HOCKEY POOLS FOR PROFIT: A SIMULATION BASED PLAYER SELECTION STRATEGY

by

Amy E. Summers

B.Sc., University of Northern British Columbia, 2003

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Fall 2005

APPROVAL

Name: Amy E. Summers

Degree: Master of Science

Title of project: Hockey Pools for Profit: A Simulation Based Player Selection Strategy

Examining Committee: Dr. Richard Lockhart

Chair

Dr. Tim Swartz

Senior Supervisor

Simon Fraser University

Dr. Derek Bingham

Supervisor

Simon Fraser University

Dr. Gary Parker

External Examiner

Simon Fraser University

Date Approved: Sept. 29/2005

Abstract

The goal of this project is to develop an optimal player selection strategy for a common playoff hockey pool. The challenge is to make the strategy applicable in real time. Most selection methods rely on the draftee's hockey knowledge. Our selection strategy was created by applying appropriate statistical models to regular season data and introducing a reasonable optimality criterion. A simulated draft is performed in order to test our selection method. The results suggest that the approach is superior to several ad-hoc strategies.

"You have brains in your head.

You have feet in your shoes.

You can steer yourself any direction you choose.

You're on your own. And you know what you know.

And YOU are the guy who'll decide where to go."

— Dr. Seuss 1904-1991

Acknowledgments

I would like mention the people who helped to make this project a success. First, I would like to thank my supervisor Tim Swartz. Tim was the instructor of one of the first courses I took at SFU. I quickly learned to appreciate his “tough but fair” teaching style. He challenged me and at the same time helped me to increase my confidence. From the beginning of this project Tim was always willing to discuss, encourage and give valuable insight to the topic.

I also wish to thank my initial “partner in crime” Maria Lorenzi. The foundation of this project was a joint effort. The initial framework helped to give shape to this project and give clear direction to the writing process. In addition, this project would not have been possible without the valuable insights of “hockey guru” David Beaudoin.

Last but certainly not least I would like to thank my parents for their support and encouragement. I was instilled with the notion that I could do whatever I set my mind to, I would have been lost without my Mom’s “helpful reminders” and my Dad’s many words of wisdom; my personal favourite being “How do you eat an elephant? - One bite at a time”.

Many thanks to all of you. - Amy

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sports and gambling have been associated with one another for generations. From wagers on bare knuckle boxing and cockfights, sports gambling has evolved into a multi-million dollar industry. There are web sites dedicated to providing point spreads for virtually every game in any major sporting event (eg. www.pinnaclesports.com).

The appeal of sports gambling is not limited to die hard fans; even people who admittingly know next to nothing about sports are willing to wager a few dollars. There are office pools for major sports such as basketball, football and hockey. The popularity of office pools resides in the camaraderie between participants and the notion that winning is a matter of luck rather than skill. With little doubt the most popular office pools in Canada are hockey pools. Afterall, who has not heard the phrase "Hockey Night in Canada"?

The idea for this project came from the course STAT 890, Statistics in Sport, offered in the summer of 2004. One of the requirements was to research and present an original project on statistics in sport. One idea thrown out in a brainstorming session was to find a winning strategy for hockey pools, and since the National Hockey

League (NHL) playoffs were right around the corner it seemed like kismet. As the project progressed we discovered that it was beyond the scope of a class project but a fantastic idea for a Masters project.

Note that unlike some other sports pools, hockey pools are usually concerned with selecting players who accumulate points rather than selecting teams. Many people believe that a playoff hockey pool is a boom or a bust. Not only does one have to consider which players to pick, it is important to think about which teams will advance in the playoffs. Some general advice provided by hockey pool veterans is to begin by choosing the best players from the best teams and part-way through the draft, opt for high scoring players from teams which may play only one or two rounds. However, not everyone uses this strategy. Some people choose players based on very abstract qualities. People have chosen players based on the colour of their uniforms or the numbers that they wear. The most bizarre strategy that we heard of involved the selection of players with the last name Sutter. At the time, the six Sutter brothers may all have been playing in the league. When the draftee ran out of Sutters, he then opted for players with last names that sounded like Sutter, for instance Suter or Sutherland. The goal of this project is to develop an optimal player selection strategy by applying statistical methods to the data available. In addition, we want to be able to use this strategy in real time. This implies that any calculations performed during a draft must be fast.

There is a wide variety of hockey pools, some are available online to all takers while others are held between friends or co-workers. One particular type of hockey pool is a fantasy league. Often these are set up at the beginning of the regular season. Participants pick players (both skaters and goalies) to make up their fantasy teams and participants can choose the same players. Players are awarded points subject to a given scoring method. For example, one particular pool found online at www.bluerodeo.com/br/hockey.html had the following the scoring system:

- 1.0 point for each goal
- 1.0 point for each assist
- 0.2 points for each penalty minute
- 2.0 points for each win (goalie and team)
- 1.0 point for each tie (goalie and team)
- 0.5 points for each loss in overtime

Often fantasy leagues will specify positions that must be filled, for example, one must choose 2 goalies, 5 forwards and 3 defensemen. Bonus points can also be awarded for certain events. In the same on-line pool the following bonus points were included in the scoring system:

- 1.0 point for each shutout (goalies and team)
- 1.0 point for each shorthanded goal
- 1.0 point for each overtime goal
- 1.0 point for each game-winning goal (per player)
- 1.0 point for each hat trick (per player)

Playoff pools often tend to be smaller than fantasy leagues and usually follow slightly different rules. In particular, once a player is chosen he becomes ineligible and is removed from the draft. But really there are a multitude of different scoring systems that can be employed. In both fantasy leagues and playoff pools the team with the most points at the end is declared the winner.

Since the 2003-2004 NHL regular season had just ended when we began our class project we decided to limit our application to Stanley Cup Playoff pools. The NHL is organized into two conferences. The Eastern conference is subdivided into three divisions, Atlantic, Northeast and Southeast. The Western conference is also subdivided into three divisions, Central, Northwest and Pacific. The number of teams qualifying for the playoffs is sixteen, eight from each conference. The three division winners in each conference are seeded one through three and wild-card teams are seeded four through eight based on their regular season point totals. The first round of the playoffs has the first seed playing the eighth seed, the second playing the seventh, the third playing the sixth, and the fourth playing the fifth. At the end of the first round, the teams in each conference are reseeded as before, with the top remaining seed playing against the fourth remaining seed, and the second remaining seed playing against the third remaining seed. In the Conference Finals, the two remaining teams play each other, with the winners playing against one another in the Stanley Cup Finals.

Teams battle in best of seven series; that is to advance to the next round a team needs four wins, so at most seven games are needed to determine a series winner. In post-season play there are no ties, instead the result is decided by sudden death overtime. Twenty minute overtime periods are played until someone scores the final goal. Each series follows a 2-2-1-1-1 home-away schedule. Home-ice advantage is given to the higher-ranked team.

In chapter 2, we describe a common playoff pool that is the focus of this project. The statistical model used to describe hockey scores and individual player performance is explained and justified. The model is then applied to the playoff pool. A description of the steps and assumptions used to simulate the Stanley Cup Playoffs is given which includes different methods for obtaining team win probabilites as well as determining a player's potential worth. We also give a criterion for optimal drafting. We had planned to test our simulation based selection strategy by running our own Stanley Cup Playoff pool in the SFU Statistics and Actuarial Science Department. However, the 2004-2005 NHL regular season was cancelled because salary disputes between the players and the team owners could not be resolved. So we have postponed testing our player selection method until the lockout ends. In chapter 3, we present a simulation study designed to investigate our player selection method. We conclude with a discussion in chapter 4.

Chapter 2

Optimal Drafting

2.1 A Common Hockey Playoff Pool

As described in the Introduction, we are interested in developing an optimal (or nearly optimal) drafting strategy for hockey playoff pools. However, there are many different scoring systems and rules that impact the draft. The focus of this project is a common playoff pool with the following rules:

- The draft is of skaters only (i.e. no goalies)
- The scoring method is 1 point per goal and 1 point per assist
- The number of draftees, K, is fixed before the draft begins
- The number of rounds in the draft, m, is dependent upon K (more participants would mean fewer rounds)
- Once a player is drafted he cannot be drafted again
- There are no trades between draftees and no replacement players; if a player is injured that is too bad

- The draft order is randomized before the first round; afterwards the order is reversed in each subsequent round

In order to avoid confusion we use the term "lineup" to refer to a unique group of players drafted by a particular draftee and the term "team" is reserved for actual hockey teams.

2.2 Statistical Modelling

2.2.1 Hockey Scores

Possession of the puck is key, since whenever a team controls the puck they have an opportunity to attack the opposing team's net and score a goal. We assume that the probability of scoring a goal on a particular possession, p, is constant. Naturally, this is a simplification which does not account for situations such as power plays. Final scores in most hockey games are relatively low and there are many possessions; this supports the claim that p is quite small. In addition, we assume that all possessions of the puck are independent. Letting X be the number of goals scored by a team in the game, then $X \sim Binomial(n, p)$, where the number of possessions in a game, n, is large but unknown. Since p is small and n is large we set $\theta = np$ and apply the Poisson approximation

$$P(X = x) \approx \frac{e^{-\theta}\theta^x}{x!} \quad x = 0, 1, \ldots$$

The parameter θ can be interpreted as a measure of a team's offensive ability and its opponent's defensive ability. This model has been used previously (Berry, 2000) to investigate statistical applications in hockey. An advantage of the Poisson model over the Binomial model is that there is one less parameter.

2.2.2 Individual Player Performance

Consider player i who is the i^{th} player drafted to a lineup of m players. Let X_i be the number of points obtained by player i in a particular game. Then as before, we use the Poisson approximation (Berry, Reese and Larkey, 1999), and obtain

$$P(X_i = x_i) \approx \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!} \quad x_i = 0, 1, \ldots$$

where θ_i can be considered a measure of player i's ability. The parameter θ_i can be estimated by a combination of regular season results and subjective tweaking. A straightforward method for estimating θ_i is

$$\theta_i = \frac{\text{number of points obtained by player } i \text{ in the regular season}}{\text{number of games played by player } i \text{ in the regular season}}.$$

A possible improvement to this estimator is to emphasize a player's recent performances by giving more weight to the latter half of the season. Additional subjective modifications can be made based on personal knowledge. For instance, perhaps a player with a large θ breaks his leg in the last week of the regular season. Unable to play in the post season, you would not choose this player in the draft; therefore you could set his θ equal to zero. For the remainder of the project, we will assume that the θ's are known values that have been determined in some manner.

2.2.3 Playoff Pool Extension

Let Y_{ki} be the number of points accumulated in the playoffs by the player chosen in round i by draftee k. Then,

$$P(Y_{ki} = y) = \sum_{g_{ki}} P(Y_{ki} = y|g_{ki})P(G = g_{ki}) \qquad (2.1)$$

where g_{ki} is the number of games played in the playoffs by the i^{th} player chosen by draftee k. In the playoffs of the National Hockey League (NHL), there are a maximum of four best of 7 rounds which implies that the summation in (2.1) ranges from 4 to 28. Recall from the previous section that we used the Poisson approximation to model the number of points accumulated by a player in a single game. The random variable Y_{ki} is a sum of independent Poisson variables; that is

$$Y_{ki} \equiv X_{ki1} + \ldots + X_{kig_{ki}}$$

where X_{kij} is the number of points accumulated in the j^{th} game by the i^{th} player selected by draftee k. It is well known that a sum of independent Poisson variables is also Poisson; therefore $Y_{ki}|g_{ki} \sim Poisson(\theta_{ki}g_{ki})$. This, in turn gives the unconditional distribution of Y_{ki} in (2.1) as a finite mixture of Poissons.

2.2.4 Some Expectations

It turns out that various expectations are required for our drafting strategy. These expectations make use of the conditional expectation formulae. The expected number of points scored by the i^{th} player selected by draftee k is given by

$$\begin{aligned} E(Y_{ki}) &= E_{g_{ki}}(E(Y_{ki}|g_{ki})) \\ &= E_{g_{ki}}(\theta_{ki}g_{ki}) \\ &= \theta_{ki}E(g_{ki}). \end{aligned} \tag{2.2}$$

Next, we extend the calculations for a given lineup. Consider the total points accumulated by draftee $k = 1, \ldots, K$,

$$T_k \equiv Y_{k1} + \ldots + Y_{km}$$

where m is the number of rounds in the draft, or in other words, the number of players per lineup. Then,

$$\begin{aligned} E(T_k) &= \sum_{i=1}^{m} E(Y_{ki}) \\ &= \sum_{i=1}^{m} \theta_{ki}E(g_{ki}). \end{aligned} \tag{2.3}$$

Next,

$$
\begin{aligned}
Var(T_k) &= \sum_{i=1}^{m} Var(Y_{ki}) + 2\sum_{i<j} Cov(Y_{ki}, Y_{kj}) \\
&= \sum_{i=1}^{m} [E_{g_{ki}}(Var(Y_{ki}|g_{ki})) + Var_{g_{ki}}(E(Y_{ki}|g_{ki}))] + 2\sum_{i<j} Cov(Y_{ki}, Y_{kj}) \\
&= \sum_{i=1}^{m} [E_{g_{ki}}(\theta_{ki} g_{ki}) + Var_{g_{ki}}(\theta_{ki} g_{ki})] + 2\sum_{i<j} Cov(Y_{ki}, Y_{kj}) \\
&= \sum_{i=1}^{m} [\theta_{ki} E(g_{ki}) + \theta_{ki}^2 Var(g_{ki})] + 2\sum_{i<j} Cov(Y_{ki}, Y_{kj}). \qquad (2.4)
\end{aligned}
$$

We assume conditional independence when expanding the covariance term in (2.4). Therefore,

$$
\begin{aligned}
Cov(Y_{ki}, Y_{kj}) &= E[(Y_{ki} - E(Y_{ki}))(Y_{kj} - E(Y_{kj}))] \\
&= E[Y_{ki}Y_{kj} - E(Y_{ki})Y_{kj} - E(Y_{kj})Y_{ki} + E(Y_{ki})E(Y_{kj})] \\
&= E(Y_{ki}Y_{kj}) - E(Y_{ki})E(Y_{kj}) \\
&= E_{g_{ki}g_{kj}}[E(Y_{ki}Y_{kj}|g_{ki}, g_{kj})] - E(Y_{ki})E(Y_{kj}) \\
&= E_{g_{ki}g_{kj}}[E(Y_{ki}|g_{ki})E(Y_{kj}|g_{kj})] - E(Y_{ki})E(Y_{kj}) \\
&= \theta_{ki}\theta_{kj}E(g_{ki}g_{kj}) - E(Y_{ki})E(Y_{kj}) \\
&= \theta_{ki}\theta_{kj}E(g_{ki}g_{kj}) - \theta_{ki}\theta_{kj}E(g_{ki})E(g_{kj}) \\
&= \theta_{ki}\theta_{kj}Cov(g_{ki}, g_{kj}) \\
&= \theta_{ki}\theta_{kj}(E(g_{ki}g_{kj}) - E(g_{ki})E(g_{kj})) \qquad (2.5)
\end{aligned}
$$

Putting (2.4) and (2.5) together we have,

$$
\begin{aligned}
Var(T_k) &= \sum_{i=1}^{m} [\theta_{ki} E(g_{ki}) + \theta_{ki}^2 Var(g_{ki})] \\
&+ 2\sum_{i<j} \theta_{ki}\theta_{kj}(E(g_{ki}g_{kj}) - E(g_{ki})E(g_{kj})). \qquad (2.6)
\end{aligned}
$$

Another relevant quantity for our optimal drafting procedure is the covariance between two lineups. We have,

$$
Cov(T_k, T_l) = E(T_kT_l) - E(T_k)E(T_l)
$$

$$
\begin{aligned}
&= E(\sum_{i=1}^{m} Y_{ki} \sum_{j=1}^{m} Y_{lj}) - (\sum_{i=1}^{m} \theta_{ki} E(g_{ki}))(\sum_{j=1}^{m} \theta_{lj} E(g_{lj})) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} E(Y_{ki} Y_{lj}) - (\sum_{i=1}^{m} \theta_{ki} E(g_{ki}))(\sum_{j=1}^{m} \theta_{lj} E(g_{lj})) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} E_{g_{ki} g_{lj}} E(Y_{ki} Y_{lj} | g_{ki} g_{lj}) - (\sum_{i=1}^{m} \theta_{ki} E(g_{ki}))(\sum_{j=1}^{m} \theta_{lj} E(g_{lj})) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} E_{g_{ki} g_{lj}} (\theta_{ki} g_{ki} \theta_{lj} g_{lj}) - (\sum_{i=1}^{m} \theta_{ki} E(g_{ki}))(\sum_{j=1}^{m} \theta_{lj} E(g_{lj})) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \theta_{ki} \theta_{lj} E(g_{ki} g_{lj}) - (\sum_{i=1}^{m} \theta_{ki} E(g_{ki}))(\sum_{j=1}^{m} \theta_{lj} E(g_{lj})) \qquad (2.7)
\end{aligned}
$$

due to the conditional independence assumption. An important point is that the $E(T_k)$, $Var(T_k)$ and $Cov(T_k, T_l)$ expressions in (2.3), (2.6) and (2.7) involve the terms $E(g_{ki})$, $Var(g_{ki})$ and $E(g_{ki}g_{kj})$, and these terms are found in advance of the draft by simulation (see section 2.3). Therefore, we can calculate $E(T_k)$, $Var(T_k)$ and $Cov(T_k, T_l)$ for every lineup in the draft quickly. One further point is that the covariance in (2.7) assumes that lineups k and l have the same number of players. This expression is easily modified for two lineups with an unequal numbers of players, and this is required for optimal drafting.

2.3 Simulating the Stanley Cup Playoffs

In order to calculate the terms $E(T_k)$, $Var(T_k)$ and $Cov(T_k, T_l)$, we have written an S-Plus program to simulate the Stanley Cup Playoffs and estimate $E(g_{ki})$, $Var(g_{ki})$ and $E(g_{ki}g_{kj})$. We now explain how this is done.

2.3.1 The Probability Matrix P

We estimate the terms $E(g_{ki})$, $Var(g_{ki})$ and $E(g_{ki}g_{kj})$ by simulating the Stanley Cup Playoffs. These rely on estimates for the probability of every series outcome. We require a 16×16 matrix of win probabilites, $\mathbf{P}$. The entry $P(i, j)$ is the probability that team i wins a best of seven series against team j for $i, j = 1, 2, \ldots, 16$. Therefore, $P(j, i) = 1 - P(i, j)$ for $i \neq j$ and since a team cannot play against itself, $P(i, i)$ is left undefined.

We consider three methods for estimating these win probabilites. The first method was proposed for use in the NCAA Men's Basketball March Madness tournament (Breiter and Carlin, 1997); it is based solely on team seedings. In the NCAA tournament, the teams are ranked by a selection committee according to their relative strength. The win probabilites are given by

$$P(i, j) = \frac{rank(j)}{rank(i) + rank(j)}$$

where $rank(i)$ is the seeding of team i with $rank(i) = 1$ denoting the "best" team. Applying this formula to the Stanley Cup Playoffs, we suggest that it can be used for both within and between conference matchups. However, an implicit assumption of this formula is that both conferences are equally strong, so that for example, the probability of two equally ranked teams beating one another is equivalent to $1/2$. In addition, it suggests that a higher ranked team (ie. a team with a lower $rank()$ value) is always stronger than a lower ranked team. This is not a desirable quality because hockey teams are not ranked by a committee; seeding is determined by the regular season point totals. Sometimes a higher ranked team is not stronger than a lower ranked team. For example, team i may have lost every regular season game against an opponent j, but still finish with more points and hence $P(i, j) > 1/2$.

The second method is based on a two step approach that uses the regular season point totals to estimate win probabilites (Monahan and Berger, 1977). First, the probability that team i wins a particular game against team j in Stanley Cup play is estimated by

$$p = \text{P(team } i \text{ wins a game against team } j) = \frac{i\text{'s total points}}{i\text{'s total points}+j\text{'s total points}} \quad (2.8)$$

Assuming the outcome of a game is a Bernoulli random variable with probability p, given in (2.8) then the series win probability is given by

$$P(i,j) = p^4 + \binom{4}{3}p^4(1-p) + \binom{5}{3}p^4(1-p)^2 + \binom{6}{3}p^4(1-p)^3.$$

Notice that this method allows for overall differences in conference strength; teams with the same seeding from the two conferences may have very different regular season point totals. However, this method also suffers from the same inadequacy as the March Madness method. Simply because team i has a greater regular season point total than team j does not necessarily suggest that team i is more likely to beat team j. For example, perhaps team i has suffered a recent rash of injuries.

The final method that we consider is a graph theory approach that uses first round Sportsbook odds and the draftee's subjective hockey knowledge. We believe that this method is superior to the other two approaches because it is designed to make the series win probabilites as realistic as possible. Sportsbooks try to determine public opinion in order to balance bets (Insley, Mok and Swartz, 2004). By balancing the bets the Sportsbooks are able to guarantee a profit regardless of the winning team. So using series win probabilites derived from Sportsbook odds seems like a reasonable idea. Figure 2.1 gives the layout for the first round of the playoffs. Each line in the graph indicates that the probability $P(i,j)$ between the two connected teams is available. That is, since the betting odds for the eight first round series are available, these odds can be transformed to win probabilities. Sportsbook odds are reported in the form $Odds(i,j):1$ where $Odds(i,j)$ is the payout in dollars on a winning one dollar bet on team i and $P(i,j) = 1/(Odds(i,j)+1)$. For example, 3:1 betting odds

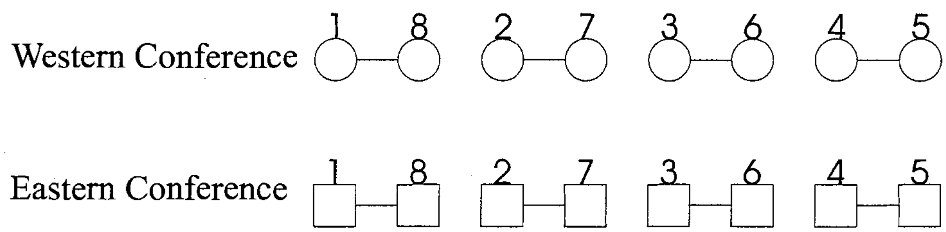

Figure 2.1: First round format

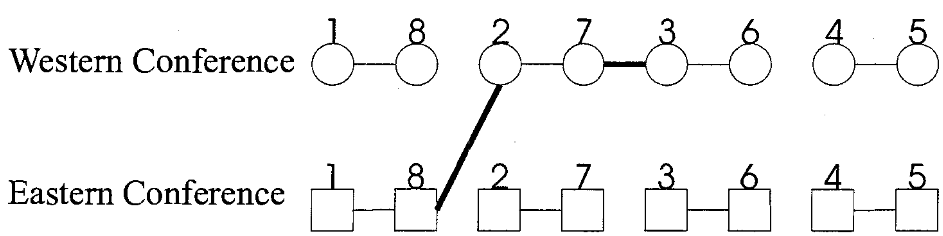

Figure 2.2: Draftee's probabilites in bold

for team i defeating team j corresponds to $P(i, j) = 0.25$.

Now, the only glitch is that we need to complete the probability matrix $\mathbf{P}$ before the playoffs begin and we only have the Sportsbook odds for the first round. To complete the probability matrix, we use the draftee's subjective hockey knowledge. For example, perhaps the draftee is a die hard Canucks fan and can meticulously predict the probabilities of all possible matchups against the Canucks. The bold lines in Figure 2.2 represent two of the draftee's subjective probabilites.

Fortunately, it is not necessary for the draftee to complete the remainder of the probability matrix $\mathbf{P}$. By assuming that the odds are "transitive", we can use the Sportsbook odds and the draftee's subjective odds to determine the odds and corresponding probabilites of other matchups. For example, referring to Figure 2.2, by transitivity we can calculate $P(8E, 7W)$ by following the line from $8E$ to $7W$ whereby

$$Odds(8E, 7W) = Odds(8E, 2W) \bullet Odds(2W, 7W)$$

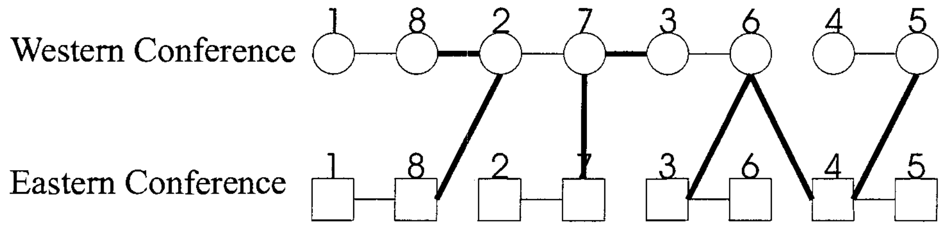

Figure 2.3: Example of a connected graph

The goal is to create a connected graph; a graph is connected if there exists a path between each pair of vertices. In this situation, the draftee must specify a minimum of seven subjective probabilities in order to connect the graph and complete the probability matrix. An example of a connected graph is given in Figure 2.3.

Of course, if one is going to complete the probability matrix **P** via transitivity there should not be different paths that lead to different probability calculations. The draftee must be "transitivity coherent" in his or her subjective probability assignments. For example, one should not have $P(4,5) = 0.5$ and then assign $P(1,4) = 0.4$ and $P(1,5) = 0.6$. Note that transitivity is a strong assumption that is not always applicable in sports. For example, one could imagine particular matchups where team i is favoured over team j, team j is favoured over team k, yet team k is favoured over team i. We believe that the transitivity assumption is fairly sensible in hockey. If one is adamant that transitivity is inappropriate, they should then simply complete the entire probability matrix **P**.

2.3.2 Calculating the Number of Games in a Series

After determining the matrix **P**, for any possible matchup we have the probability that a particular team will win the entire best of seven series. However, we do not know the probability distribution for the number of games that are played in the series. Let p_{ij}

be the probability that team i beats team j in a single game on neutral ice. However, the playoff games are not played on neutral ice. Recall from the Introduction that each series follows a 2-2-1-1-1 home away schedule where the home ice advantage is given to the higher-ranked team. We define home ice advantage as the increase in probability ϵ of team i beating team j in a single game at home compared to neutral ice. We obtain an estimate of ϵ common to the league by considering the results of the regular season and setting

$$\epsilon = \frac{(\text{number of home team wins}) + \frac{1}{2}(\text{number of tied games})}{\text{total number of regular season games}}.$$

From the probability matrix $\mathbf{P}$, we have estimates for $P(i,j)$, the probability that team i wins the series over team j for $i \neq j$. We want to find an estimate for p_{ij}, the probability that team i beats team j in a single game on neutral ice. Since each series is a best of seven games we have

$$\begin{aligned} P(i,j) &= P(i \text{ wins in } 4) + P(i \text{ wins in } 5) \\ &+ P(i \text{ wins in } 6) + P(i \text{ wins in } 7) \end{aligned} \tag{2.9}$$

This is not as straightforward as it appears because we must take into account the 2-2-1-1-1 schedule. We examine each term in the right hand side of (2.9) separately.

First we assume that team i is the higher ranked team. The simplest scenario to calculate is team i sweeping the series by winning the first four games.

$$\begin{aligned} P(i \text{ wins in } 4) &= P(i \text{ wins the first 4 games}) \\ &= P(i \text{ wins 2 games at home and 2 games away}) \\ &= (p_{ij}+\epsilon)^2(p_{ij}-\epsilon)^2. \end{aligned} \tag{2.10}$$

If team i were to win the series in five games there are two possible outcomes to consider; team i loses a home game or team i loses an away game. Of course there are different combinations to consider as well. If team i were to lose a home game this means that they must lose either the first or second game of the series. Similarly, if team i were to lose an away game they must lose either the third or fourth game of the series. Therefore,

$$\begin{aligned} P(i \text{ wins in } 5) &= 2(p_{ij}+\epsilon)(1-(p_{ij}+\epsilon))(p_{ij}-\epsilon)^2(p_{ij}+\epsilon) \\ &+ 2(p_{ij}+\epsilon)^2(p_{ij}-\epsilon)(1-(p_{ij}-\epsilon))(p_{ij}+\epsilon). \end{aligned} \tag{2.11}$$

If team i were to win the series in six games we must consider three possible outcomes; team i loses two home games, team i loses two away games, or team i loses one home game and one away game. Abiding by the 2-2-1-1-1 schedule there are three combinations (games 12, 15, 25) of two home losses, one combination (games 34) of two away losses and six combinations (games 13, 14, 23, 24, 35, 45) of one home loss and one away loss. Therefore,

$$\begin{aligned} P(i \text{ wins in } 6) &= 3(p_{ij}+\epsilon)(1-(p_{ij}+\epsilon))^2(p_{ij}-\epsilon)^2(p_{ij}-\epsilon) \\ &+ (p_{ij}+\epsilon)^3(1-(p_{ij}-\epsilon))^2(p_{ij}-\epsilon) \\ &+ 6(p_{ij}+\epsilon)^2(1-(p_{ij}+\epsilon))(p_{ij}-\epsilon)^2(1-(p_{ij}-\epsilon)). \end{aligned} \tag{2.12}$$

If team i were to win the series in seven games we must consider four possible outcomes; team i loses three home games (games 125), team i loses three away games (games 346), team i loses two home games and one away game (nine combinations games 123, 124, 126, 135, 145, 156, 235, 245, 256) and team i loses one home game and two away games (nine combinations consisting of games 134, 136, 146, 234, 236, 246, 345, 356, 456). Therefore,

$$\begin{aligned} P(i \text{ wins in } 7) &= (1-(p_{ij}+\epsilon))^3(p_{ij}-\epsilon)^3(p_{ij}+\epsilon) \\ &+ (p_{ij}+\epsilon)^3(1-(p_{ij}-\epsilon))^3(p_{ij}+\epsilon) \\ &+ 9(1-(p_{ij}+\epsilon))^2(1-(p_{ij}-\epsilon))(p_{ij}+\epsilon)(p_{ij}-\epsilon)^2(p_{ij}+\epsilon) \\ &+ 9(1-(p_{ij}+\epsilon))(1-(p_{ij}-\epsilon))^2(p_{ij}+\epsilon)^3(p_{ij}-\epsilon). \end{aligned} \tag{2.13}$$

The equation given in (2.9) is expanded by substituting equations (2.10), (2.11), (2.12) and (2.13) and we note that p_{ij} is the only unknown in the expanded equation. In order to obtain p_{ij} we use the Newton-Raphson algorithm and set the initial value ${p_{ij}}^{(0)} = 0.5$. We then substitute p_{ij} into the expressions (2.10)-(2.13) to obtain $P(i$ wins in 4), $P(i$ wins in 5), $P(i$ wins in 6) and $P(i$ wins in 7). Now using equations (2.10)-(2.13) and $p_{ji} = 1 - p_{ij}$, we can similarly obtain $P(j$ wins in 4), $P(j$ wins in 5), $P(j$ wins in 6) and $P(j$ wins in 7). We then have a discrete probability distribution with 8 cells describing the outcome of the series between teams i and j.

To simulate the total number of games played by team i in the playoffs, we simulate each round of the playoffs using the 8-cell discrete probability distributions and keep a running total of the games played for team i. This is done for each of the 16 teams. After completing many (thousands of) simulations, S-Plus has built-in functions that are used to estimate the terms $E(g_{ki})$, $Var(g_{ki})$ and $Cov(g_{ki}, g_{kj})$.

2.4 Optimality Criterion

Recall from section 2.2.4 that we defined T_k as the total number of points accumulated by draftee $k = 1, \ldots, K$. In addition, we defined g_{ki} to be the number of games played in the playoffs by the i^{th} player selected in the draft by draftee k. In section 2.2.3 we argued that $Y_{ki}|g_{ki} \sim Poisson(\theta_{ki} g_{ki})$. Furthermore, it can be argued (Summers, Swartz and Lockhart 2005) that T_k can be approximated by a normal distribution, with parameters $E(T_k)$, $Var(T_k)$ and $Cov(T_k, T_l)$ known in advance of the draft and given by (2.3), (2.4) and (2.7) respectively. Ideally, and without loss of generality, we would like to draft a lineup T_1 so as to maximize

$$P(T_1 = \max_{j=1,\ldots,K} T_j) = P(T_1 > T_2, \ldots, T_1 > T_K)$$

$$= P(T_1 - T_2 > 0, \ldots, T_1 - T_K > 0) \qquad (2.14)$$

Now assuming that $(T_1, \ldots, T_K)$ is multivariate normal $\mathbf{N}_K(\mu, \Sigma)$, the probability (2.14) is equal to $P(Q > 0)$ where $Q \sim \mathbf{N}_{K-1}(\mu_Q, \Sigma_Q)$ with parameters $\mu_Q = \mathbf{X}\mu$ and $\Sigma_Q = \mathbf{X}\Sigma\mathbf{X}'$ where

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & \\ 1 & 0 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

This is known as an orthant probability and it is notoriously difficult to approximate in moderate/high dimensions in reasonable computing times (Evans and Swartz, 1988).

We want an optimality criterion that is logical and effective in real time. We therefore attempt to maximize

$$P^* = \frac{1}{K-1}[P(T_1 > T_2) + \cdots + P(T_1 > T_K)]. \qquad (2.15)$$

We interpret P^* as the average probability that lineup 1 beats one of its competitors. Note that the terms in P^* are easily obtained via

$$\begin{aligned} P(T_1 > T_j) &= P(T_1 - T_j > 0) \\ &= P(Z > -\mu_j/\sigma_j) \\ &= \Phi(\mu_j/\sigma_j) \end{aligned}$$

where $T_1 - T_j \sim N(\mu_j, {\sigma_j}^2)$ with $\mu_j = (1 \quad 0 \cdots -1 \quad 0 \cdots 0)$ and ${\sigma_j}^2 = (1 \quad 0 \cdots -1 \quad 0 \cdots 0)\,\Sigma\,(1 \quad 0 \cdots -1 \quad 0 \cdots 0)'$.

2.5 Putting Theory into Action

The optimality criterion established in section 2.4 was the last step in our theoretical development. The next step is to test our player selection method by entering a NHL playoff pool. In chapter 3 we simulate a playoff pool of m rounds with K draftees. In order to implement our selection method we need to create a realistic probability matrix $\mathbf{P}$ to determine estimates for $E(g_{ki})$, $Var(g_{ki})$ and $Cov(g_{ki}, g_{kj})$ via simulation as described in section 2.3. In addition, regular season data is used to determine estimates of θ for eligible players.

Before the draft begins the order of player selection is randomized. After the first round of drafting is complete the order of drafting is then reversed for the second round, and the process continues for m rounds. We let T_1 correspond to the total number of points accumulated by our lineup chosen by our optimality criterion even if we are not the first draftee. The question that we want to answer is, "which player should we choose next?" By keeping track of all the draftees' lineups we calculate $P(T_1 > T_j)$ for $j \neq 1$. Following the optimality criterion (2.15) we choose the player from those remaining in the draft who maximizes P^*. Once the draft is complete we tally the points obtained by each player in the playoffs and determine the winning lineup(s).

Chapter 3

Simulation

3.1 Simulation Prerequisites

Due to the cancellation of the 2004-2005 NHL season, we were unable to conduct a real office playoff pool to test our player selection method. To test the performance of our methods, we simulate a playoff pool of m rounds with K draftees. Using the data from the 2003-2004 regular season we consider the 2004 NHL playoffs. The team summaries at the end of the regular season are given in Table 3.1 where W denotes wins, L denotes losses, T denotes ties, OTL denotes overtime losses, Pts denotes points and W% denotes win percentage. To calculate W% we take the number of points and divide it by twice the number of games played.

Table 3.1: 2003-2004 Regular season team summary

Team	**Conference**	**Seed**	**W**	**L**	**T**	**OTL**	**Pts**	**W%**
DET	Western	1	48	21	11	2	109	0.665
TAM	Eastern	1	46	22	8	6	106	0.646

–continued on next page

– continued from previous page

Team	Conference	Seed	W	L	T	OTL	Pts	W%
BOS	Eastern	2	41	19	15	7	104	0.634
SJ	Western	2	43	21	12	6	104	0.634
TOR	Eastern	4	45	24	10	3	103	0.628
OTT	Eastern	5	43	23	10	6	102	0.622
PHI	Eastern	3	40	21	15	6	101	0.616
VAN	Western	3	43	24	10	5	101	0.616
NJ	Eastern	6	43	25	12	2	100	0.610
COL	Western	4	40	22	13	7	100	0.610
DAL	Western	5	41	26	13	2	97	0.591
CAL	Western	6	42	30	7	3	94	0.573
MON	Eastern	7	41	30	7	4	93	0.567
NYI	Eastern	8	38	29	11	4	91	0.555
STL	Western	7	39	30	11	2	91	0.555
NAS	Western	8	38	29	11	4	91	0.555
EDM	Western	9	36	29	12	5	89	0.543
BUF	Eastern	9	37	34	7	4	85	0.518
MIN	Western	10	30	29	20	3	83	0.506
LOS	Western	11	28	29	16	9	81	0.494
ATL	Eastern	10	33	37	8	4	78	0.476
CAR	Eastern	11	28	34	14	6	76	0.463
ANA	Western	12	29	35	10	8	76	0.463
FLA	Eastern	12	28	35	15	4	75	0.457
NYR	Eastern	13	27	40	7	8	69	0.421
PHO	Western	13	22	36	18	6	68	0.415

–continued on next page

– continued from previous page

Team	Conference	Seed	W	L	T	OTL	Pts	W%
CLB	Western	14	25	45	8	4	62	0.378
WAS	Eastern	14	23	46	10	3	59	0.360
CHI	Western	15	20	43	11	8	59	0.360
PIT	Eastern	15	23	47	8	4	58	0.354

Since the 2004 playoffs have already been completed, the Sportsbook odds for round one are no longer available. We employed the assistance of a "hockey guru" (David Beaudoin) to construct realistic first round odds and the subjective odds of seven hypothetical matchups given in Table 3.2.

Using the odds given in Table 3.2 we are able to complete the probability matrix using the transitivity assumption described in Section 2.3.1. In Table 3.3, the $(i, j)^{th}$ entry is the probability that team i wins the series against team j given that team i has the home ice advantage, and in parentheses is the probability that team i wins a single game against team j on neutral ice. Recall the estimate of home ice advantage given in section 2.3.2; for the 2003-2004 regular season data we found $\epsilon = 0.05$.

3.2 The Draft

For our playoff pool simulation we choose $K = 10$, $m = 10$ and restrict our attention to the 10 players per team who had the highest θ's as estimated via the method described in section 2.2.2. A sample of the 160 eligible players, their teams, regular season points, games played and corresponding θ's are given in Table 3.4.

Table 3.2: Subset of Sportsbook and subjective odds

Team i **vs Team** j	**Odds:1**	$P(i,j)$
First Rounds Odds		
Detroit vs Nashville	0.190	0.84
San Jose vs St. Louis	0.724	0.58
Vancouver vs Calgary	0.786	0.56
Colorado vs Dallas	0.818	0.55
Tampa Bay vs NY Islanders	0.449	0.69
Boston vs Montreal	0.667	0.60
Philadelphia vs New Jersey	0.923	0.52
Toronto vs Ottawa	1.083	0.48
Subjective Odds		
Detroit vs Vancouver	0.639	0.61
Vancouver vs Colorado	0.923	0.52
Tampa Bay vs Boston	0.923	0.52
Boston vs Vancouver	0.887	0.53
Calgary vs Ottawa	1.326	0.43
Detroit vs San Jose	0.786	0.56
Calgary vs New Jersey	1.222	0.45

To simulate a playoff pool we need to create "virtual" draftees. Draftees follow specific rules to determine their lineups.

- Draftee 1 $\sim$ chooses players using the optimality criterion (2.15).
- Draftee 2 $\sim$ chooses players with the largest θ values. If there is a tie then the draftee chooses the player with the most regular season points.
- Draftee 3 $\sim$ chooses players with the largest expected number of points during the playoffs. The expected points are the product of the player's θ with his expected number of games played obtained from the Stanley Cup playoff simulation.
- Draftee 4 $\sim$ is an advocate of numerology. The draftee believes that the numbers 8 and 9 are lucky, and chooses players with the most regular season point totals

Table 3.3: Probability matrix

	DET	SJ	VAN	COL	DAL	CAL	STL	NSH	TAM	BOS	PHI	TOR	OTT	NJ	MON	NYI
DET		0.56	0.61	0.63	0.67	0.67	0.64	0.84	0.56	0.58	0.6	0.62	0.6	0.62	0.68	0.74
		(0.54)	(0.57)	(0.58)	(0.60)	(0.60)	(0.58)	(0.70)	(0.54)	(0.55)	(0.56)	(0.57)	(0.56)	(0.57)	(0.60)	(0.64)
SJ	0.44		0.55	0.57	0.62	0.61	0.58	0.81	0.50	0.52	0.54	0.56	0.54	0.56	0.62	0.69
	(0.48)		(0.54)	(0.55)	(0.57)	(0.57)	(0.55)	(0.68)	(0.51)	(0.52)	(0.53)	(0.54)	(0.53)	(0.54)	(0.57)	(0.61)
VAN	0.39	0.45		0.52	0.57	0.56	0.53	0.77	0.45	0.47	0.49	0.51	0.49	0.51	0.57	0.65
	(0.46)	(0.49)		(0.52)	(0.55)	(0.54)	(0.53)	(0.66)	(0.49)	(0.50)	(0.51)	(0.52)	(0.51)	(0.52)	(0.55)	(0.59)
COL	0.37	0.43	0.48		0.55	0.54	0.51	0.76	0.43	0.45	0.47	0.49	0.47	0.49	0.55	0.63
	(0.45)	(0.48)	(0.50)		(0.54)	(0.53)	(0.52)	(0.65)	(0.48)	(0.49)	(0.50)	(0.51)	(0.50)	(0.51)	(0.54)	(0.58)
DAL	0.33	0.38	0.43	0.45		0.49	0.46	0.72	0.38	0.40	0.42	0.44	0.42	0.44	0.50	0.58
	(0.43)	(0.45)	(0.48)	(0.49)		(0.51)	(0.49)	(0.63)	(0.45)	(0.46)	(0.47)	(0.48)	(0.47)	(0.48)	(0.51)	(0.55)
CAL	0.33	0.39	0.44	0.46	0.51		0.47	0.73	0.39	0.41	0.43	0.45	0.43	0.45	0.51	0.59
	(0.43)	(0.46)	(0.48)	(0.49)	(0.52)		(0.50)	(0.63)	(0.46)	(0.47)	(0.48)	(0.49)	(0.48)	(0.49)	(0.52)	(0.56)
STL	0.36	0.42	0.47	0.49	0.54	0.53		0.75	0.42	0.44	0.46	0.48	0.46	0.48	0.54	0.62
	(0.44)	(0.47)	(0.50)	(0.51)	(0.53)	(0.53)		(0.64)	(0.47)	(0.48)	(0.49)	(0.50)	(0.49)	(0.50)	(0.53)	(0.57)
NSH	0.16	0.19	0.23	0.24	0.28	0.27	0.25		0.20	0.21	0.22	0.24	0.22	0.24	0.28	0.35
	(0.33)	(0.35)	(0.37)	(0.38)	(0.40)	(0.39)	(0.38)		(0.35)	(0.36)	(0.37)	(0.38)	(0.37)	(0.38)	(0.4)	(0.44)
TAM	0.44	0.50	0.55	0.57	0.62	0.61	0.58	0.80		0.52	0.54	0.56	0.54	0.56	0.62	0.69
	(0.48)	(0.51)	(0.54)	(0.55)	(0.57)	(0.57)	(0.55)	(0.68)		(0.52)	(0.53)	(0.54)	(0.53)	(0.54)	(0.57)	(0.61)
BOS	0.42	0.48	0.53	0.55	0.6	0.59	0.56	0.79	0.48		0.52	0.54	0.52	0.54	0.60	0.67
	(0.47)	(0.50)	(0.53)	(0.54)	(0.56)	(0.56)	(0.54)	(0.67)	(0.50)		(0.52)	(0.53)	(0.52)	(0.53)	(0.56)	(0.60)
PHI	0.40	0.46	0.51	0.53	0.58	0.57	0.54	0.78	0.46	0.48		0.52	0.50	0.52	0.58	0.66
	(0.46)	(0.49)	(0.52)	(0.53)	(0.55)	(0.55)	(0.53)	(0.66)	(0.49)	(0.50)		(0.52)	(0.51)	(0.52)	(0.55)	(0.59)
TOR	0.38	0.44	0.49	0.51	0.56	0.55	0.52	0.76	0.44	0.46	0.48		0.48	0.50	0.56	0.64
	(0.45)	(0.48)	(0.51)	(0.52)	(0.54)	(0.54)	(0.52)	(0.65)	(0.48)	(0.49)	(0.50)		(0.50)	(0.51)	(0.54)	(0.58)
OTT	0.40	0.46	0.51	0.53	0.58	0.57	0.54	0.78	0.46	0.48	0.50	0.52		0.52	0.58	0.66
	(0.46)	(0.49)	(0.52)	(0.53)	(0.55)	(0.55)	(0.53)	(0.66)	(0.49)	(0.50)	(0.51)	(0.52)		(0.52)	(0.55)	(0.59)
NJ	0.38	0.44	0.49	0.51	0.56	0.55	0.52	0.76	0.44	0.46	0.48	0.50	0.48		0.56	0.64
	(0.45)	(0.48)	(0.51)	(0.52)	(0.54)	(0.54)	(0.52)	(0.65)	(0.48)	(0.49)	(0.50)	(0.51)	(0.50)		(0.54)	(0.58)
MON	0.32	0.38	0.43	0.45	0.50	0.49	0.46	0.72	0.38	0.40	0.42	0.44	0.42	0.44		0.58
	(0.42)	(0.45)	(0.48)	(0.49)	(0.51)	(0.51)	(0.49)	(0.63)	(0.45)	(0.46)	(0.47)	(0.48)	(0.47)	(0.48)		(0.55)
NYI	0.26	0.31	0.35	0.37	0.42	0.41	0.38	0.65	0.31	0.33	0.34	0.36	0.34	0.36	0.42	
	(0.39)	(0.42)	(0.44)	(0.45)	(0.47)	(0.47)	(0.45)	(0.59)	(0.42)	(0.43)	(0.43)	(0.44)	(0.43)	(0.44)	(0.47)	

Table 3.4: Sample of player information

Player	**Team**	**Pts**	**GP**	θ
Martin St. Louis	TAM	94	82	1.1463
Joe Sakic	COL	87	81	1.0741
Markus Naslund	VAN	84	78	1.0769
Marian Hossa	OTT	82	81	1.0123
Patrick Elias	NJ	81	82	0.9878
Daniel Alfredsson	OTT	80	77	1.0390
Cory Stillman	TAM	80	81	0.9877
Alex Tanguay	COL	79	69	1.1449
Robert Lang	DET	79	69	1.1449
Brad Richards	TAM	79	82	0.9634
Milan Hejduk	COL	75	82	0.9146
Mark Recchi	PHI	75	82	0.9146
Mats Sundin	TOR	75	81	0.9259
Joe Thorton	BOS	73	77	0.9481
Jarome Iginla	CAL	73	81	0.9012

that are divisible by 8 or 9. If there is a tie the draftee chooses the player with the largest θ value.

- Draftee 5 $\sim$ roots for the underdog by choosing players with the most points, alternating between the lowest seeded teams in the Eastern and Western Conferences.

- Draftee 6 $\sim$ alternates between the two top seeded teams in the Eastern and Western Conferences, choosing the player with the most regular season points.

- Draftee 7 $\sim$ chooses players with the most regular season points. If there is a tie the draftee chooses the player that belongs to the highest seeded team. If there is still a tie the draftee chooses the player with the largest θ value.

- Draftee 8 $\sim$ is a Vancouver Canucks "Superfan". The draftee always picks Canuck players in order of regular season points. If there is a tie the draftee chooses the Canuck with the largest θ value.

- Draftee 9 $\sim$ chooses players with the most regular season points whose first names begin with the letter S. If there is a tie between players, the draftee chooses the player with the larger θ value.

- Draftee 10 $\sim$ chooses players with the highest θ values from the top four seeded teams in the first four rounds of the draft. For the remaining six rounds the draftee chooses players with the highest θ values regardless of the team seeding.

We make special note of Draftee 3. Draftee 3 can be considered a "ringer" because the expected points are based on the probability matrix and our playoff simulation also uses the probability matrix. In a sense, Draftee 3's underlying knowledge of players' playoff production is perfect in the same way as Draftee 1. In a real playoff pool we would not be so willing to share our results with other draftees. Thus, a draftee who chooses players based on expected points would have to conduct an independent simulation with results that are unlikely to match ours. This raises the question why do we include a draftee with such a competitive edge? We use this draftee to check our player selection method. To be more specific, when we use the optimality criterion to choose the next player in our lineup do we inadvertantly choose the player with the most expected points?

We also want to investigate whether or not the draftee's position in the player selection order is influential on the final results. In order to test this we performed a second draft modifying the order by moving Draftee 1 from the desirable position of choosing first to choosing last. The lineups for the first draft are given in Table 3.5 where the order of the players within each lineup corresponds to their order of drafting. In order to save space the lineups from the second draft have been omitted.

Table 3.5: Lineups: Draftee 1 ~ first pick

Player	**Team**	θ	**Expected Points**
	Draftee 1		
Robert Lang	DET	1.145	17.211
Martin Havlat	OTT	1.000	10.965
Vincent Lecavalier	TAM	0.815	10.586
Ray Whitney	DET	0.642	9.648
Alexei Zhamnov	PHI	0.837	9.221
Sergei Samsonov	BOS	0.690	8.215
Brian Leetch	TOR	0.708	7.393
Paul Kariya	COL	0.706	7.685
Marco Sturm	SJ	0.641	7.631
Jonathan Cheechoo	SJ	0.580	6.912
	Draftee 2		
Peter Forsberg	COL	1.410	15.353
Alex Tanguay	COL	1.145	12.464
Keith Tkachuk	STL	0.947	8.965
Jarome Iginla	CAL	0.901	8.659
Doug Weight	STL	0.867	8.207
Joe Nieuwendyk	TOR	0.781	8.154
Jason Arnott	DAL	0.781	7.591
Alexei Yashin	NYI	0.723	5.892
Bryan McCabe	TOR	0.707	7.375
Rob Blake	COL	0.622	6.767

–continued on next page

– continued from previous page

Player	Team	θ	Expected Points
Draftee 3			
Martin St. Louis	TAM	1.146	14.894
Brett Hull	DET	0.840	12.620
Henrik Zetterberg	DET	0.705	10.596
Milan Hejduk	COL	0.915	9.957
Kris Draper	DET	0.597	8.974
Jeremy Roenick	PHI	0.758	8.349
John Leclair	PHI	0.733	8.077
Jason Spezza	OTT	0.705	7.731
Radek Bonk	OTT	0.667	7.310
Alexander Korolyuk	SJ	0.587	6.996
Draftee 4			
Patrick Elias	NJD	0.988	10.278
Daniel Alfredsson	OTT	1.039	11.392
Michael Ryder	MON	0.778	7.172
Chris Pronger	STL	0.675	6.392
Scott Niedermayer	NJD	0.667	6.937
Owen Nolan	TOR	0.738	7.707
Gary Roberts	TOR	0.667	6.958
Brian Rolston	BOS	0.585	6.972
Alex Kovalev	MON	0.577	5.320
Pierre Turgeon	DAL	0.526	5.117
Draftee 5			
Steve Sullivan	NAS	0.913	5.875

–continued on next page

– continued from previous page

Player	Team	θ	Expected Points
Trent Hunter	NYI	0.662	5.395
Marek Zidlicky	NAS	0.646	4.162
Oleg Kvasha	NYI	0.630	5.129
Martin Erat	NAS	0.645	4.151
Mariusz Czerkawski	NYI	0.605	4.927
David Legwand	NAS	0.573	3.691
Jason Blake	NYI	0.627	5.104
Kimmo Timonen	NAS	0.571	3.679
Adrian Aucoin	NYI	0.543	4.425
Draftee 6			
Cory Stillman	TAM	0.988	12.832
Pavel Datsyuk	DET	0.907	13.629
Fredrik Modin	TAM	0.695	9.031
Brendan Shanahan	DET	0.646	9.716
Ruslan Fedotenko	TAM	0.506	6.581
Mathieu Schneider	DET	0.590	8.865
Dan Boyle	TAM	0.500	6.496
Nicklas Lidstrom	DET	0.469	7.052
Dave Anderychuk	TAM	0.476	6.179
Pavel Kubina	TAM	0.432	5.614
Draftee 7			
Joe Sakic	COL	1.074	11.693
Marian Hossa	OTT	1.012	11.100
Mark Recchi	PHI	0.915	10.073

–continued on next page

– continued from previous page

Player	Team	θ	Expected Points
Mats Sundin	TOR	0.926	9.664
Bill Guerin	DAL	0.841	8.181
Glen Murray	BOS	0.741	8.823
Michael Handzus	PHI	0.707	7.790
Nils Ekman	SJ	0.671	7.989
Tony Amonte	PHI	0.663	7.297
Richard Zednik	MON	0.617	5.692
Draftee 8			
Markus Naslund	VAN	1.077	12.150
Todd Bertuzzi	VAN	0.870	9.811
Brendan Morrison	VAN	0.732	8.255
Daniel Sedin	VAN	0.659	7.430
Martin Rucinsky	VAN	0.549	6.191
Henrik Sedin	VAN	0.553	6.235
Brent Sopel	VAN	0.525	5.923
Geoff Sanderson	VAN	0.450	5.077
Trevor Linden	VAN	0.439	4.953
Mattias Ohlund	VAN	0.415	4.678
Draftee 9			
Scott Gomez	NJD	0.875	9.104
Scott Walker	NAS	0.893	5.752
Sergei Gonchar	BOS	0.817	9.730
Saku Koivu	MON	0.809	7.458
Simon Gagne	PHI	0.563	6.195

–continued on next page

– continued from previous page

Player	Team	θ	Expected Points
Sergei Zubov	DAL	0.545	5.303
Shean Donovan	CAL	0.512	4.921
Steve Konowalchuk	COL	0.488	5.311
Sheldon Souray	MON	0.556	5.123
Scott Hartnell	NAS	0.559	3.601
Draftee 10			
Brad Richards	TAM	0.963	12.517
Joe Thornton	BOS	0.948	11.292
Patrick Marleau	SJ	0.713	8.487
Steve Yzerman	DET	0.680	10.222
Pavol Demitra	STL	0.853	8.077
Mike Ribeiro	MON	0.802	7.400
Valeri Bure	DAL	0.765	7.435
Craig Conroy	CAL	0.746	7.168
Steven Reinprecht	CAL	0.659	6.332
Peter Bondra	OTT	0.636	6.977

We want to check that our player selection method is not identical to that of Draftee 3 otherwise there is little point in using our complicated methodology based on the optimality criterion. We therefore compare whom Draftee 3 would have selected had he been drafting in the position of Draftee 1. As we can see in Table 3.6 the player selection strategies employed by Draftees 1 and 3 are not identical. The other issue we wish to investigate is whether we need to consider all available players or only the players with the largest θ values from each team. In both of the drafts, the players

Table 3.6: Draftee 3 versus Draftee 1

	Draftee 1 ∼ first pick	
Round	**Draftee 3**	**Draftee 1**
1	Robert Lang	Robert Lang
2	Martin Havlat	Martin Havlat
3	Henrik Zetterberg	Vincent Lecavalier
4	Ray Whitney	Ray Whitney
5	Alexei Zhamnov	Alexei Zhamnov
6	Sergei Samsonov	Sergei Samsonov
7	John Leclair	Brian Leetch
8	Paul Kariya	Paul Kariya
9	Marco Sturm	Marco Sturm
10	Jonathan Cheechoo	Jonathan Cheechoo
	Draftee 1 ∼ last pick	
Round	**Draftee 3**	**Draftee 1**
1	Pavel Datsyuk	Pavel Datsyuk
2	Brett Hull	Alex Tanguay
3	Steve Yzerman	Mark Recchi
4	Steve Yzerman	Steve Yzerman
5	Kris Draper	Glen Murrray
6	Kris Draper	Kris Draper
7	Nils Ekman	Nils Ekman
8	Jason Spezza	Jason Spezza
9	Alexander Korolyuk	Peter Bondra
10	Alexander Korolyuk	Alexander Korolyuk

Table 3.7: Draft results: Draftee 1 $\sim$ first pick

	Draftee									
Finish	1	2	3	4	5	6	7	8	9	10
1st	1113	2093	1501	244	117	1635	740	1925	11	621
2nd	2061	1072	2125	708	188	1225	1244	435	42	900
3rd	2452	873	2097	684	172	780	1438	277	99	1128
4th	2023	941	1423	735	164	839	1533	312	220	1810
5th	1111	1093	936	1003	215	1062	1723	357	321	2179
6th	625	1445	706	1590	259	945	1452	624	611	1743
7th	372	1162	561	2243	288	820	1092	721	1737	1004
8th	188	823	442	2048	555	854	615	600	3403	472
9th	54	450	180	619	2396	979	141	1878	3173	130
10th	1	48	29	126	5646	861	22	2871	383	13
$E(LD)$	6.54	5.60	6.40	3.99	1.12	4.92	5.49	3.41	2.17	5.35

selected in each round had the largest available θ value on their teams. This seems to imply that we need not consider all available players; this would help to reduce the computational time necessary for implementing the optimality criterion, but proving this property has turned out to be somewhat problematic. More discussion on this problem is given in the companion paper by Summers, Swartz and Lockhart (2005).

3.3 Simulation Draft Results

We ran simulations for $N = 10000$ iterations in order to see how the optimality criterion performs against other player selection strategies. In Tables 3.7 and 3.9 we see the frequency of finishing in first to last place for the two drafts. In addition, from these tables we can extract $E(LD)$, the expected number of lineups that lineup j defeats for $j = 1, \ldots, 10$. Tables 3.8 and 3.10 show the cumulative probabilities associated with the drafts.

In draft one, Draftees 1 and 3 performed very well with the largest $E(LD)$ values,

Table 3.8: Cumulative probabilities: Draftee 1 $\sim$ first pick

	Draftee									
Finish	1	2	3	4	5	6	7	8	9	10
1st	0.111	0.209	0.150	0.024	0.012	0.164	0.074	0.192	0.001	0.062
2nd	0.317	0.316	0.363	0.095	0.030	0.286	0.198	0.236	0.005	0.152
3rd	0.563	0.404	0.572	0.164	0.048	0.364	0.342	0.264	0.015	0.265
4th	0.765	0.498	0.715	0.237	0.064	0.448	0.496	0.295	0.037	0.446
5th	0.876	0.607	0.808	0.337	0.086	0.554	0.668	0.331	0.069	0.664
6th	0.938	0.752	0.879	0.496	0.112	0.649	0.813	0.393	0.130	0.838
7th	0.976	0.868	0.935	0.721	0.140	0.731	0.922	0.465	0.304	0.939
8th	0.994	0.950	0.979	0.926	0.196	0.816	0.984	0.525	0.644	0.986
9th	1.000	0.995	0.997	0.987	0.435	0.914	0.998	0.713	0.962	0.999
10th	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

6.54 and 6.40 respectively. However, Draftee 1 was fifth in term of finishing in first place. Despite the relatively small percentage of finishing first using the optimality criterion (11.1%), Draftee 1 is "in the money" (finishing first, second or third) 56.3% of the time. This is a higher percentage than any of the other draftees excluding Draftee 3. Draftee 3 is very similar finishing in the top three 57.2% of the time. Note that the Canucks Superfan (Draftee 8) followed the "eggs in one basket" philosophy; when the Canucks did well/poorly, his lineup did well/poorly.

In the second draft, Draftee 1 was the last to pick. Unfortunately, Draftee 1 does not have the largest $E(LD)$ as in the first draft. However, the same two draftees (1 and 3) have the largest $E(LD)$ values, 6.16 and 6.70 respectively. Draftee 1's number of first place finishes is slightly larger than in the first draft (1291 versus 1113), but Draftees 2, 3, 6 and 8 still have a greater number of first place finishes. The increase in Draftee 1's number of first place finishes may be attributed to the fact that only 10 players have been eliminated before Draftee 1 makes his second choice whereas in the first draft 19 players have been eliminated before Draftee 1's second choice.

Table 3.9: Draft results: Draftee 1 $\sim$ last pick

	Draftee									
Finish	2	3	4	5	6	7	8	9	10	1
1st	1321	1826	237	125	2168	715	1829	15	473	1291
2nd	1199	2742	556	138	1006	1148	447	44	931	1789
3rd	1122	1809	712	187	734	1533	411	79	1457	1956
4th	1408	1161	876	182	735	1488	322	196	1965	1667
5th	1566	820	1171	225	732	1466	422	370	2063	1165
6th	1636	653	1545	302	801	1357	577	741	1599	789
7th	1152	472	2095	330	852	1068	641	1838	930	622
8th	413	353	2086	516	998	914	621	3209	444	446
9th	174	147	599	2420	1029	278	1864	3162	128	199
10th	9	17	123	5575	945	33	2866	346	10	76
$E(LD)$	5.66	6.70	3.99	1.14	4.96	5.33	3.43	2.21	5.42	6.16

Despite the increase in the number of first place finishes we can see that the probability of finishing "in the money" is only 50.4%. Also, note that we are never able to overcome Draftee 3. This seems to indicate that the selection order is quite important in determining the end result.

Table 3.10: Cumulative probabilities: Draftee 1 ~ last pick

	Draftee									
Finish	2	3	4	5	6	7	8	9	10	1
1st	0.132	0.183	0.024	0.012	0.217	0.072	0.183	0.002	0.047	0.129
2nd	0.252	0.457	0.079	0.026	0.317	0.186	0.228	0.006	0.140	0.308
3rd	0.364	0.638	0.150	0.045	0.391	0.340	0.269	0.014	0.286	0.504
4th	0.505	0.754	0.238	0.063	0.464	0.488	0.301	0.033	0.483	0.670
5th	0.662	0.836	0.355	0.086	0.538	0.635	0.343	0.070	0.689	0.787
6th	0.825	0.901	0.510	0.116	0.618	0.771	0.401	0.144	0.849	0.866
7th	0.940	0.948	0.719	0.149	0.703	0.878	0.465	0.328	0.942	0.928
8th	0.982	0.984	0.928	0.200	0.803	0.969	0.527	0.649	0.986	0.972
9th	0.999	0.998	0.988	0.442	0.906	0.997	0.713	0.965	0.999	0.992
10th	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Chapter 4

Conclusion

The goal of this project was to create a real time optimal drafting strategy for hockey playoff pools. First, we explored the theoretical background including the distributions of points scored and games. Next, we described the Stanley Cup Playoff simulation employing a transitivity assumption to complete the probability matrix. The optimality criterion

$$P^* = \frac{1}{K-1}[P(T_j > T_1) + \cdots + P(T_j > T_K)]$$

where $T_1, \ldots, T_K$ correspond to the current lineups in round m was developed in Section 2.4. P^* is interpretted as the average probability that lineup j accumulates more points than one of the other lineups. Using the approximation

$$T_j - T_k \sim N(\mu_{jk}, {\sigma_{jk}}^2)$$

where $\mu_{jk} = E(T_j) - E(T_k)$, and ${\sigma_{jk}}^2 = Var(T_j) + Var(T_k) - 2Cov(T_j, T_k)$ we approximate P^* by

$$\frac{1}{K-1}[\Phi(\frac{\mu_{j1}}{\sigma_{j1}}) + \cdots + \Phi(\frac{\mu_{jK}}{\sigma_{jK}})]$$

where Φ is the cumulative distribution function of the standard normal distribution. We choose the available hockey player that maximizes P^*.

The optimality criterion was evaluated using a program written in S-Plus. We were able to select players for our lineup in real time, but as the lineups became larger and larger the program began to lag, taking between five to seven minutes to finish. Although this is not excessive in terms of our simulated drafts, taking five minutes to choose the player in a real draft may be frowned upon by the other draftees. One way to help alleviate this problem would be to try a different programming language (e.g. C++) that handles loops more efficiently. Another suggestion is to reduce the number of players under consideration. As suggested in Section 3.2 we may only want to consider the available player with the top θ value from each team. A retrospective analysis of our player drafts revealed that this strategy would have worked in our drafts. However, we spent a fair amount of time trying to prove this property in general, only to come up empty handed.

The results from the simulated drafts given in Section 3.3 suggest that the optimality criterion is an effective drafting strategy. In the "first pick" draft, we did not finish in first place as often as we had hoped. If the pool only has a single prize then the best strategies involve loading up on players from one team or alternatively choosing players from the top seeded teams in the Eastern and Western conferences. The success of picking players based on the expected number of points in the playoffs depends upon how closely your model matches reality. However, if the pool has multiple prizes then the optimality criterion does very well, quite often finishing in the money (first, second or third place).

In the "last pick" draft, the number of first place finishes is somewhat larger.

However, the cumulative probability of finishing in first, second or third was not as impressive as in the first draft. We still finish in the top three quite regularly, but we are unable to overcome our nemesis Draftee 3, who picks second in the draft. Based on the observed results, it appears that a draftee's position influences the final outcome of the draft.

One thing that is usually known before the draft begins is the prizes; one should consider the prize distribution before settling on a particular selection strategy. If there is only one prize you may want to use a more aggressive strategy or if there are multiple prizes you may want to use a strategy that increases your probability of winning a prize.

The final test for our optimality criterion will be an authentic office playoff pool. With the National Hockey League Players Association (NHLPA) and team owners having reached an agreement to end the lockout we eagerly anticipate the 2006 NHL Stanley Cup playoffs.

Bibliography

[1] Berry, S.M. (2000), "My triple crown. In the column, A Statistician Reads the Sports Pages." *Chance*, 13 (3), 56-61.

[2] Berry, S.M., Reese, C.S. and Larkey, P.L. (1999), "Bridging Different Eras in Sports." *Journal of the American Statistical Association*, 94, 661-686.

[3] Breiter, D.J. and Carlin, B.P. (1997), "How to Play the Office Pools if You Must." *Chance*, 10 (1), 5-11.

[4] Monahan, J.P. and Berger P.D. (1977), "Playoff Structures in the National Hockey League." *Optimal Strategies in Sports*, S. Ladany and R. Machol (editors), 123-128.

[5] Insley, R., Mok, L. and Swartz, T.B. (2004), Issues Related to Sports Gambling." *The Australian and New Zealand Journal of Statistics*, 46 (2), 219-232.

[6] Evans, M. and Swartz, T.B. (1988), "Monte Carlo computation of multivariate normal probabilities." *Journal of Statistical Computation and Simulation*, 30, 117-128.

[7] Summers, A.E., Swartz, T.B. and Lockhart R.A. (2005) "Optimal Drafting in Hockey Pools." *manuscript*