Joseph Elliott

Social Network Analysis

Optional Programming Assignment

Analysis of the Major League Baseball Trade Network

1 Introduction

Professional sports teams are large, well established organizations that participate in a sort of commerce similar to corporations or nations: trade. The player trades that occur between professional sports teams potentially can be analyzed to determine relationships between the trading teams. Major League Baseball (MLB) is of particular interest, as it has a high trade volume and relatively large roster. This project will determine if a social structure exists between MLB teams by utilizing the concepts of Social Network Analysis.

2 Methods

The data collection and initial processing was completed using the R programming language, including the XML and igraph libraries (Lang, 2013; Csardi, 2013; The R Project for Statistical Computing, 2013). Once the data was properly organized into a graph (GML) format, it was exported from R and imported into Gephi for social network analysis and visualization (The Gephi Consortium, 2012).

2.1 Data Collection

Historical data on trades in Major League Baseball is surprisingly difficult to obtain through online sources. The website Spotrac.com "is the largest online sports player contract system on the internet. Including over 5,000 MLB, NFL, NBA, and NHL contracts, year by year salaries, and up to date transactions" (Spotrac.com, 2013). This provides an easy method for obtaining regularly formatted data on transactions in major league sports. An HTTP request to the website listed below provides an HTML stream of transaction data on particular sports, using the following formula.

http://www.spotrac.com/transactions/more/'index'/'sport'/

#	Sport
1	MLB (baseball)
2	NBA (basketball)
3	NFL (football)
4	NHL (hockey)
6	MLS (soccer)

The data on transactions in a sport is obtained in pages, with 25 items on a page, starting with the most current data at index=0. For this project, pages were searched for trade data between October 31, 2012 and October 31, 2013, using sport=1 (MLB).

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

A 'transaction' can include player trades, dropping players to free agency, salary increases, issuance of new contracts, and other data. A sample of the data obtained can be seen in the Appendix B. Once the transaction data was obtained from the SportsTrac.com, it was filtered using a regex to remove all data except for player trades between two teams. The filtered data was then saved to a local data store for later post-collection processing. The source code for obtaining the data can be seen in the Appendix A.

2.2 Data Processing

To extract the team names, SportsTrac.com provides a standard format; each player traded is listed individually with both the trading team and acquiring team names. This standard format made it fairly easy to extract the trading team name and the acquiring team name. For example:

```
<tr>
  <td class="transaction  ">
    <a href="http://www.spotrac.com/mlb/seattle-mariners/xavier-avery/" class="deadlink">Xavier Avery</a>
    <span class="at">at</span>
    <span class="data">Traded to <b>Seattle (SEA)</b> from <b>Baltimore (BAL)</b> for <a href="http://www.spotrac.com/player/redirectPlayer/7548/" class="tag player-tag">Mike Morse</a> </span>
  </td>
  <td class="playerinfo  ">
    <span class="team">Seattle Mariners</span>
    <span class="position">Left Field</span>
  </td>
  <td class="date  " style="width:100px">Aug 31, 2013</td>
</tr>
```

In total, 166 trades were detected and processed over the time period specified (one trade was removed due to processing error, leaving 165 valid trades). Each of the 30 MLB teams formed the nodes of our social network graph. Each trade of players between two teams was assumed to create an undirected edge between the nodes. Multiplayer and multiteam trades in particular presented a problem. For multiplayer and multiteam trades, often several players were listed separately from each team, resulting in multiple edges between the teams. Further issues with the data processing can be seen in Section 3.2. Each edge is assigned a weight corresponding to the number of player trades that took place between the two teams.

Using these qualifications, the cleaned data frame was formulated into an edgelist, and this edgelist was then imported into igraph to form a graph object. The graph was then simplified to transform duplicate edges into edge weights, and exported as a GML file to be further analyzed with Gephi.

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

3 Results

3.1 Analysis

The social graph of MLB teams was analyzed in Gephi using four different social analysis metrics to try and determine the relationship between the teams. The degree, weighted degree, modularity, and betweeness were all analyzed for the MLB team social graph, and a visualization can be seen below that sizes the nodes according to degree, and colors them according to the communities structure based on the modularity algorithm. In addition to the social analysis metrics, linear regressions of the data were performed to attempt to detect relationships between real-world factors and social network parameters. The layout of the below visualization is Force Atlas 2 (with Label Adjust).

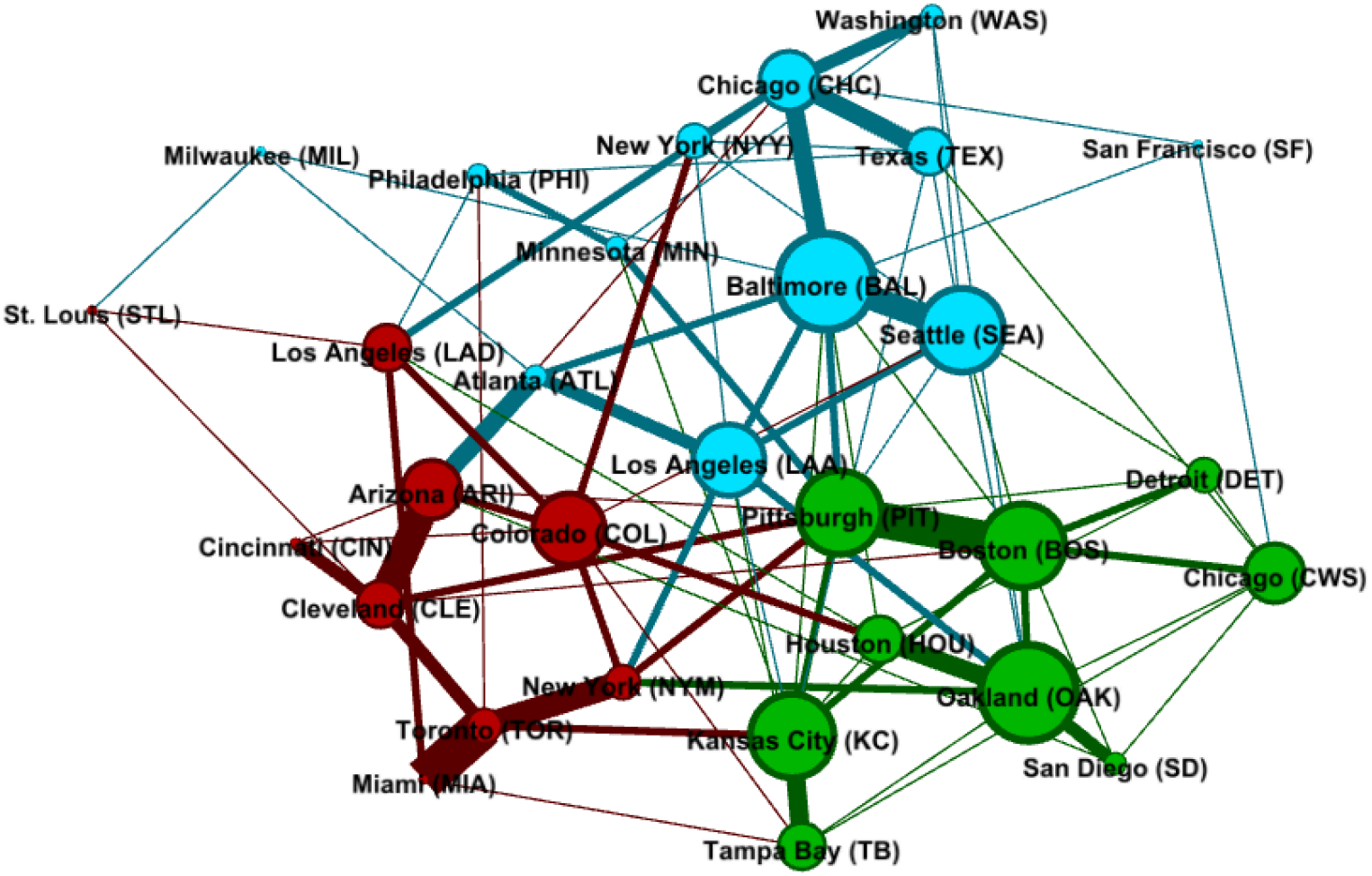

Figure 1 - MLB Trade Network 2013

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

3.1.1 Degree/Weighted Degree

The first parameter analyzed was the degree (undirected). The unweighted degree evaluates the number of other teams traded with over the time period analyzed. The highest degree (10) belongs to Baltimore and Oakland, who traded with one third of the total teams over the course of the season.

The weighted degree was also evaluated, which approximates the number of trades (and number of players) that were traded between each team over the time period analyzed. Pittsburg and Baltimore had the highest weighted degree, with 20 player trades each.

3.1.2 Community Structure

The MLB trade network from 2013 had an overall modularity of 0.371. The community finding algorithm from Gephi detected 3 communities which can be seen in Figure 1. All the teams form a single weakly connected component.

Analyzing the communities which formed along with 'natural communities' such as MLB conference or division, no easily-explainable patterns could be detected. The social communities that were formed in the trade network fall outside the real-world divisions of Major League Baseball and geography.

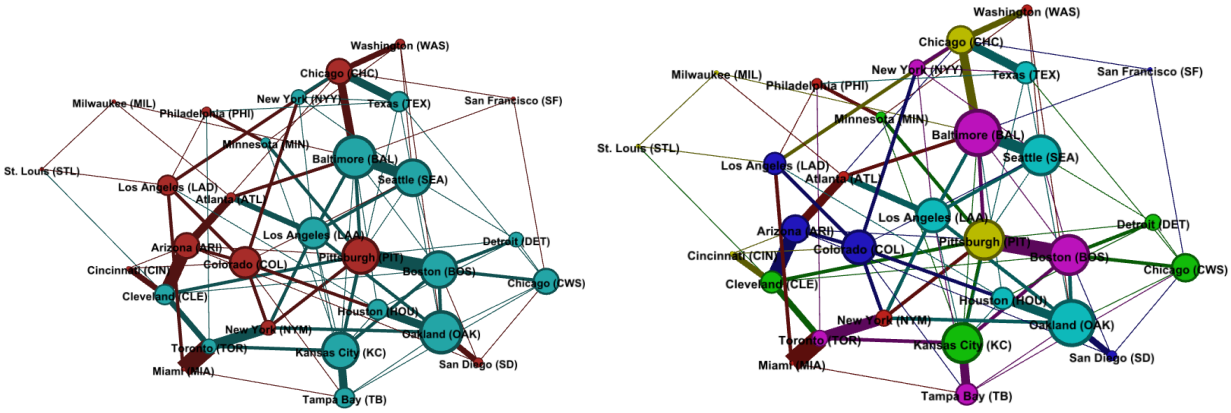

Figure 2 - MLB Trade Network colored by conference (left) and division (right)

The clustering coefficient was 0.219, with average shortest path of 2.032, compared with an equivalent random graph clustering coefficient of 0.257 with average shortest path 2.025. Interestingly, the similarity between the actual graph and the equivalent random graph implies that the MLB trade network is not a small-world (Adamic, 2013).

3.1.3 Betweeness

The betweeness of the MLB network was also analyzed. The betweeness could show if some teams are acting as a 'broker' between other teams. The MLB trade network showed that nodes with high degree are also likely to have high betweeness, including Baltimore, Kansas City, and Colorado. Therefore,

more prolific traders were more likely to have high betweeness. Interestingly, Los Angeles Dodgers (LAD) had the fourth highest betweeness, but the 12th highest degree, however, this appears to be the result of trading with two other high-degree nodes (Chicago and Colorado).

3.1.4 Linear Regression Analysis

In addition to the network analysis techniques, linear regression analysis was applied to a few real-world metrics of MLB teams, including number of wins and total team payroll. This was done to determine if any patterns exist between these parameters and the social network parameters of degree, weighted degree, and betweenness. The figures below show the linear regressions performed.

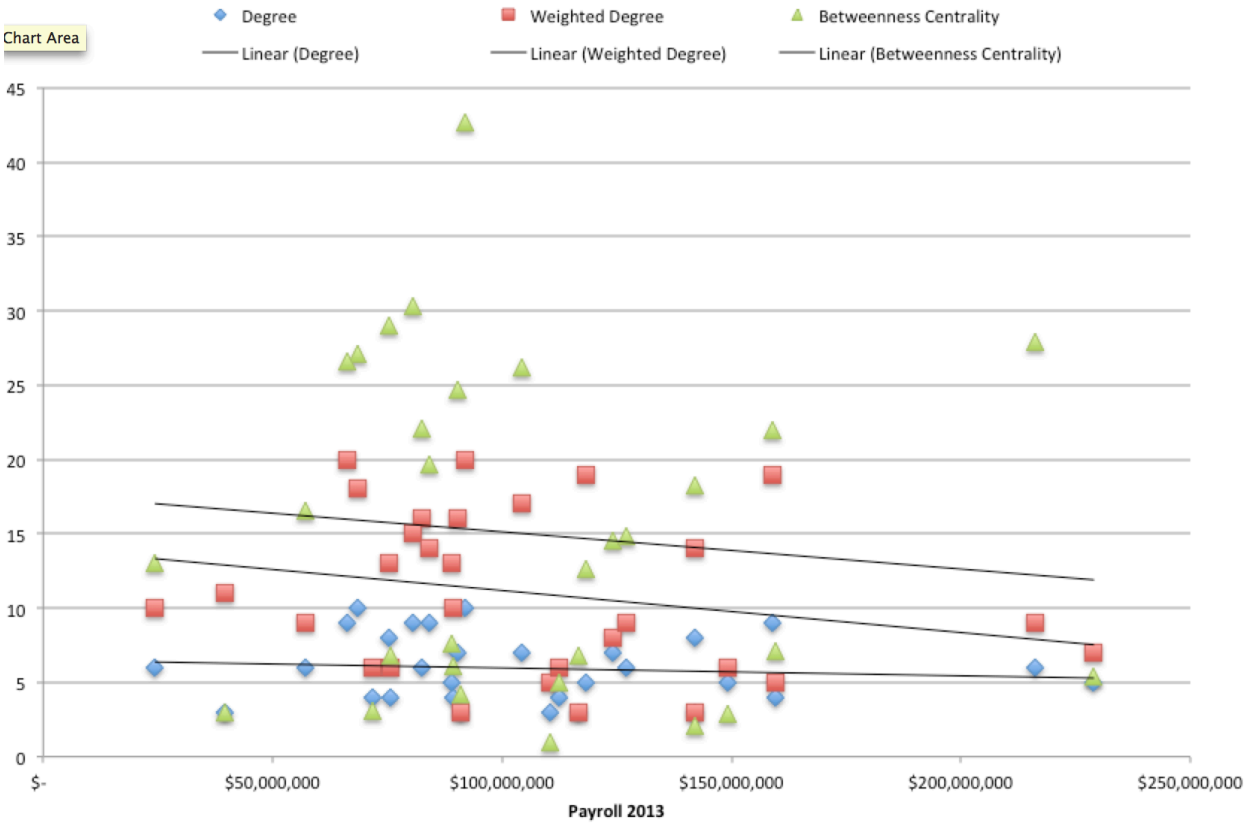

Figure 3 Linear Regression MLB 2013 Payroll

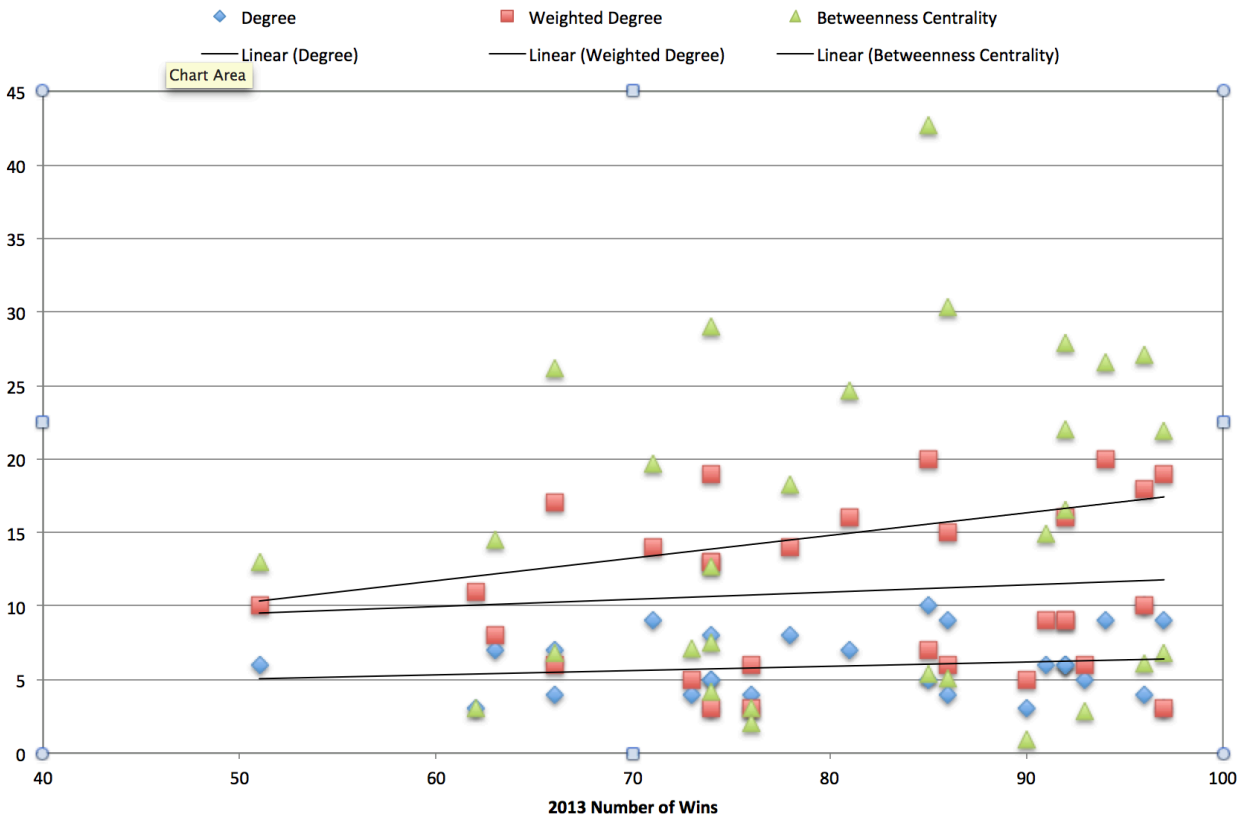

Figure 4 Linear Regression MLB 2013 Number of Wins

All linear regressions performed showed very low correlation coefficients (R^2 = <0.2), leading to the conclusion that there is no relationship between the factors mentioned above.

3.2 Limitations

The data source contributed to some of the limitations of this research. Sports trades are complex and occur under varying contract conditions, often with multiple teams involved. Trades involving multiple teams or multiple players resulted in duplicate rows of data for the same single trade date. For example, "Player 1 traded from Team A to Team B for Player 2" was sometimes additionally listed as "Player 2 traded from Team B to Team A for Player 1". This was especially common with multi-player trades, as these trades can often be very complex. This issue resulted in higher weighted degree for teams that participated in the multi-player or multi-team trades.

The data also suffered from some inconsistencies that caused the regex that parsed the team names to fail for certain date ranges. Due to this, trades that could not be parsed were dropped from the dataset. A single data point was eliminated due to this issue.

4 Conclusion

The social network analysis of Major League Baseball's trade network in 2013 provides some interesting insights into the structure of the league. Using the trade network, it is easy to detect teams that are particularly frequent trading partners, and teams that have high volume of trades. Surprisingly, the MLB trade network does not exist along conference or divisional lines. The clustering behavior of the network approximates that of the random graph, and no patterns were detected between social network behavior and real-world parameters such as team performance or payroll. Future research should be undertaken to detect underlying patterns in MLB social network.

5 Works Cited

Adamic, L. (2013). *SNA 5: small world networks.* Retrieved from Coursera.org.

Csardi, G. (2013, October 28). *igraph: Network analysis and visualization*. Retrieved November 29, 2013, from The Comprehensive R Archive Network (CRAN): http://cran.r-project.org/web/packages/igraph/index.html

ESPN Internet Ventures. (2013, September 29). *MLB Standings - 2013.* Retrieved November 20, 2013, from ESPN: http://espn.go.com/mlb/standings

Lang, D. T. (2013, June 20). *XML: Tools for parsing and generating XML within R and S-Plus*. Retrieved November 29, 2013, from The Comprehensive R Archive Network (CRAN): http://cran.r-project.org/web/packages/XML/index.html

Petchesky, B. (2013, March 29). *2013 Payrolls And Salaries For Every MLB Team.* Retrieved November 29, 2013, from Deadspin: http://deadspin.com/2013-payrolls-and-salaries-for-every-mlb-team-462765594

Spotrac.com. (2013). *Spotrac*. Retrieved November 20, 2013, from http://www.spotrac.com/

The Gephi Consortium. (2012). *Gephi*. Retrieved November 29, 2013, from https://gephi.org/

The R Project for Statistical Computing. (2013). Retrieved 2013 йил 12-February from http://www.r-project.org/

6 Appendix A – Code

```
downloadData<-function() {
      import(XML)

      #first gather the data from the pages and put it into a table
      #Action (Traded), team1, team2, date
      maxPages<-53
      n<-2
      tradeData<-data.frame()
      dates<-c()
      tradeFrame<-data.frame()
      dateDownloaded <- date()

      while(n <= maxPages) {
                html<-
            htmlTreeParse(paste0("http://www.spotrac.com/transactions/more/",
            as.character(n),"/1/"),useInternalNodes=TRUE)
              scrapedDataEven<-xpathSApply(html, "//tr/td[@class='transaction
            even']/span[@class='data']", xmlValue)
                scrapedDateEven<-xpathSApply(html, "//tr/td[@class='date
      even']", xmlValue)
                scrapedDataOdd<-xpathSApply(html,
      "//tr/td[@class='transaction  ']/span[@class='data']", xmlValue)
                scrapedDateOdd<-xpathSApply(html, "//tr/td[@class='date  ']",
      xmlValue)
                evenFrame<-
      data.frame(date=scrapedDateEven,data=as.character(scrapedDataEven))
                oddFrame<-
      data.frame(date=scrapedDateOdd,data=as.character(scrapedDataOdd))
                fullFrame<-rbind(evenFrame,oddFrame)
                tradedStrings<-grep("Traded ",fullFrame$data)
             tradeDataTemp<-fullFrame[tradedStrings,]
             #create a master list of all the trade data.
             tradeData<-rbind(tradeData,tradeDataTemp)
             n<-n+1
      }
      print(tradeData)
      print(str(tradeData))
      save(tradeData,dateDownloaded,file="tradeDataRaw2013.rda")
}
```

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

```
cleanDataSet<-function(tradeData){
      #use regex to break down the trade data into team1 and team2
      #use unlist and sapply to eliminate missing data points, rather than
      NA.
      team1<-
      unlist(sapply(regmatches(as.character(tradeData$data),regexec("Traded
      to (.*? \\([A-Z]{1,5}\\))
      from",as.character(tradeData$data))),function(x) x[2]))
        team2<-
      unlist(sapply(regmatches(as.character(tradeData$data),regexec(".* from
      (.*?\\))",as.character(tradeData$data))),function(x) x[2]))


      #Build data frame of just teams
      teams<-
      data.frame(tradedTo=team1,tradedFrom=team2,date=as.Date(tradeData$date,
      "%b %d, %Y"))
        teams

      #get rid of rows with NA
      cleanTeams<-na.omit(teams)

      #for some reason, some teams have two abbreviation types or alternative
      capitalization:
      cleanTeamName<-function(x) {
          if(x == "Baltimore (BALT)") {
            as.factor("Baltimore (BAL)")
          }else if(x=="Tampa Bay (TAM)"){
            as.factor("Tampa Bay (TB)")
          }else if(x=="Florida (FLA)"){
            as.factor("Miami (MIA)")
          }else if(x=="Chicago (CHW)"){
            as.factor("Chicago (CWS)")
          }else if(x=="PIttsburgh (PIT)"){
            as.factor("Pittsburgh (PIT)")
          }else{x}
      }
      #create cleanTeams data frame
      cleanTeams<-
      data.frame(tradedTo=factor(sapply(cleanTeams$tradedTo,cleanTeamName)),t
      radedFrom=factor(sapply(cleanTeams$tradedFrom,cleanTeamName)),
      date=cleanTeams$date)

      #check to make sure we have 30 teams
      print(sort(unique(c(levels(cleanTeams[1,1]),levels(cleanTeams[1,2])))))

      #save the cleaned data set.
      save(cleanTeams,file="cleanTeams2013.rda")
}
```

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

```
exportGMLFile<-function(cleanTeams){
      import(igraph)

      #use edgelist to create graph.  Initialize weights to zero.
      g<-graph.edgelist(as.matrix(cleanTeams[,1:2]),directed=FALSE)
      E(g)$weight <- 1

      #simplify the graph to eliminate duplicates and sum edge weights
      sg<-simplify(g)

      #node names
      print(V(sg)$name)

      #edges
      print(cbind( get.edgelist(sg) , round( E(sg)$weight, 3 )))

      #turn into a gml file.
      write.graph(sg,'mlb2013.gml',format='gml')
}
```

7 Appendix B – Sample Raw Data

```
<tr>
  <td class="transaction  ">
    <a href="http://www.spotrac.com/mlb/san-diego-padres/will-venable/">Will Venable</a>
    <span class="at">at</span>
    <span class="data">Signed a 2 year $8.5 million extension with San Diego (SD) </span>
  </td>
  <td class="playerinfo  ">
    <span class="team">San Diego Padres</span>
    <span class="position">Right Field, Left Field</span>
  </td>
  <td class="date  " style="width:100px">Sep 7, 2013</td>
</tr>

<tr>
  <td class="transaction  even">
    <a href="http://www.spotrac.com/mlb/philadelphia-phillies/miguel-alfredo-gonzalez/">Miguel
Alfredo Gonzalez</a>
    <span class="at">at</span>
    <span class="data">Signed a 3 year $12 million contract with Philadelphia (PHI) </span>
  </td>
  <td class="playerinfo  even">
    <span class="team">Philadelphia Phillies</span>
    <span class="position">Starting Pitcher</span>
  </td>
  <td class="date  even" style="width:100px">Aug 31, 2013</td>
</tr>

<tr>
  <td class="transaction  ">
    <a href="http://www.spotrac.com/mlb/seattle-mariners/xavier-avery/" class="deadlink">Xavier
Avery</a>
    <span class="at">at</span>
    <span class="data">Traded to Seattle (SEA) from Baltimore (BAL) for <a
href="http://www.spotrac.com/player/redirectPlayer/7548/" class="tag player-tag">Mike Morse</a>
</span>
  </td>
  <td class="playerinfo  ">
    <span class="team">Seattle Mariners</span>
    <span class="position">Left Field</span>
  </td>
  <td class="date  " style="width:100px">Aug 31, 2013</td>
</tr>
```

Joseph Elliott

Social Network Analysis

Optional Programming Assignment

8 Appendix C - Results

Id	Label	Conf	Division	Payroll 2013	Wins 2013	Win %	Degree	Weighted Degree	Betweenness Centrality
16	Arizona (ARI)	NL	NL West	$90,158,496	81	0.5	7	16	24.642
17	Atlanta (ATL)	NL	NL East	$89,288,192	96	0.593	4	10	6.134
4	Baltimore (BAL)	AL	AL East	$91,793,336	85	0.525	10	20	42.673
10	Boston (BOS)	AL	AL East	$158,967,280	97	0.599	9	19	21.955
13	Chicago (CHC)	NL	NL Central	$104,150,728	66	0.407	7	17	26.196
21	Chicago (CWS)	AL	AL Central	$124,065,280	63	0.389	7	8	14.555
28	Cincinnati (CIN)	NL	NL Central	$110,565,728	90	0.556	3	5	0.916
15	Cleveland (CLE)	AL	AL Central	$82,517,296	92	0.568	6	16	22.050
27	Colorado (COL)	NL	NL West	$75,449,072	74	0.457	8	13	29.024
20	Detroit (DET)	AL	AL Central	$149,046,848	93	0.574	5	6	2.877
19	Houston (HOU)	AL	AL West	$24,328,538	51	0.315	6	10	13.005
9	Kansas City (KC)	AL	AL Central	$80,491,728	86	0.531	9	15	30.352
22	Los Angeles (LAA)	AL	AL West	$142,165,248	78	0.481	8	14	18.214
2	Los Angeles (LAD)	NL	NL West	$216,302,912	92	0.568	6	9	27.896
29	Miami (MIA)	NL	NL East	$39,621,900	62	0.383	3	11	3.011
7	Milwaukee (MIL)	NL	NL Central	$91,003,368	74	0.457	3	3	4.203
1	Minnesota (MIN)	AL	AL Central	$75,562,496	66	0.407	4	6	6.807
8	New York (NYM)	NL	NL East	$88,877,000	74	0.457	5	13	7.553
26	New York (NYY)	AL	AL East	$228,995,952	85	0.525	5	7	5.392
11	Oakland (OAK)	AL	AL West	$68,577,000	96	0.593	10	18	27.085
3	Philadelphia (PHI)	NL	NL East	$159,578,208	73	0.451	4	5	7.133
0	Pittsburgh (PIT)	NL	NL Central	$66,289,524	94	0.58	9	20	26.617
24	San Diego (SD)	NL	NL West	$71,689,904	76	0.469	4	6	3.033
25	San Francisco (SF)	NL	NL West	$142,180,336	76	0.469	3	3	2.026
5	Seattle (SEA)	AL	AL West	$84,295,952	71	0.438	9	14	19.672
6	St. Louis (STL)	NL	NL Central	$116,702,088	97	0.599	3	3	6.836
23	Tampa Bay (TB)	AL	AL East	$57,030,272	92	0.564	6	9	16.586
18	Texas (TEX)	AL	AL West	$127,197,576	91	0.558	6	9	14.876
14	Toronto (TOR)	AL	AL East	$118,244,040	74	0.457	5	19	12.650
12	Washington (WAS)	NL	NL East	$112,431,768	86	0.531	4	6	5.030