Support Based Measures Applied to Ice Hockey Scoring Summaries: Extended Version

Bradley P. Kram, James A. Hall and
Howard J. Hamilton

Technical Report CS-2000-05
August, 2000

Department of Computer Science
University of Regina
Regina, Saskatchewan, CANADA
S4S 0A2

ISSN 0828-3494
ISBN 0-7731-0409-7

Support Based Measures Applied to Ice Hockey Scoring Summaries: Extended Version

Bradley P. Kram, James A. Hall, and Howard J. Hamilton
University of Regina
Regina, Saskatchewan, Canada, S4S 0A2
{kram, hallj, hamilton}@cs.uregina.ca

Subject Areas: AI Applications, AI in Data Mining, AI in Multimedia

Abstract

We present the Hockey Line Extraction (HLE) Algorithm, which examines ice hockey scoring summaries in an attempt to determine a team's lines. The players on a hockey team are divided into units called "lines" that appear together on the ice. While many statistics are available on the Internet for hockey players and teams in the National Hockey League (NHL), accurately identifying current lines for a team is difficult. The HLE algorithm uses single link clustering, support based measures and positional information to identify lines of players. The HLE Algorithm and appropriate visualization techniques have been implemented in the Hockey Lines software that enables users to view relationships between players on a team.

1. Introduction

Increasingly, data mining techniques are being applied to sports data [5, 8]. For example, IBM developed the Advanced Scout data mining software for analyzing basketball games by collecting statistics on shots attempted, shots blocked, etc. and also detecting patterns in these statistics, such as which plays are most effective with which players and under what circumstances [8]. Sports data is plentiful on the World Wide Web and frequently analyzed for summaries and predictions. In particular, the Web provides ice hockey fans with a tremendous amount of information and statistics about National Hockey League (NHL) players and teams. The demand for statistics about NHL hockey players has prompted web sites to provide detailed information, such as a player's ice time, body checks, and blocked shots for every player in the

NHL. However, some information is difficult to obtain. The players on a team are divided into units called *lines* that appear together on the ice. Currently, a hockey fan has no easy way of learning which players are on lines together. Also, as teams make changes and players are injured or traded, lines often change. For these reasons, obtaining current, accurate information about the lines for the teams in the NHL is difficult.

Data mining and visualization techniques can be applied to available NHL hockey statistics in an attempt to determine the lines for a team. Jagadish and Ng analyze NHL scoring statistics in an attempt to group records in a database that share a set of common characteristics [5]. The most accessible NHL statistics are the scoring summaries for NHL games. These summaries outline the players who scored goals, and the players who assisted on goals. Since two players who play on the same line are likely to assist on each other's goals, these summaries can be analyzed to determine which players appear on lines together. The support-based measures that are commonly used to analyze market basket data are well suited to solving this problem.

We introduce the Hockey Line Extraction Algorithm (HLE) to cluster players who commonly contribute to scoring the same goal according to NHL scoring summaries. For our purposes, a goal is a transaction, and the players involved in the goal are the items in the transaction. Due to the dynamic nature of hockey lines, we present the result of our algorithm to the user in a fashion that enables the user to view the clusters, as well as the relationships among the clusters, over varying time periods.

The following section introduces the Hockey Line Extraction Algorithm. Section 3 explains visualization techniques used to present the relationships among players to the user.

Section 4 discusses the results of these techniques using detailed examples as illustrations. Section 5 outlines the conclusions that can be drawn from this project and potential future work.

2. Approach

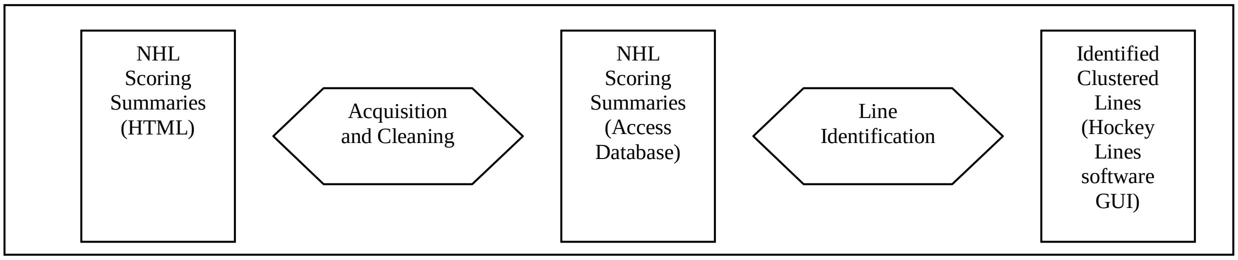

NHL Scoring Summaries (HTML)
Acquisition and Cleaning
NHL Scoring Summaries (Access Database)
Line Identification
Identified Clustered Lines (Hockey Lines software GUI)

Figure 1: Approach

The Hockey Lines software acquires NHL scoring summaries from the Internet, stores these summaries in a database, and analyzes these summaries to identify lines of players. The lines are presented to the user along with graphical depiction of the relationships between players on a team. Figure 1 depicts this process.

2.1. Data Acquisition and Cleaning

The Hockey Lines software first acquires data concerning teams, players, and goals. Since we assume that the only available source of the information is the web, the data must be parsed from HTML into a usable format for insertion into an Access database with tables for players and goals. Since no one web site had all the information required in a convenient format, we used data from both http://www.nhl.com and http://cbs.sportsline.com to fully populate the database.

A major concern was formatting inconsistencies between sites and even within one site. NHL.com, where we obtained the majority of the player, game, and goal information, had numerous changes in formatting during the 1999-2000 season, without an associated adjustment for games earlier in the season. This meant the software must convert the data from HTML to our format needed to understand not only all of the formats throughout the season, but needed to synchronize these differences. For example, the format of a player name changed midway through the season, but we needed to know players' statistics across the entire year, not broken up at some artificial date when the format changed. For this reason, an automated tool such as YAT [4], was not appropriate, and a dedicated data cleaning program was written to ensure data met all requirements.

Data downloaded from the web needed to be cleaned. In addition to formatting difficulties and spelling inconsistencies, the data source does not adequately maintain time-varying data such as the history of teams a player has played for. If a temporal database [6] had been used by the websites, these problems could have been avoided. In its absence, the dedicated cleaning program to had to resolve these problems as much as possible. Support Based Measures

Support based measures are used on large databases of transactions in an attempt to determine which items often occur together [2]. These groups of items, called *itemsets*, can be examined through a measure called *support*, defined as the ratio of the number of transactions containing a particular itemset over the total number of transactions [3]. An *association rule*, A→B between two itemsets A and B, states that when A occurs, B also occurs. The *confidence* of an association rule based on a database of transactions shows the ratio of the support of $A \cup B$ to the support of A [2]. This confidence indicates the likelihood of one item occurring given that another item occurs.

The support based measures are adapted to a database containing information about goals scored in the National Hockey League. Each goal is considered a transaction involving one to three players (the goal scorer and up to two assisting players). Over many transactions (goals), we determine the support of the every two-player itemset for a particular team. These measures will provide insight into which players are playing together. Both the Hockey Line Extraction Algorithm in Section 2.2 and the visualization techniques of Section 3 utilize these support values.

2.2. Hockey Line Extraction Algorithm.

The Hockey Line Extraction (HLE) Algorithm clusters players into lines of a given size. The HLE Algorithm is a modified version of the Single Link Clustering Algorithm [1, 7]. The Single Link Clustering Algorithm groups items in a database into clusters of items that commonly occur together. The goal is to form clusters of items such that any two-itemset formed from the items in a cluster is likely to have a relatively high support value. Support values are calculated for every possible two-itemset. The algorithm begins with an edgeless graph with each node representing an item. The *distance* between two nodes is defined as the inverse of the support of the two-itemset containing the items at the two nodes. Edges are added in order of increasing distance. Nodes connected by an edge form a *connected component*. The *mass* of a connected component is defined as the sum of the supports for the individual items (nodes) in the connected component. The *critical mass* is a threshold value for the mass of a connected component. The critical mass is defined as a percentage of the total support for all items. Once the mass of a connected component reaches the critical mass, the connected component is removed from the graph and forms a cluster. The algorithm continues until all edges have been added or all items have been removed from the graph. Any remaining connected components are considered clusters even if their mass does not exceed the critical mass.

In the HLE Algorithm, as shown in Figure 2, each node represents a player. Edges are added in order of decreasing support to form connected components. However, instead of using a critical mass threshold as a stopping condition as is done in the Single Link Clustering Algorithm, we examine the requested line size, denoted *L*. Once the cardinality of a connected component reaches *L*, the connected component is removed and forms a line in the final result.

```
Input: $L$ as a number indicating the desired line size.

Output: Result as a list of player lists.  Each list of players in the outer list represents a line.

ConnectedComponents is a list of player lists.  Each list of player lists represents a working connected component
that has not yet formed a line.

For a given team and date range, calculate the distance between each combination of players on the team.  Sort these
values in order of increasing distance and store them in LinkInfo.

For each link $I$ in LinkInfo connecting player $A$ and player $B$:
{
    If $A$ and $B$ are both in Result, continue.
    If neither $A$ nor $B$ is in ConnectedComponents, create a new connected component containing $A$ and $B$.  If
        this new connected component has cardinality $L$, move the connected component to Result.
    If one of $A$ and $B$ is already in ConnectedComponents, add the other player to the same connected
        component.  If the given connected component has cardinality $L$, move the connected component
        to Result.
    If both $A$ and $B$ are already in ConnectedComponents, find the components that contain $A$ and $B$ known as
        $C_A$ and $C_B$ respectively.  Let $S$ be the set of all player combinations of cardinality $L$ consisting of
        the players in $C_A$ and $C_{B.}$  Let $P_C$ be the number of duplicate player positions in a combination, $C \in$
        $S$. Let $M_S$ be the set of player combinations that minimize $P_{C.}$
    {
        If Cardinality( $M_S$ ) = 1, remove the players in the single combination in $M_S$ into Result.
        Else
        {
            Let $S_C$ be the average support for each two-itemset in a player combination $C \subset M_S$.
            Remove the players in the player combination that minimized $S_C$ from the graph.  Move
                these players into Result.
        }
    }
}
```

Figure 2: The Hockey Line Extraction Algorithm

Consider the following table of goals for a given time range with *L* = 3. The three players in a line may include any number of Left Wingers (LW), Right Wingers (RW) and Centers (C).

Player	Position
P1	C
P2	LW
P3	RW
P4	LW
P5	RW
P6	RW
P7	LW
P8	C
P9	LW
P10	C

Goal ID	Goal Scorer	Assist 1	Assist 2
1	P1	P2	P3
2	P1	P2	P3
3	P1	P2	P3
4	P1	P2	P3
5	P1	P2	P3
6	P1	P2	P3
7	P1	P2	
8	P4	P5	P7
9	P4	P5	P7
10	P4	P5	P7

Goal ID	Goal Scorer	Assist 1	Assist 2
11	P4	P5	P7
12	P4	P5	
13	P4	P5	
14	P6	P7	P8
15	P6	P7	P8
16	P6	P7	P8
17	P6	P7	
18	P6	P7	
19	P9	P10	P8
20	P9	P10	

Since P1 and P2 appear together in 7 out of 20 goals, the support, denoted s, of { P1, P2 } is 0.35. The two-itemsets in order of decreasing support are as follows:

Itemset	Support
{P1,P2}	0.35
{P1,P3}	0.30
{P2,P3}	0.30
{P4,P5}	0.30

Itemset	Support
{P6,P7}	0.25
{P4,P7}	0.20
{P5,P7}	0.20
{P6,P8}	0.15

Itemset	Support
{P7,P8}	0.15
{P9,P10}	0.10
{P8,P9}	0.05
{P8,P10}	0.05

The addition of the first two edges to the graph creates a connected component consisting of three players. We remove these players from the graph and the players {P1, P2, P3} form a line in the final assessment. Figure 3 depicts the addition of the first two edges.

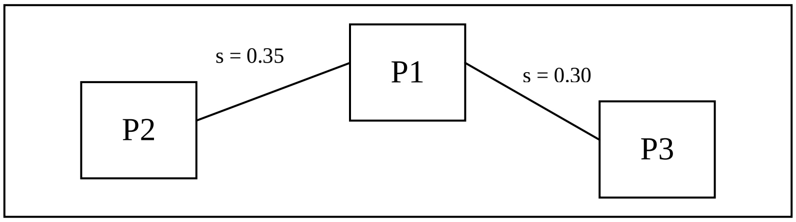


Figure 3: The Formation of a Line

The addition of an edge may form a connected component that is larger than the requested line size. In this case, we must decide how to break the connected component into two components. As shown in the diagram in Figure 4, the edge between Player PW and Player PY is added to form a connected component consisting of four players. Since we are looking for a line of three players, we must break the connected component into smaller sub-components. Let the cardinality of Connected Component 1 be denoted *C1* and the cardinality of Connected Component 2 be denoted *C2*. To split the large connected component, we examine the set *S* of $C_{L}^{(C1 + C2)}$ combinations of players from Connected Components *C1* and *C2*.

Rule 1: We first examine the positions of the players within each combination of players in *S* to determine which player to select for a line. The player position information is a highly accurate indication of the position a player is playing on the ice, although it may be incorrect due to injuries or other reasons. Let P_C be the number of duplicate player positions in a combination, $C \in S$. Let M_S be the set of player combinations that minimize $P_{C.}$

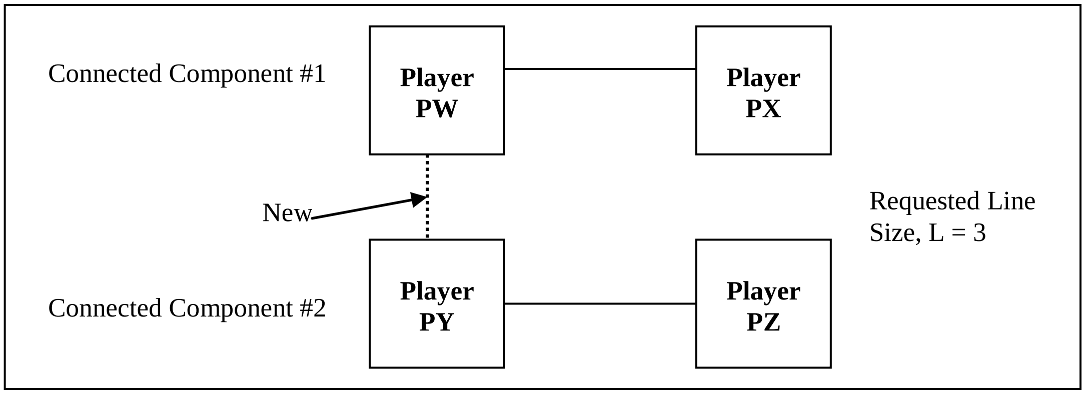


Figure 4: A New Edge Forming a Large Connected Component

Rule 1: If Cardinality(M_S) = 1, the players in the single player combination in M_S are removed from the graph and form a line.

This rule attempts to minimize the number of players in a line that have the same position. The unfortunate side effect of this decision is that it may separate two players whose two-itemset has a high support value.

Rule 2: If Cardinality(M_S) > 1, we examine the average support for each two-itemset consisting of the players in each combination in M_S. Let S_C be the average support for each two-itemset in a player combination $C \in M_S$.

Rule 2: The players in the player combination with the maximum S_C are removed from the graph and form a line.

This rule attempts to maximize the average support of the two-itemsets in a line.

Returning to our example, we find that the addition of the edge between P4 and P7 forms a connected component with four players (Figure 5). We examine the four three-player combinations: {P4, P5, P6}, {P4, P6, P7}, {P4, P5, P7}, {P4, P5, P6}. The number of duplicate player positions for the combinations, $P_{\{P4, P5, P6\}}$, $P_{\{P4, P6, P7\}}$, $P_{\{P4, P5, P7\}}$, and $P_{\{P4, P5, P6\}}$, is 1, so rule 1 does not apply. We examine the values of $S_{\{P4, P5, P6\}}$, $S_{\{P4, P6, P7\}}$, $S_{\{P4, P5, P7\}}$, and $S_{\{P4, P5, P6\}}$, which are as follows: $S_{\{P4, P5, P6\}} = 0.10$; $S_{\{P4, P6, P7\}} = 0.15$; $S_{\{P4, P5, P7\}} = 0.17$; $S_{\{P4, P5, P6\}} = 0.10$. Since the number of duplicate players is a tie, and $S_{\{P4, P5, P7\}}$ is highest, by rule 2 players P4, P5 and P7 form a line and are removed from the graph.

When an edge added between players P8 and P9, a connected component is formed consisting of four players (Figure 6). We examine the four three-player combinations: {P6, P8, P9}, {P8, P9, P10}, {P6, P9, P10}, and {P6, P8, P10}. We find $P_{\{P6, P8, P9\}} = 0$; $P_{\{P8, P9, P10\}} = 1$; $P_{\{P6, P9, P10\}} = 0$; $P_{\{P6, P8, P10\}} = 1$. We examine $S_{\{P6, P8, P9\}}$ and $S_{\{P6, P9, P10\}}$ to find that $S_{\{P6, P8, P9\}} = 0.07$ and $S_{\{P6, P9, P10\}} = 0.03$. We move players P6, P8 and P9 into a line in the final result.

The lines determined by the algorithm are {P1, P2, P3}, {P7, P4, P5} and {P6, P8, P9}.

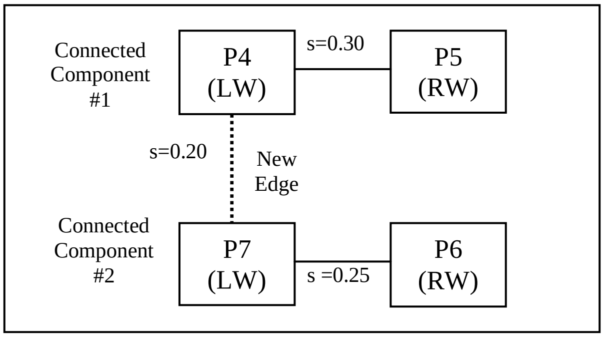


Figure 5: Adding Edge between P4 and P7

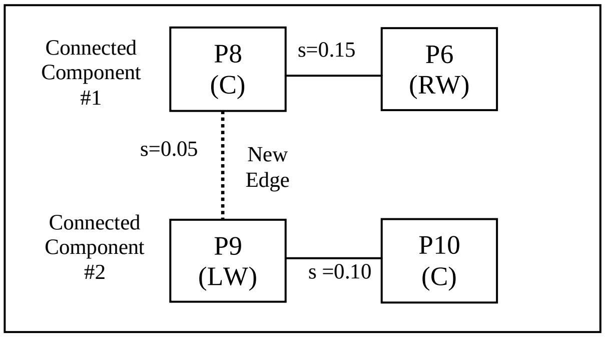


Figure 6: Adding Edge Between P8 and P9

3. Visualization

The software designed for this project enables the user to select a team, a date range, and one or more goal types for analysis. The software displays pictures of the players on the team and allows the user to view the relationships between the players according to support or association confidence. The following sections outline the visualization techniques used in the software.

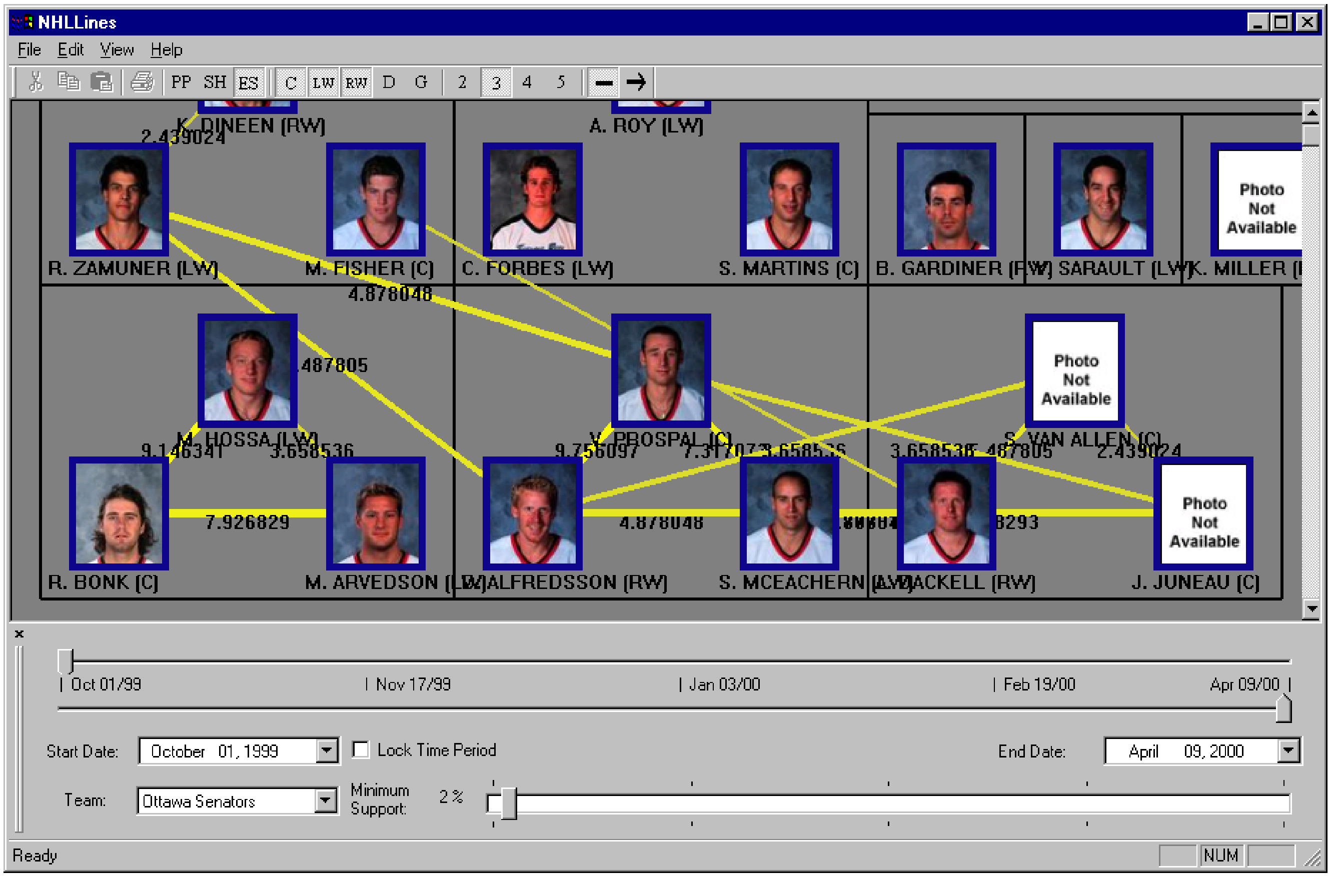
3.1. Automatic Placement

The Hockey Lines software attempts to display players on the screen in a manner such that players who commonly appear in the same goal appear close together on the screen. Using the Hockey Line Extraction Algorithm, the screen is segmented with a bordered space for each line. The space for each line is then filled with the pictures of the players on that line. Only players involved in at least one goal for the given date range are shown.

The use of a bordered space for each line highlights inter-line relationships. A player in one line may have a strong relationship with a player in a different line. Such a connection may indicate that the lines changed significantly during the time period being visualized and no sharp division into lines is possible. Such a connection may also indicate the Hockey Line Extraction Algorithm incorrectly determined the lines, for example, by separating two players based on position even though the players had a strong relationship. These cases are easily noticed because of the thick lines crossing the borders.

3.2. Supports

When the Hockey Lines software is in “View Supports” mode as shown in Figure 7, the lines drawn between players represent the support of the two-itemset containing the two players. The user may optionally view the values associated with each link and filter low support values by adjusting the Minimum Support Slider. The relationships over time may be viewed through the use of the date slider controls. The “Lock Time Period” feature enables the user to slide a fixed size time window over the season and view how the associations between players change over time.



Start Date: October 01, 1999

□ Lock Time Period

End Date: April 09, 2000

Team: Ottawa Senators

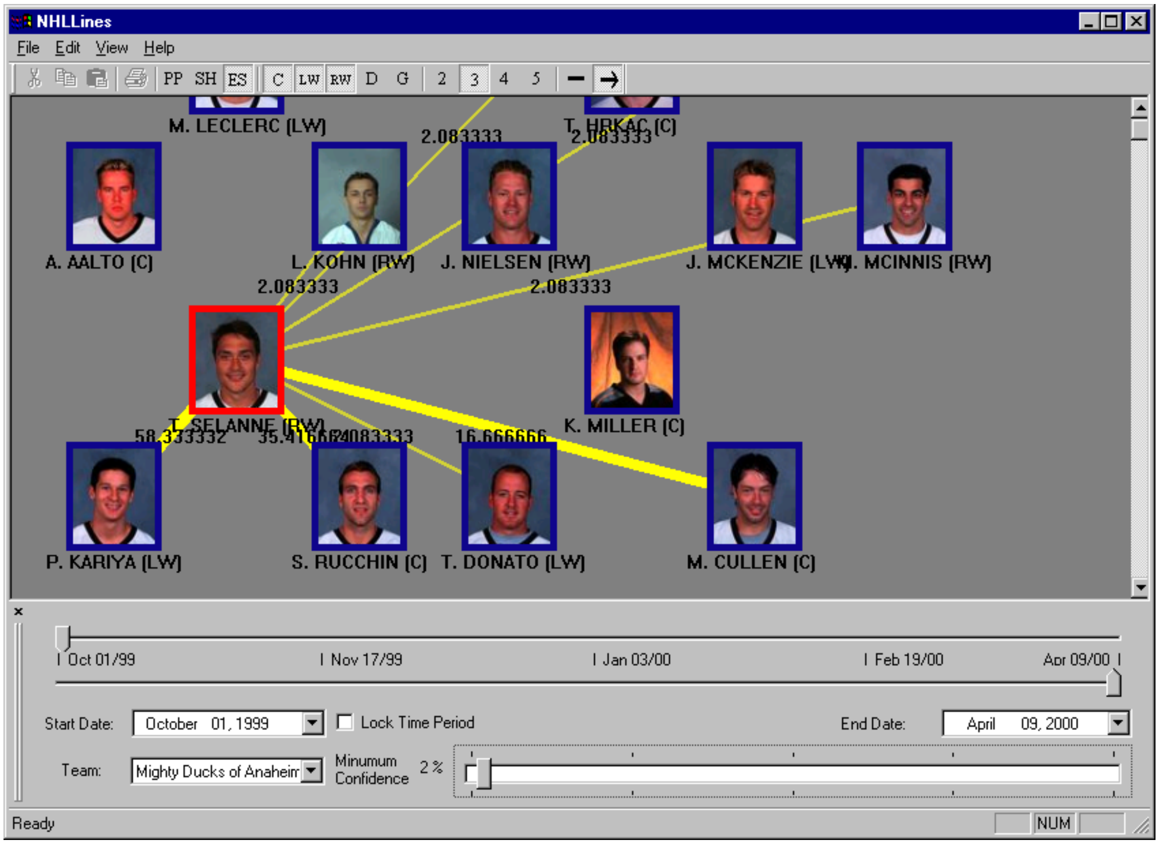
Minimum Support: 2 %

Ready

Figure 7: Visualization of Support Values

3.3. Association Rule Confidence

When the software is in “View Association Rule Confidence Mode” as shown in Figure 8, the user may view the confidence of the association rule between pairs of players. If one player is involved in a goal, how likely is it that another player is involved in the same goal? To view the association confidences, the user selects a player. The selected player is marked with a different colored border. The line connecting the selected player (denoted *S*) to another player *P* indicates the confidence of the association rule *S*→*P*. These confidence values can be filtered using the Association Confidence Slider in a fashion similar to the Minimum Support Slider described in Section 3.2.



Start Date: October 01, 1999

□ Lock Time Period

End Date: April 09, 2000

Team: Mighty Ducks of Anaheim

Minumum Confidence 2 %

Ready

Figure 8: Visualization of Association Confidence Values

3.4. Customization

The Hockey Lines software enables the user to customize the appearance of the links between players. The software displays links in different colors depending on the value associated with the link. In addition, the width of lines is varied depending on the value associated with the link. The user may customize these settings by changing the line width or color of a link for a particular value. The value associated with the link refers to the support when the software is in Support Mode and the confidence of the association rule when it is in Association Confidence Mode. This feature makes it easier for the user to identify relationships among players on the screen.

4. Results

Testing was performed under a Windows NT platform on a PC with an Intel Celeron 500MHz CPU and 128MB of RAM. All goals for a season for the current team are stored in memory, allowing for smooth interaction with the software as the user slides the date range sliders. Each time the date range is changed, a full re-calculation of all player supports and lines is required. However, since an NHL team only scores approximately 300 goals in an NHL season, these calculations are fast enough to be done repeatedly while the user is dragging a scroll bar.

The accuracy of Hockey Line Extraction Algorithm is difficult to measure since detailed records are not available; this was, after all, one of the motivating factors for the project. Furthermore, well-known lines are usually high scoring lines. Published accounts of games emphasize well-known lines that are also usually high scoring ones. We evaluated the Hockey Lines software based on its ability to identify well-known lines. The following examples illustrate both accurate and inaccurate identification of actual lines

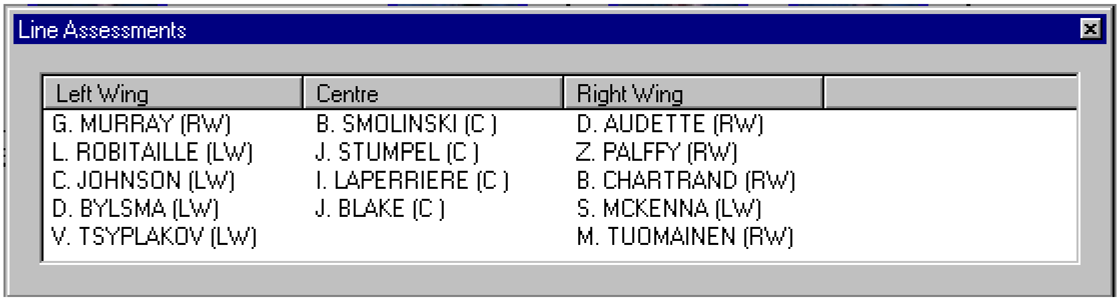
While we include test cases depicting various lines, the focus of our testing was to identify forward lines of three players. To achieve this, the majority of the testing limited the goals to even strength goals and limits the players to Left Wingers, Right Wingers, and Centers.

Testing was performed on a variety of goal types. A goal may be scored in one of three situations. An Even Strength (ES) goal is a goal scored when neither team is being penalized. A Short Handed (SH) goal is a goal scored by a team that is being penalized and has one fewer player on the ice. A Power Play (PP) goal is a goal scored by a team whose opponent is being penalized.

4.1. Test Case 1: Line Identified

Team	**Dates**	**Goal Types**	**Positions**
LA Kings	Oct 1, 1999 to Oct 31, 1999	ES	LW, C, RW

The line of Ziggy Palffy, Josef Stumpel, and Luc Robitaille was one of the most effective lines for the month of October, 1999. The team's second line consisted of Glen Murray, Brain Smolinski and Donald Audette. The Hockey Lines software clearly identifies these two lines as depicted in Figure 9.



Line Assessments

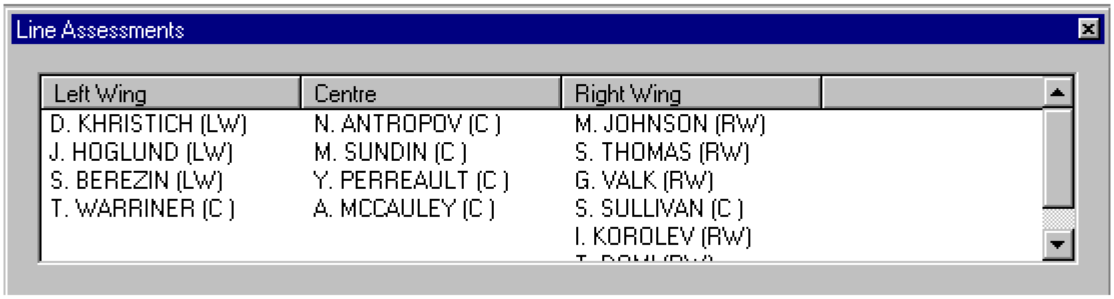
Left Wing	Centre	Right Wing
G. MURRAY (RW)	B. SMOLINSKI (C)	D. AUDETTE (RW)
L. ROBITAILLE (LW)	J. STUMPEL (C)	Z. PALFFY (RW)
C. JOHNSON (LW)	I. LAPERRIERE (C)	B. CHARTRAND (RW)
D. BYLSMA (LW)	J. BLAKE (C)	S. MCKENNA (LW)
V. TSYPLAKOV (LW)		M. TUOMAINEN (RW)

Figure 9: Accurate Line Idnetification

4.2. Test Case 2: Injured Player Not Identified

Team	**Dates**	**Goal Types**	**Positions**
Toronto Maple Leafs	Oct 1, 1999 to Nov 11, 1999	ES	LW, C, RW

The Hockey Lines software does not paint an accurate picture of the lines for the period of October 1, 1999 to November 11, 1999 for the Toronto Maple Leafs. From October 1 to October 9, Toronto's first line consisted of Jonus Hoglund, Mats Sundin, and Steve Thomas, and they scored many goals. However, on October 9, Sundin was hurt, and Yanic Perrault took his place on the line.



Left Wing	Centre	Right Wing
D. KHRISTICH (LW)	N. ANTROPOV (C)	M. JOHNSON (RW)
J. HOGLUND (LW)	M. SUNDIN (C)	S. THOMAS (RW)
S. BEREZIN (LW)	Y. PERREAULT (C)	G. VALK (RW)
T. WARRINER (C)	A. MCCAULEY (C)	S. SULLIVAN (C)
		I. KOROLEV (RW)
		T. DOMI (RW)

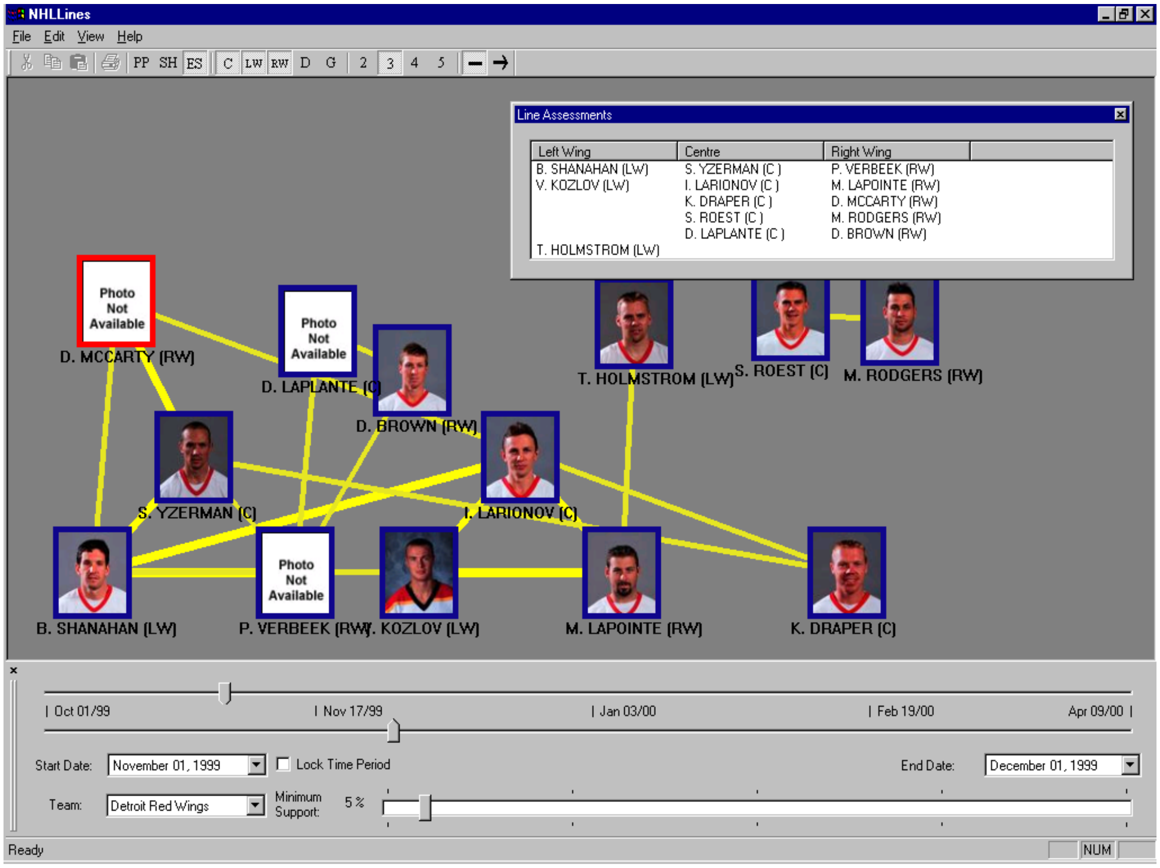
Figure 10: Injured Player Causing Inaccurate Line Identification

As shown in Figure 10, the Hockey Lines software identifies Hoglund, Sundin, and Thomas as playing on a line for October 1 to November 11, even though Sundin was hurt for most of this time period. Hoglund and Thomas scored very few points while Sundin was hurt. Since the line scored several goals during the period that they were playing together, and then the Hoglund and Thomas scored few goals from October 10 to November 11, the support values are highest for the Hoglund, Sundin and Thomas line.

4.3. Test Case 3: Line Change Identified

Team	Dates	Goal Types	Positions
Detroit Red Wings	Nov 1, 1999 to Dec 1, 1999	ES	LW, C, RW

In mid-November, the Detroit Red Wings moved Verbeek onto a line with Yzerman and Shanahan. Verbeek replaced McCarty on the line. The Hockey Lines software determines that the team's first line consists of Yzerman, Shanahan and Verbeek. The visualization techniques indicate that there were other relationships during this time period. As Figure 11 indicates, there is a relationship between Yzerman and McCarty. A user can use the date slide bars to further analyze this relationship over time. If the range November 1 to November 15 is selected, the software determines that Yzerman, Shanahan, and McCarty are on a line. When the ranges November 15 to December 1, the software determines that Yzerman, Shanahan and Verbeek are on a line. This demonstrates how the visualization techniques can identify potential problems with the line assessment and enable the user to investigate further.



Line Assessments

Left Wing	Centre	Right Wing	
B. SHANAHAN (LW)	S. YZERMAN (C)	P. VERBEEK (RW)	
V. KOZLOV (LW)	I. LARIONOV (C)	M. LAPOINTE (RW)	
	K. DRAPER (C)	D. MCCARTY (RW)	
	S. ROEST (C)	M. RODGERS (RW)	
	D. LAPLANTE (C)	D. BROWN (RW)	
T. HOLMSTROM (LW)			

Start Date: November 01, 1999

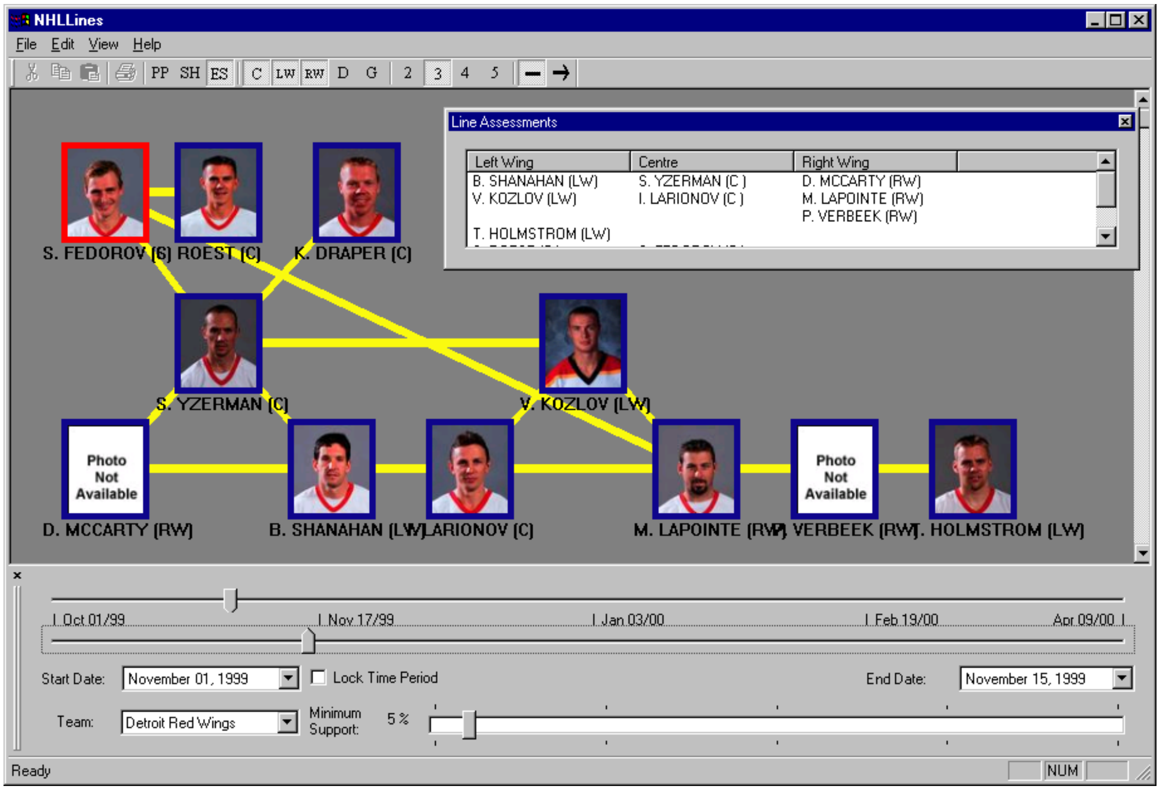
□ Lock Time Period

End Date: December 01, 1999

Team: Detroit Red Wings

Minimum Support: 5 %

Figure 11: Line Change in November, 1999


NHLLines
File Edit View Help
PP SH ES C LW RW D G 2 3 4 5
S. FEDOROV (C) ROEST (C) K. DRAPER (C)
S. YZERMAN (C) V. KOZLOV (LW)
Photo Not Available
D. MCCARTY (RW) B. SHANAHAN (LW) LARIONOV (C) M. LAPOINTE (RW) VERBEEK (RW) T. HOLMSTROM (LW)
Oct 01/99 Nov 17/99 Jan 03/00 Feb 19/00 Apr 09/00

Line Assessments

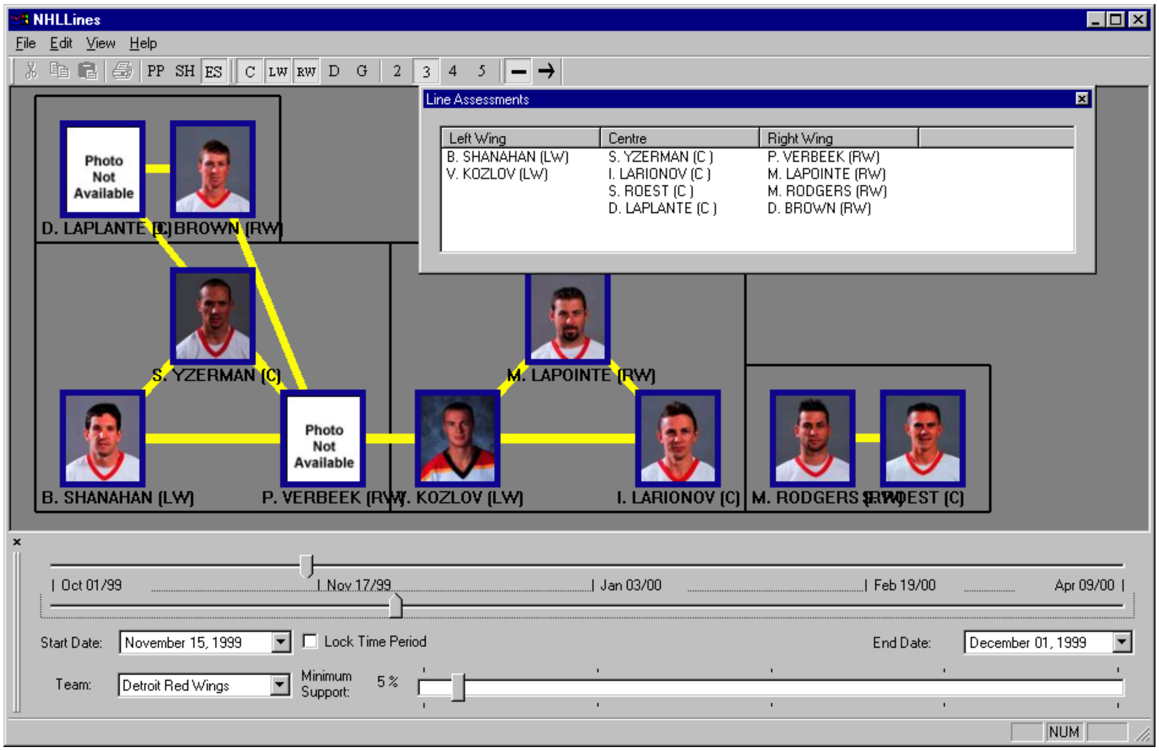
Left Wing	Centre	Right Wing
B. SHANAHAN (LW)	S. YZERMAN (C)	D. MCCARTY (RW)
V. KOZLOV (LW)	I. LARIONOV (C)	M. LAPOINTE (RW)
		P. VERBEEK (RW)
T. HOLMSTROM (LW)		

Start Date: November 01, 1999 ☐ Lock Time Period End Date: November 15, 1999

Team: Detroit Red Wings Minimum Support: 5 %

Ready NUM

Figure 12: First Half of November, 1999


NHLLines
File Edit View Help
PP SH ES C LW RW D G 2 3 4 5
Photo Not Available
D. LAPLANTE (C) BROWN (RW)
S. YZERMAN (C) M. LAPOINTE (RW)
B. SHANAHAN (LW) P. VERBEEK (RW) V. KOZLOV (LW) I. LARIONOV (C) M. RODGERS (RW) ROEST (C)
Oct 01/99 Nov 17/99 Jan 03/00 Feb 19/00 Apr 09/00

Line Assessments

Left Wing	Centre	Right Wing
B. SHANAHAN (LW)	S. YZERMAN (C)	P. VERBEEK (RW)
V. KOZLOV (LW)	I. LARIONOV (C)	M. LAPOINTE (RW)
	S. ROEST (C)	M. RODGERS (RW)
	D. LAPLANTE (C)	D. BROWN (RW)

Start Date: November 15, 1999 ☐ Lock Time Period End Date: December 01, 1999

Team: Detroit Red Wings Minimum Support: 5 %

NUM

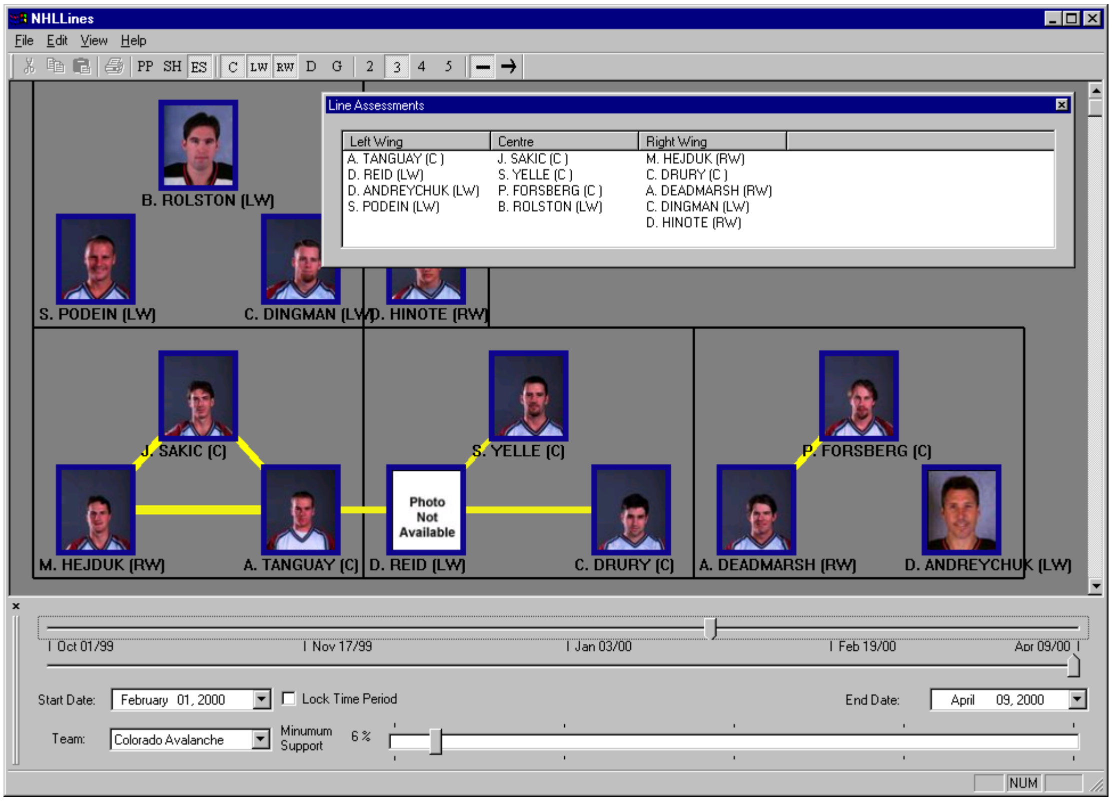
Figure 13: Second Half of November, 1999

4.4. Test Case 4: Line Identified With Power Play

Team	Dates	Goal Types	Positions
Colorado Avalanche	Feb 1, 2000 to Apr 9, 2000	ES and ES PP	LW, C, RW

When a team is on the power play, the coach may often decide to play four forwards and only one defenseman. For this reason, erroneous lines may be identified when two players who do not play on the same line are involved in a goal together on the power play.

The Colorado Avalanche are an example of a team that plays an additional forward when they are on the power play. In fact, Peter Forsberg usually joins the top line on the ice. As depicted in Figure 14, Colorado's lines for February 1 to April 9, 2000, consist of {Sakic, Hejduk, Tanguay}, {Reid, Yelle, Drury} and {Forsberg, Deadmarsh, Andreychuk}. However, as shown in Figure 15, when power play goals are included in the visualization, there is a strong relationship between Peter Forsberg and Joe Sakic because the two players often are involved in the same power play goals. This demonstrates that including power play goals in the analysis can be misleading if a team adjusts its lines when on the power play.



Line Assessments

Left Wing	Centre	Right Wing
A. TANGUAY (C)	J. SAKIC (C)	M. HEJDUK (RW)
D. REID (LW)	S. YELLE (C)	C. DRURY (C)
D. ANDREYCHUK (LW)	P. FORSBERG (C)	A. DEADMARSH (RW)
S. PODEIN (LW)	B. ROLSTON (LW)	C. DINGMAN (LW)
		D. HINOTE (RW)

Start Date: February 01, 2000

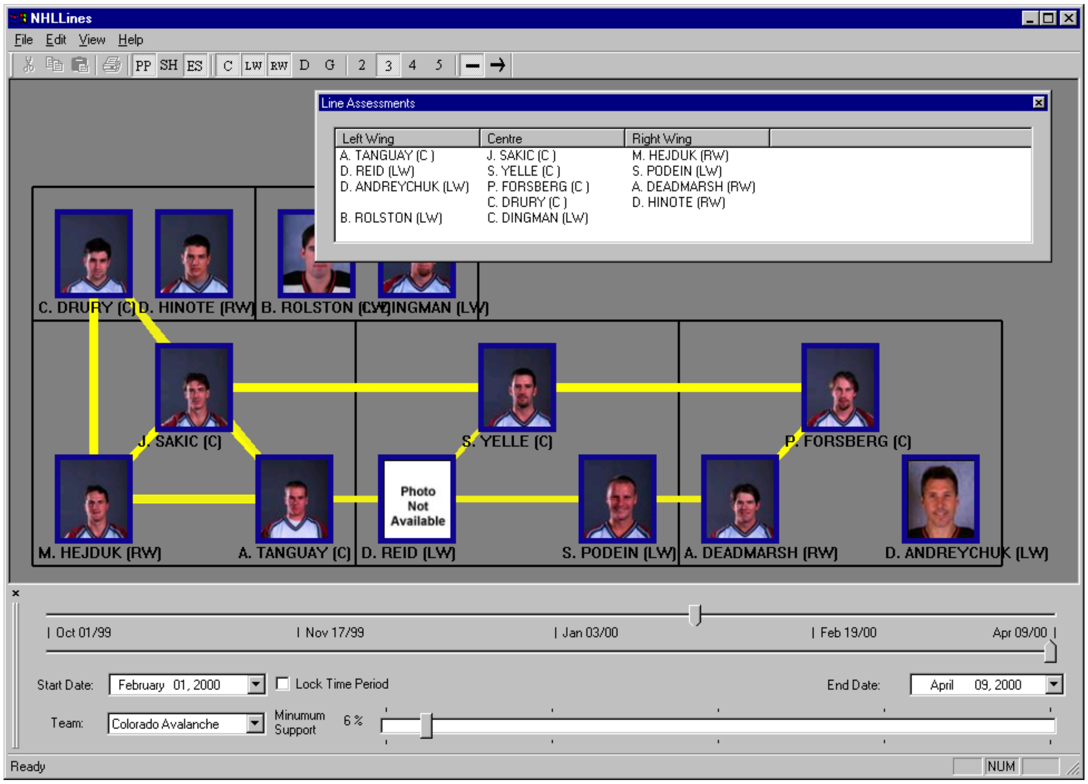
□ Lock Time Period

End Date: April 09, 2000

Team: Colorado Avalanche

Minumum Support 6 %

Figure 14: Power Play Goals Not Included



Line Assessments

Left Wing	Centre	Right Wing
A. TANGUAY (C)	J. SAKIC (C)	M. HEJDUK (RW)
D. REID (LW)	S. YELLE (C)	S. PODEIN (LW)
D. ANDREYCHUK (LW)	P. FORSBERG (C)	A. DEADMARSH (RW)
	C. DRURY (C)	D. HINOTE (RW)
B. ROLSTON (LW)	C. DINGMAN (LW)	

Start Date: February 01, 2000

□ Lock Time Period

End Date: April 09, 2000

Team: Colorado Avalanche

Minumum Support 6 %

Figure 15: Power Play Goals Included

4.5. Other Scenarios

A variety of other testing was performed in an attempt to identify power play units, short handed units and defensive pairings.

The software is able to identify some known power play units consisting of five players. This is achieved by considering only power play goals and searching for units of five players consisting of forwards and defencemen. The software is less effective when attempting to identify defensive pairings and penalty killing units. A typical NHL team only scores eight to ten short-handed goals in a season. For this reason, analysis of short-handed goals is unable to yield reliable information about the penalty killing lines. Since defencemen are involved in relatively few goals it is unlikely that two defencemen will be involved in the same goal. As a result, determining defensive line pairings is generally not successful.

As outlined in Section 2.2, the Hockey Line Extraction Algorithm makes decisions about how to break up large clusters. The decision is first made based on player position, and then based on the average support between all players in the lines. Testing revealed that these decisions are rare. When calculating the 3-player lines of forwards for all 28 NHL teams over a date range of the entire season, this scenario was observed seven times. This is encouraging because it indicates that it is common that the HLE Algorithm groups players into relatively tight clusters based on their support values and rarely creates large groups of players that have higher two-itemset support values.

In the instances when the HLE Algorithm made a decision based on position, we were unable to measure the effectiveness of the decision. The difficulty arises from the fact that these instances deal with players on the third and fourth lines. These lines are not well known for any teams because the players on these lines change often and score infrequently. For these reasons, it is difficult to compare the results with the actual lines. During our testing, we were unable to find a known line that was affected by this decision.

5. Conclusion

We created an application that extracts hockey lines and displays the relationships between hockey players. A modified version of the Single Link Clustering Algorithm is used to make line assessments. Our visualization techniques enable the user to view the relationships between players and customize the software's appearance to study these relationships with ease. While further testing is necessary, initial testing seems to indicate that the software is capable of yielding accurate results.

The Hockey Line Extraction Algorithm may not yield accurate results when a player is injured for a period of time or the lines change significantly. While the visualization techniques are helpful for identifying these situations there are a few areas for improvement. The Hockey Line Extraction Algorithm could provide a measure of how tightly clustered lines are. Furthermore, players who are injured for a significant amount of time could be marked with a different colour in the software. This may help the user understand why there are relationships between players who are not in the same line.

For the 1999-2000 season, the NHL changed its overtime rules. If two teams are tied after three periods, the teams play a five minute overtime period where both teams only play four players. Teams may combine two forwards from different lines for the overtime period that may affect results. A future version of the software will allow the user to exclude goals from different periods.

Enabling the user to delete links and players can further enhance the visualization. This would enable the user to customize the display by deleting links and players that are not currently of interest.

As discussed in the previous section, accurate assessments of actual NHL lines are difficult to obtain. Future work would involve more rigorous testing. This could be accomplished by manually keeping track of the lines for one or more teams for a significant period of time. This data could be used for comparison with the line identifications made by the Hockey Lines software.

6. References

[1] Aggarwal, C., Wolf, J. and Yu, P., “A New Method for Similarity Indexing of Market Basket Data.” *SIGMOD Record*, 28(2):407-418, 1999.

[2] Agrawal, R. and Srikant, R., “Fast Algorithms for Mining Association Rules.” *Proceedings of the 20th VLDB Conference*, pages 487-499, 1994.

[3] Carter, C., Hamilton, H. and Cercone, N., “Share Based Measures for Itemsets.” *Proceedings of Principles of Data Mining and Knowledge Discovery: First European Symposium*, pages 14-24, 1997.

[4] Cluet, S., Delobel, C., Simeon, J. and Smaga, K., “Your Mediators Need Data Conversion.” *SIGMOD Record*, 27(2):177-187, 1998.

[5] Jagadish, H. and Ng, R., “Incompleteness in Data Mining.” *2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 1-10, 2000.

[6] Ozosoyoglu, G. and Snodgrass, R., “Temporal and Real-time Databases: A Survey.” *IEEE Transactions on Knowledge and Data Engineering*, 7(4):513-532, 1995.

[7] Sibson, R., “SLINK: An Optimally Efficient Algorithm for the Single Link Clustering Method.” *The Computer Journal*, 16(1):30-34, 1973.

[8] International Business Machines. "Data Mining: Advanced Scout." http://www.research.ibm.com/scout/home.html, 1995.