration to many, and a dear friend.

Introduction

This study hypothesizes that a robust weighted passing metric, that measures the total number of passes completed through the application of network theory analysis, will have significant correlation with success. The transfer of the ball from one player to another in the game of soccer creates a network of passes. The intricacies of this network can be summed and used to create a rating that is more powerful than simple passing statistics - such as possession percentage and total passes completed. In theory, a more varied passing network allows for more robust ball movement and forces defenses to adapt to more diverse passing sequences. The econometric model suggests that the proposed rating has no significant positive correlation with success. Possibly, the outcome of this model is due to the inability for the proposed passing rating to account for stylistic differences unique teams adopt.

As the sport of soccer is an extremely subjective game at face value, more objectivity can be brought to the sport through complex analyses. One coach can watch a match and think his team dominated, while the opposing coach can watch the same game and feel as if his team dominated. Soccer as its played and coached is largely based around style. A player's skill level has an impact on the extent to which the player is able to utilize the manager's style, but style is fundamental to the application of a player's skill level. No matter which style is used, all styles have the same end goal: success. Whether it's a tie against the league leader or a thumping of an inferior opponent, each team plays to find success in some fashion. To accomplish success, some teams may put many defenders behind the ball and make passing lanes congested. Alternatively, another

team may use a short passing style in hopes of keeping the ball for the majority of the game and patiently wearing down the opposition. Soccer is a relatively low scoring game (an average of 2.62 goals per game for both teams combined in the 2012 MLS Season sample), so an entire match analysis cannot be done based on the final score alone. Because of this, the basic statistics of possession, shots, and total passes have been used for years to study the game. These statistics have provided little statistical support in their correlation with success because of the differences in play style (Jones, James, & Mellalieu, 2004). As arbitrary as play style can be, new more complex measurements, such as expected goals (Rathke, 2017), try to take out some of the game's subjectivity in hopes of building ratings that all coaches and spectators can agree upon.

This study attempts to use network theory to take subjectivity out of the game as well. Network theory is a way to analyze graphs and their connections to better understand the entire network. Network theory can be applied to a soccer team's passing network to better understand the team's passing intricacies. This analysis can put a numerical value on the passing network, which can be compared across a season, or against an opponent. There are many practical applications of this type of analysis for coaches, players, general managers, scouts, and even the average viewer. When looking at possession statistics specifically, the entire picture cannot be told by a simple percentage of the time that one team has the ball versus the other. The basic possession statistic does not account for where on the pitch the team has the ball (Jones et al., 2004). Because of the stylistic issue raised previously, simply counting the number of passes a team has does not have correlation with success. Even with a high number of the passes completed, these passes may be completed in the defensive third of the field and have

little chance of leading to a goal. This paper proposes a solution to the shortcomings of basic passing statistics with a more practical measure called tuned passing influence (TPI). By tuning a passing rating based on where on the field the pass is completed, this rating would theoretically be a fix to the basic possession statistic and number of total passes statistic. TPI applies the idea that if a team completes passes higher up the field they have a higher likelihood of scoring on their opponent. TPI would be a movement toward the idea that not all passes are completed equal (Singh, 2018). The application of a weighted network allows the rating system to account for this very idea. The proposed TPI rating also accounts for how many unique teammates a player passes to throughout a game. This study has found that even with the implementation of a weighted network that gives more value to passes completed higher up the field, the proposed passing rating still has very small, and even a negative correlation with success in the game of soccer. TPI, however, can be used as a building block for future studies in analysis of soccer. As the rating has successfully applied a score to the network interaction a team uses throughout a game, the rating could be used in coordination with a stylistic quantification of play, the TPI rating would add value to certain quantifications of style.

Literature Review

The use of big data, specifically positional data, has been paramount in the increased analysis of sports data (Fry & Ohlmann, 2012). Positional data includes data captured in the form of X-Y coordinate player positions on the coordinate plane which can be coded as a field (Memmert & Raabe, 2018). In soccer, many applications of X-Y coordinate data have allowed literature to provide a lack of statistical support for the quality of currently employed key performance indicators, such as shots and passes and their correlation with success. More complex metrics, such as expected goals (xG) (Rathke, 2017), have provided more statistical support for a correlation with success as they are built out of the new spatial data. Expected goals measures the quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance. A measure such as expected goals is only calculatable because of X-Y coordinate data. Each element that makes up the total rating uses position on the field, and position of opposition, to quantify the quality of the opportunity to score.

The introduction of digital X-Y coordinate data naturally lends itself to network theory because the field is coded granularly as one large coordinate plane. Social network scholars use network theory to capture relationships between nodes (Opsahl, Agneessens, & Skvoretz, 2010). Network theorists study graphs as a representation of either symmetric relations or asymmetric relations between discrete objects. Network theory applications exist throughout society in the form of railroads, phone lines, electronic circuits, and social connections. To better understand more complex relational states

between nodes and edges, scholars have defined a weighted network (Opsahl et al., 2010). In doing so, the literature has extended the rudimentary concept of degree to the more complex idea of tie strength. Node relationships can be better understood when defining a weighted network in which ties are not just present or absent, but have some form of weight assigned to the edge that forms the connection (Granovetter, 1973). Network theory analyses such as closeness, degree, pagerank, and clustering were studied in the context of soccer to determine whether a team with higher measurements was more successful (Pena & Touchette, 2012). With a small sample size, Pena et al. showed the applications of these variables as a possible indicator of success. Early applications of soccer passing network scores centered around static passing networks. These networks provided an indication of key individuals, and a visual representation of key areas on the pitch. Furthermore, they allowed opposing teams to find weaknesses in overall connectivity (Pena et al., 2012). Edge connectivity, which is defined as the minimum number of edges one needs to remove in a network to make the network disconnected, was another score that has been studied throughout the literature (Pena et al., 2012). In theory, edge connectivity would provide value when applied to the game of soccer because as a team has a higher edge connectivity score, they have a more varied network of passes. Pena and Touchette's static model fails to account for position on the pitch where the pass is completed. Field position can be an indication of the difficulty of the pass as there is an increasing number of defenders as the ball progresses up the field. Research progressed to consider a dynamic passing network, which is based on an aggregated model where growing time windows are considered (Barghi, 2015). Eigenvector centrality was explored looking at the 2018 FIFA World Cup (Hamilton,

2018). Eigenvector centrality identifies which individual players are most important to their team's passing network. The study points out however that it is rare to see out-and-out forwards as central to the passing network. Only three forwards appeared in their team's top three eigenvector centrality scores in the 2018 FIFA World Cup data set. Usually, the eigenvector centrality metric pulled deep-lying players such as center backs and holding midfielders the study notes.

This study hopes to build upon current network theory soccer passing analyses and create a metric that uses a tuning measure to give a higher value for a pass completed higher up the field. This tuning measure will be experimented with to find if passes higher up the field do indeed have more correlation with success. Different attempts at weighted networks have been built using attack, defense, backward, sideways, and long ball weighted networks (Singh, 2018). Many of these early models have been built using a continuous weight that changes throughout the game. This paper will use the average passing position throughout the match to determine the weight for that individual player. When calculating the total team passing rating, the weight (calculated by the average passing position) will be used for every pass completed by that individual. The dependent variable of success for a given team has been found to have a positive correlation with the team's associated city, college, and even stock exchange (Pope & Pope, 2009) (Ashton, Gerrard, & Hudson, 2003). The paper hopes to find the level of correlation between success and a rating that accounts for not only the number of teammates a player passes to, and the number of passes between each unique passer and receiver, but also where on the field these passes take place. The link between a more robust passing network, as indicated by Tuned Passing Influence (TPI), and success will be explored in this paper.

Defining Success

The only way to win games is to score goals. This concept presents a problem for defining success. There is an inherent randomness to the timing of a goal, which is the means by which teams win games. Goals are hard to predict because the same chain of events does not always lead to the same outcome. Passing motifs are a way to categorize passing outcomes. A passing motif is a chain of completed passes between teammates. If Player A passes to Player B and then Player B passes to Player C, the passing motif would be written as ABC. As some goals are the result of a long passing motif (Gyarmati, Kwak, & Rodriguez, 2014), other goals are the result of an individual duel (Player A may win possession deep in his attacking third) won high up the field leading to a shot just seconds after possession is won. Furthermore, a specific passing motif of ABCB, where A, B, and C and individual players on the field, leading to a goal once in a match can be replicated in the following series and may not create the same result. Goals follow a relative Poisson Distribution. Though some research has shown that when a goal has been scored, the team who just conceded gets a relative boost (Capgemini, 2016), goals are still largely randomly distributed throughout a game. Therefore, it is clear goals cannot be the only measure of the success of a passing network. A commonly implemented methodology in sports is that the best defensive strategy is keeping the ball offensively. However, for soccer, the ability to retain possession of the ball for prolonged periods of time has not been shown to have statistical correlation with success. Possession characteristics have been shown to change throughout a match as the losing team retains more possession. Possession may be indicative of more skilled players and not

necessarily indicative of strategy differences (Jones, 2017). Possession alone, therefore, cannot be the sole indicator of success. Another commonly implemented success statistic is goal differential, which can be summarized as a total of all goals and goals against throughout a season with +1 for a goal scored and -1 for a goal conceded. However, goal differential is only positive when a team wins a game, and winning a game by more than one goal still does not give the team more points. Because of these reasons, this paper will define success in a very specific way. Success will be defined as a win, tie or loss. Represented numerically: three points for a win, one for a tie and zero for a loss. This measure attempts to take into account for the failures of the other success measures discussed.

Theory

There are many practical applications of network theory in the world of sports, specifically passing from one player to another in team sports that involve ball movement, such as in the case of soccer. The paper will define a passing network to be a series of edges and nodes where a node represents a player and an edge between two nodes represents a successfully completed pass between the two players. The sum of this network is defined as the graph (Appendix A). This paper looks to answer the question does more network interaction, which produces higher network theory measurements, lead to more success in the game of soccer?

Network Theory Measurements

This paper will consider network theory measurements by computing a global network invariant (Pena et al., 2012). Global invariants numerically describe the team, while local invariants provide numerical insights into an individual player's contribution to the overall network. Previously referenced was eigenvector centrality which is a local invariant (Hamilton, 2018). Such a rating would allow a team scouting a player to give value to that player's impact on his team's passing network. To determine the global invariant this paper adopts an Opsahl et al. equation, applying it to soccer passing networks, to create a variable we will call Tuned Passing Influence (TPI). TPI is a combination of how many total passes are completed, and the number of players passed to, with a tuning variable. To build TPI, measures of degree and edge strength are used (Appendix B). These two values will be the building blocks of the final equation of

degree centrality used to determine the global invariant measurement. Other measurements of closeness, betweenness, pagerank, and clustering have been explored in previous literature (Pena et al., 2012). Research has found a high correlation between high scores in closeness, pagerank and clustering and the overall media and scouting professional's perception of a player's quality of play in the tournament. Though the analysis of local invariant scores is thorough, the lack of an objective variable given to the entire team network is a shortcoming of other literature. Even many local invariant scores are skewed because the position of a pass is not considered. For example, low local invariant scores for attacking players are common as a result.

Degree Centrality

$$TPI = DC^{\alpha}(i) = k_i \left(\frac{s_i}{k_i}\right)^{\alpha} \quad (1)$$

Table 1:

Degree Centrality Building Blocks	
k_i	Total number of unique teammates the given player has successfully passed to
s_i	Total number of passes completed between the given player and the unique teammate
a	Tuning Measure to account for position on the field the pass was completed into

To combine both degree and edge strength, Opsahl et al. created an adapted version of the equation above. For this paper, $DC^{\alpha}(i)$, which is the measure of Degree Centrality subject to the given node *i,* is created to determine the value of pass subject to its location on the field. Effectively, without the application of *alpha,* $DC^{\alpha}(i)$ would consider each pass on the field to be equal. Practical application of the equation to soccer would reason that a pass from one center back to another is less important and difficult than a pass from one striker to another in the opponent's 18-yard box. Without *alpha,* TPI fails to give weight for where on the pitch the pass took place. To take this into account we use the tuning measure *alpha*. The tuning measure allows for an analysis of where on the pitch the number of passes completed or the number of unique players passes to is more important. An underlying assumption is that it is more important for players higher up the pitch to complete a high volume of passes sacrificing number of unique players passes to, while players in defensive positions have a greater influence on the passing network if they pass to a higher volume of unique players, sacrificing the number of total passes completed. To account for the assumption that passes higher up the field have a higher impact on success, an inverted TPI score is calculated.

When $a = 0$ and $\left(\frac{s_i}{k_i}\right) = 1$, TPI only accounts for the number of players passed to. When $0 < a < 1$, the number of players passed to is more important than the number of passes. When $a = 1$ then $TPI = S_i$. When $a > 1$, the number of passes are more important than the number of teammates passed to. The idea behind *alpha* is to award players for completing passes higher up the field to a variety of different teammates. What this measure hopes to avoid is allowing for center backs who are passing mainly to one another or to their keeper, with a low chance of the ball being intercepted, finishing games with their team's highest TPI ratings. Center backs certainly can finish with their team's highest TPI, but with the proper tuning measurement, they would be encouraged to complete a higher number of passes to a more varied group of teammates higher up the field. With this testable hypothesis in mind, this study will use a variety of tuning measures and run the model on each. Then, after looking at the results of the distinct models, the study will determine whether increasing alpha as passes are completed higher up the field has a higher correlation with success. If weighting the network doesn't matter, then the unweighted degree centrality rating and the weighted TPI scores should be statistically the same.

Methodology/Data

The data for this research project is courtesy of Opta, which is a data supplier for media, teams and other analytics firms. The data provided was from the 2012 MLS season when there were just 19 teams in the league. The data is Opta's F24 event level data which captures basic statistics like goals shots on target and passes, but with more details than a normal data feed. X, y & z coordinates are attached to each shot on target for example. This study's use of the F24 files allows for the creation of the TPI because each pass has a coordinate associated with the passer and the receiver. Because of limited resources, this study was unable to use the F7 Opta data feed which would give the player names associated with each player ID. In the case of this study, players are anonymous. To create the TPI rating, an R script was written to parse the F24 .xml files. Once the data was arranged into a data frame, the TPI rating (profiled in the degree centrality heading) was created using for-loops to build individual TPI scores for each player. The individual scores were then added to build a team TPI score for each match.

Variation of Alpha

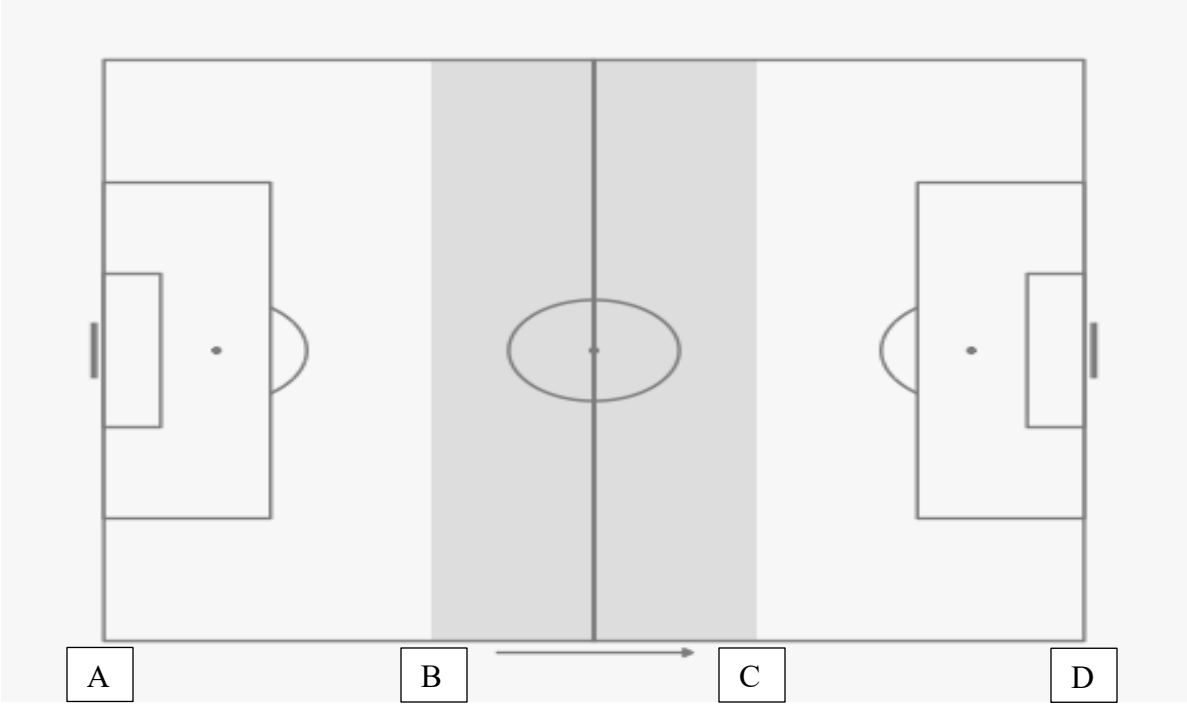
In Opsal et al, the alpha tuning parameter allows the user to weight the network. In the case of this study, multiple alphas were experimented with to build the four different TPI scores.

Table 2: TPI Weighting System

TPI Weighting System				
TPI Name	Baseline	Weight 1	Weight 2	Reverse Weight
TPI Number	TPI Base	TPI 2	TPI 3	TPI 4
Alpha Associated	1, 1, 1	.5, 1, 1.5	.25, 1, 1.75	1.5, 1, .5

In Table 1, different variations of alpha are shown for the TPI rating where $TPI = DC^{\alpha}(i) = k_i \left(\frac{s_i}{k_i}\right)^{\alpha}$ This alpha represents the weight that will be put on an individual based on their average passing position throughout a match. This study uses average passing position as an alternate to a continuous weighted network (Singh, 2018). The application of an average passing position may have a higher correlation with success, which will be explored in this study. The line titled *Alpha Associated* in Table 2 represents (in order) the alpha assigned if the player's average passing position is in the defending third, middle third, or attacking third of the field. Figure one displays this concept. The team is attacking from left to right for this graphic.

Figure 1: Weighted Alpha Zones



When looking at Weight 1 (TPI 2) in Table 2 with *Alpha Associated* of .5, 1, 1.5, the weighting system would work as follows. If a player's average passing position throughout a match is between line A and line B in Figure 1, then the alpha score assigned to that player is a .5 in the case of TPI 2. If a player's average passing position throughout a match is between line B and line C in Figure 1, then the alpha score assigned to that player is 1 for TPI 2. If a player's average passing position throughout a match is between line C and line D in Figure 1, then the alpha score assigned to that player is a 1.5 for TPI 2. Because the Opta data is formatted on a 0 to 100 left to right x coordinate grid, this study used a cutoff of 0 to less than 33.3 for the A to B zone, 33.3 to less than 66.6 for the B to C zone, and greater than or equal to 66.6 for the C to D zone. This study did not look at Y coordinates as a potential weight to the network. To choose the alphas that are to be assigned to the four different TPI scores, different variations

were hypothesized. First, to create a baseline non-weighted network, an alpha of 1 was given to players in each passing zone. The TPI Base was then calculated using those scores. To test the hypothesis that the level and location of network interaction in soccer that a team employs affects success, three different weighted TPI scores were constructed. TPI 2 was the first weighted network measurement and was built using a difference of .5 for each different average passing position zone. Because these differences are arbitrary, TPI 3 was built using a weighted network with a greater spread of .75 between each different average passing position zone. TPI 4 was constructed using a decreasing alpha. Instead of increasing the TPI score as players complete passes higher up the pitch, the score penalized players for completing passes higher up the field using 1.5 for the defensive third alpha, 1 for the middle third alpha and .5 for the attacking third alpha. The reversal of the theorized alpha accounts for the assumption that passes further up the field have more impact on success than passes completed in the defensive zone. To note is the idea that TPI 4 and TPI 2 could be negligible if there were a similar number of passers in the defending third and attacking third that completed a similar number of total passes.

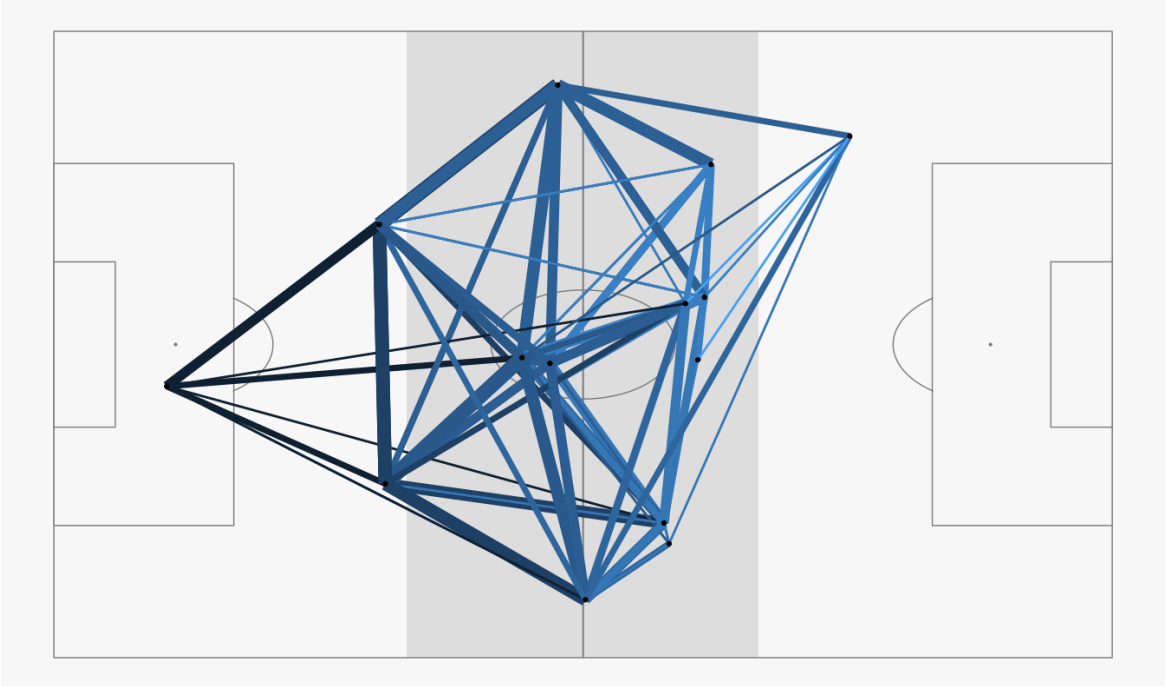
Table 3: Summary of the four proposed variations of the TPI Rating System

Summary of TPI Variables								
Variable	N	Mean	Standard Deviation	Min	1st Quartile	Median	3rd Quartile	Max
TPI Base	668	348.68	83.78	161	288.5	340.5	401.5	600
TPI 2	668	338.03	85.16	152.45	275.01	333.24	389.43	677.77
TPI 3	668	337.37	89.29	149.42	271.94	329.97	388.81	775.86
TPI 4	668	376.15	96.5	165.6	305.19	366.67	442.31	670.02

Visualization

To better understand the TPI rating system a visualization is helpful. Though not currently very common in the public sphere, passing network visualizations are used widely by soccer analysts looking to visualize an opponent's, or their own team's, passing outlook. Passing networks are helpful in identifying key players that make the network tick. When using an average passing position, as this study does, one can visualize passing tendencies of players. Maybe a winger likes to start wide on the touchline and then come inside and complete a pass regularly. If one were to just look at this player's starting position throughout a match, this tendency may not be possible to visualize. In coding this visualization, an edge data frame was created using each pass as an edge. The thickness or strength of the line between the set of nodes (players) was determined by how many passes were completed by that unique set of players.

Figure 2: A passing map built for the opening game of the 2012 MLS Season between the Colorado Rapids and Columbus Crew. The sum of all the Rapids passes from left to right create the passing map below.



In Figure 2, a passing network is built for the 14 Colorado Rapids players who played on the field on opening day against the Columbus Crew. The thickness of the line ranges from one pass completed all the way to 12 passes completed by the most connected pair of players. The reason for their being 14 players indicated instead of 11 players is because the Rapids made all three of their available substitutions throughout the match. The use of substitutions throughout a match causes inherent problems in passing maps. Some studies complete their passing map when the first substitute is made by a given team (Singh, 2018). There are inherent issues with this technique. For example, if a substitute is forced on a team 5 minutes into a match because of injury, the passing map would show an extremely limited sample size of the game's total passes. However, if a

map uses all of the substitutes alongside each starter, the passing map can become confusing as a like-for-like substitute will have a relatively similar average passing position as evidenced above. The shading placed on Figure 2 ranges from dark to light as passes are completed higher up the field. Notice, that because each given player's alpha rating is built by their average passing position, the map attempts to show that even passes completed to players with significantly higher average passing positions are given the same shading color. Each individual pass is made independent of the outcome of the passing motif. If a pass is completed from one player to another, even if it is lost on the next touch, it is shown on the map as a completed pass and goes into the overall TPI score. Another benefit of this study's TPI metric is that because it is a global invariant instead of a local invariant, even if a substitute is made, the TPI rating is affected independently. Another potential solution to the substitute problem is to treat each player that substitutes into a game as a continuation of the player they subbed in for. As long as a field player is not substituted for a keeper who was given a red card, the treating of a player who subbed into the game as the player they subbed in for would continue to build the passing network, whether the rating was built globally, locally, or non-weighted.

There are also some major issues with current passing network visualizations. One major issue is that each network is static. There is no reflection of time or movement. Whether the passing network visualization is built using the average passing position or average receiving position, or even the starting position, the map does not reflect all of the movement the player made throughout the game. Another issue is that as a game changes, whether it be a goal scored, player substituted, or formation shift, player movement is affected. A pass completed in the first ten minutes of the game does not

hold the same weight as a pass completed in the ninetieth minute. An issue this study's passing map application fails to account for is an unsuccessful pass. Some passing networks have been built to account for an unsuccessful pass using a red arrow as an unsuccessful pass and a green arrow as a successful pass. Issues with this technique arise as a network becomes muddled with many lines throughout the network. Another issue with passing networks is there is no room for stylistic setup and changes. As different managers set their team up differently, the attempt to score, or not to concede, changes based on the style. A passing network, as a static map, cannot account for these stylistic inputs. The future may lie in the use of radar plots and passing networks in coordination. As a radar plot is the best current attempt at putting an objective value on style and passing networks, such as this study's proposed TPI, are the best attempt at putting an objective value on a pass, the two visualizations marriage is likely in the future.

Econometric Model

To test if there was a statistically significant difference in each of these four TPI ratings, paired t tested were constructed for each combination of ratings.

The t-tests showed there was not significant difference between the means of TPI 2 and TPI 3. TPI 3 was not included in the final model because of this finding. A multinomial probit model was chosen because of its ability to include a categorical dependent variable. The multinomial probit model was run using the following hypothesis test and equation.

$$\text{Success} = \beta 0 + \beta 1 \text{Tuned Passing Influence} \quad (2)$$

H_a: TPI and success have a positive significant relationship.

H_o: There is no statistically significant relationship between TPI and success.

The results of the tests that were done for this model to check for economic problems can be seen in Appendix C.

Model

As stated previously, the best econometric model for this study was determined to be a multinomial probit model. Because the dependent variable of success has three unique categories that it can fall into (win – three points, tie – one point, and loss 0 points), a categorical model was necessary.

This study also chose to only look at the home team when assessing the model for a number of reasons. First, the result of the game for the two teams playing is not independent. In a given game, only one team can receive three points. In this scenario, the opposing team is automatically awarded zero points. If one team were to receive one point, the opposing team will automatically receive one point as well. The exclusive use of the home team also takes away the variability of the home team winning more often. In the 2012 MLS data set, the home team won 25.1% more of the time than the away team. Only using the home team for the multinomial probit model also allows for the base outcome to be a win – three points, when looking at the regression output.

Table 4: Multinomial Probit Model for TPI Base. TPI Base is the unweighted TPI measure.

Output of Regression 1					
	Success	Coefficient	Standard Error	z	P>|z|
0	TPI Base	0.0017805	0.012841	1.39	0.166
	Constant	-1.25079*	0.4889932	-2.56	0.011
1	TPI Base	-0.0021115	0.0013042	-1.62	0.105
	Constant	0.1328717	4800953	0.28	0.782
3	(Base outcome)				
Note *p<.05, ** p<.01, ***p<.001					

Table 3 shows a base outcome being a win for three points. This is because the home team won 51.2% of the 334 games sampled. Neither of the loss (0 points) or tie (1 point) success outcomes were significant to the 95% confidence interval. The TPI Base loss coefficient was slightly positive, showing that a higher TPI Base score led to a slightly higher chance of losing the game. The TPI Base tie coefficient was slightly negative, showing that a higher TPI Base rating leads to slightly less chance of tying the game for the home team.

Table 5: Multinomial Probit Model for TPI 2

Output of Regression 2					
	Success	Coefficient	Standard Error	z	P>|z|
0	TPI2	0.0017559	0.0012257	1.43	0.152
	Constant	-1.225359**	0.456962	-2.68	0.007
1	TPI2	-0.0017712	0.00126119	-1.4	0.16
	Constant	-0.0049138	0.454458	-0.01	0.991
3	(Base outcome)				
Note *p<.05, ** p<.01, ***p<.001					

Similar to the unweighted TPI Base Regression, either of the loss (0 points) or tie (1 point) success outcomes were significant to the 95% confidence interval in the TPI2 weighted network measure. The TPI 2 loss coefficient was slightly positive, showing that a higher TPI 2 score led to a slightly higher chance of losing the game. The TPI 2 tie coefficient was slightly negative, showing that a higher TPI 2 rating leads to slightly less chance of tying the game for the home team.

Table 6: Multinomial Probit Model for TPI 4

Output of Regression 3					
	Success	Coefficient	Standard Error	z	P>|z|
0	TPI4	0.0010445	0.0011277	0.93	0.354
	Constant	-1.007184*	0.4637061	-2.17	0.03
1	TPI4	-0.0020858	0.0011418	-1.38	0.068
	Constant	0.1829645	0.4543863	0.4	0.687
3	(Base outcome)				
Note *p<.05, ** p<.01, ***p<.001					

The TPI 4 multinomial probit model regression showed that there was a higher chance of a loss when TPI 4 score grew. The chance of a tie is slightly negatively correlated with an

increase in TPI 4. The P-value of the TPI 4 tie coefficient was very close to being significant at the 95% confidence interval but fell just above the .05 threshold.

Table 7:

Marginal Effects TPI Base				
Success	dy/dx	Standard Error	Z	P>|z|
0	0.0006068	0.0002777	2.19	0.029
1	-0.0006452	0.0002749	-2.35	0.019
3	0.0000384	0.0003292	0.12	0.907

Table 8:

Marginal Effects TPI 2				
Success	dy/dx	Standard Error	Z	P>|z|
0	0.0005731	0.0002656	2.16	0.031
1	-0.0005652	0.0002679	-2.11	0.035
3	-0.000000193	0.0003167	-0.03	0.98

Table 9:

Marginal Effects TPI 4				
Success	dy/dx	Standard Error	Z	P>|z|
0	0.0004296	0.0002464	1.74	0.081
1	-0.0005777	0.0002413	-2.39	0.017
3	0.000148	0.0002884	0.51	0.608

When looking at the marginal effects shown in the tables 7-9, there is statistical evidence to reaffirm the findings of the multinomial probit regression outputs. For all three model outputs, there is a positive correlation with TPI and the loss success outcome. In both the TPI Base and TPI 2 models, both the loss and tie variables are significant to the 95% confidence interval.

Conclusion

This study hoped to find out if a more robust network theory passing measurement led to success in the game of soccer. As possession and other passing measures have failed to show a positive correlation with success in the previous literature, this study hoped to improve by building in an alpha variable that accounted for where on the field the player was completing a pass from. This study has found that even with the implementation of an alpha that gives more value to passes completed higher up the field, the proposed passing rating still has very small, and even a negative correlation with success. This continues to confirm the idea that because different teams implore different strategies to meet the same end (success), current team passing ratings fail to be positively correlated with success. A way to improve this model would be to somehow quantify play style. This is a limitation of this current model because the data is over six seasons old and quantifying play style is extremely difficult. With the quantification of play style, a study could build dummy variables for teams that implore a counter-attacking style for example. A limitation of this methodology would be that teams change play style based on opponent and personnel (injuries or transfers) throughout a given season and even throughout a given game.

There are many practical applications of TPI for coaches, players, and media. As the metric could be built in real-time, coaches could get a visual snapshot (as seen in Figure 2) of their entire passing network throughout a game. Coaches could then make substitutes objectively based on who has shown a high TPI in previous matches with the

personnel currently in the match. Players could also get a snapshot of where on the pitch they may hope to find more passes in to improve their TPI rating, and thus help the team build a more robust, difficult to disconnect network. Opposing teams could use the metric to find areas of the field where a their opponent has low connectivity. With that in mind, those areas could be targeted to squeeze teams high up the pitch and press in areas where the opposition has weak connectivity. Opposing teams could also use TPI to find which opposing players are most important for maintaining the teams passing flow. A team could then try to disrupt the entirety of the network by assigning their best defender or defensive midfielder to mark that player for the game. TPI could also be used to identify certain players in the transfer market who have demonstrated high TPI scores in certain areas of the field where the searching team currently has a low team TPI score.

Shortcomings/Future Studies

The general shortcomings of passing networks was discussed previously. More broadly, however, there are a number of shortcomings of this study. When looking at variables that lead to success in the game of soccer, home team, shots and goals have statistical support of being highly correlated with success. Passing and time of possession have been largely hypothesized to have little statistical significance with success. Because the TPI rating system was built from the relatively weak base of passing statistics, correlation with success was unlikely. As this study looked to improve in some manner upon a basic statistic of total passes and time of possession, and a weighted network was constructed using a network theory analysis to build a weighted degree centrality measure. There are a number of other network theory measurements that could

be weighted in order to build a weighted network. Another future study could weight more complex measures of betweenness or closeness. An isolated score of either weighted degree or weighted total edge strength could be studied. A future study could focus on the Y coordinate of the pitch and build a weighted network measure based on Y coordinate position instead of X coordinate position. Other applications of weighted networks could weight based not on field position, but instead could weight based on pass length. Another study could be constructed by building a weighted network based on the end passing motif result. For example, if the passing motif led to a shot, or to a progression from one zone to another, the weight on each of those passes could be increased with a variability of the result. Another application of weighted passing networks could be a system that creates a continuous weight for where on the field the pass was completed. Instead of a basic three zone application that creates three distinct weights that the player can be assigned to, the weight would increase continuously as the passes are completed higher up the field. Passes would then also not have to be calculated based on the average passing position of that individual player throughout a game, but could be constructed in real time based on where the individual passes were completed to, or from. The holy grail of a passing network study would be to somehow combine a variety of network theory scores, but also somehow account for style employed. This would be extremely difficult because style changes not only throughout a season but throughout a game, based on a variety of factors.

Another shortcoming of the study comes from the data. The MLS, like any other league, has a certain play style that is more common. As this model was only run on the 2012 MLS season, a future study could be done using the TPI rating system on a top-tier

European league. Different results may be found in La Liga (Spanish First Division), for example, because of the passing intricacies of a club like FC Barcelona.

Future studies of weighted passing networks could be applied when looking at multiple years of data. If a study were to attempt to find the improvement in weighted network theory interaction, a multiyear study of a managerial change could be created. As a manager hopes to employ a possession-based style, the right personnel is key. As he begins to assemble a team of players that he feels would best utilize his style, one would assume that weighted network interaction would improve over time. This study could put an objective value on the buy-in from players to play with a certain style.

References

Ashton, J. K., Gerrard, B., & Hudson, R. (2003). Economic impact of national sporting success: Evidence from the london stock exchange. *Applied Economics Letters, 10*(12), 783-785. doi:10.1080/1350485032000126712

Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A., & Giorgio Parisi. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(11), 3747-3752. doi:10.1073/pnas.0400087101

Brooks, J., Kerr, M., & Guttag, J. (Aug 13, 2016). Developing a data-driven player ranking in soccer using predictive model weights. Paper presented at the 49-55. doi:10.1145/2939672.2939695 Retrieved from http://dl.acm.org/citation.cfm?id=2939695

Capgemini. (2016). Are football teams most vulnerable after they've just scored? Retrieved from https://www.capgemini.com/gb-en/2016/09/are-football-teams-most-vulnerable-after-theyve-just-scored/

Duch, J., Waitzman, J.S., & Amaral, L.N.,. (2010). Quantifying the performance of individual players in a team activity. *PLoS One, 5*(6), e10937. doi:10.1371/journal.pone.0010937

Fry, M. J., & Ohlmann, J. W. (2012). Introduction to the special issue on analytics in sports, part I: General sports applications. *Interfaces, 42*(2), 105-108. doi:10.1287/inte.1120.0633

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380. doi:10.1086/225469

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). *Searching for a unique style in soccer* Retrieved from https://www.openaire.eu/search/publication?articleId=od_____18::868ba21ec20c424e1e95bd74fcb5cf32

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*(3), 245-251. doi:10.1016/j.socnet.2010.03.006

Jones, P.D., James, N., & Mellalieu, S.D. (2004). Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport, 4*(1), 98-102. doi:10.1080/24748668.2004.11868295

Peña, J. L., & Touchette, H. (2012). *A network theory analysis of football strategies* Retrieved from https://www.openaire.eu/search/publication?articleId=od_____18::8d66a90519b5fcfe12bc6fcda75300ba

Pope, D. G., & Pope, J. C. (2009). The impact of college sports success on the quantity and quality of student applications. *Southern Economic Journal, 75*(3), 750-780. Retrieved from http://www.econis.eu/PPNSET?PPN=593055764

Power, P., Ruiz, H., Wei, X., & Lucey, P. (Aug 13, 2017). Not all passes are created equal. Paper presented at the 1605-1613. doi:10.1145/3097983.3098051 Retrieved from http://dl.acm.org/citation.cfm?id=3098051

Rahnamai, Barghi, A. (2015). *Analyzing dynamic football passing network* doi:10.20381/ruor-6796 Retrieved from http://www.dx.doi.org/10.20381/ruor-6796

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise, 12*(Proc2) doi:10.14198/jhse.2017.12.Proc2.05

Ritsuko, K., Lawrence A, R., & 菊澤, 律. (2018). 1. introduction. *Senri Ethnological Studies = Senri Ethnological Studies, 98*, 1-8. Retrieved from https://www.openaire.eu/search/publication?articleId=jairo_____::5132a00fb3dd1a0cfe0f68c14f936315

Singh, K. (2018,). Interactive passing networks: Uncovering the hidden potential of passing networks through interactive visualizations. Retrieved from https://karun.in/blog/interactive-passing-networks.html

Szczepanski, L. (2008). Measuring the effectiveness of strategies and quantifying players' performance in football. *International Journal of Performance Analysis in Sport, 8*(2), 55-66. doi:10.1080/24748668.2008.11868435

Appendix A

A graph *G* will be defined as a finite nonempty set *V* of objects called vertices or nodes and a set *E* of 2-element subsets of *V* called edges. This basic concept of a graph is expanded upon to form the basis of network theory, a subset of graph theory, where edges and nodes make up a graph where $G = (V, E)$. Given *G*, network theory expands to create variables to tell a more complex story of *G*.

Appendix B

Degree. The degree of a vertex *v* in a given graph G is the number of edges incident with *v* and is denoted by $\deg_G v$ or simply by deg *v* if the graph G is clear from the context. Deg *v* is the number of vertices adjacent to *v*.

$$k_i = DC(i) = \sum_{j}^{N} x_{ij}$$

k_i represents the degree of the given vertex, which is defined as the total number of nodes N connected to the given node *i*. X is the adjacency matrix in which the cell x_{ij} is defined as 1 if node *i* is connected to node *j.* An adjacency matrix of G will be defined as the *n x n* matrix $A = [a_{ij}]$, where

$$a_{ij} = \begin{cases} 1 & \text{if } v_i v_j \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

while the incidence matrix of *G* is the *n x m* matrix B = $[b_{ij}]$, where

$$b_{ij} = \begin{cases} 1 \text{ if } v_i \text{ is incident with } e_j \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

Edge Strength. As an extension of degree Opsahl et al. creates a variable that represents the sum of the number of weighted edges connected to a given node *i* and labels this node strength.

$$s_i = DC^w(i) = \sum_j^N w_{ij}$$

where *w* is the weighted adjacency matrix, in which $w_{ij} > 0$ if the node *i* is connected to the node *j*, and the value represents the weight of the tie. This would also be the definition of degree if the s_i measure were implemented in a binary network where each edge has a weight of 1. Because we want our edge weights to account for edges that are greater than one, the formula is adapted where $w_{ij} > 0$. This methodology has become standard for analyzing weighted networks (Barrat, Barthélemy, Pastor-Satorras, Vespignani, & Parisi, 2004) For the application to soccer, our value s_i will represent the number of passes completed from one player to another.

Appendix C

Correlation Matrix				
Variable Name	TPI Base	TPI2	TPI3	TPI4
TPI Base	1			
TPI2	0.9743	1		
TPI3	0.9450	0.9935	1	
TPI4	0.9453	0.8550	0.8066	1

When checking for multicollinearity, the correlation matrix above found all of the rho values to be above .5 other than the correlations with goals.

The model also tested and found an adjusted chi-squared value of 41.22, which is greater than 5.99 which is the chi-squared at 2 degrees of freedom and a 5% significance level. After transforming each variable in a variety of ways, it was determined that this is a limitation of the model as the chi-squared value never dropped below 41.22 with any of the transformations. This can be accounted for by the omitted variable bias which is present in the model.