Association Rules

Assignment:

Using the Book Data dataset (https://github.com/WinVector/zmPDSwR/blob/master/Bookdata/bxBooks.RData) find association rules in user preferences.

First, I prepared the environment, loaded required libraries and the dataset:

```
> setwd("YOUR_PATH") #Set working directory for the assignment
> getwd() #Check working directory
[1] "YOUR_PATH"
>
> ###Load packages
> library(arules)
> library(arulesViz)
> library(sqldf)
> library(ggplot2)
>
> #Load data
> load("bxBooks.RData")
```

The dataset contains three data frames bxBookRatings, bsBooks, and bxUsers. bxBookRatings contains 1149780 observations of 3 variables, bxBooks contains 271379 observations of 8 variables and bxUsers includes 278858 observations of 3 variables.

```
> ###############EDA and data pre-processing################################
>
> #bxBookRatings dafa frame
> str(bxBookRatings)
'data.frame':      1149780 obs. of  3 variables:
 $ User.ID    : int  276725 276726 276727 276729 276729 276733 276736 276737 276744 276745 ...
 $ ISBN       : chr  "034545104X" "0155061224" "0446520802" "052165615X" ...
 $ Book.Rating: int  0 5 0 3 6 0 8 6 7 10 ...
```

First, I looked at the internal structure of the data frames, deleted periods from the column names for ease of manipulation and looked at the first few rows in each data frame.

```
> #delete periods in the column names
> colnames(bxBookRatings) <- gsub(".", "", colnames(bxBookRatings), fixed=T)
> #Look at the first few rows
> head(bxBookRatings)
```

```
  UserID       ISBN BookRating
1 276725 034545104X          0
2 276726 0155061224          5
3 276727 0446520802          0
4 276729 052165615X          3
5 276729 0521795028          6
6 276733 2080674722          0
```

User data:

```
> #bxUsers dataframe
> str(bxUsers)
'data.frame':      278858 obs. of  3 variables:
 $ User.ID : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Location: chr  "nyc, new york, usa" "stockton, california, usa" "moscow, yukon territory, russ
ia" "porto, v.n.gaia, portugal" ...
 $ Age     : chr  "NULL" "18" "NULL" "17" ...
> #delete periods in the column names
> colnames(bxUsers) <- gsub(".", "", colnames(bxUsers), fixed=T)
> #Look at the first few rows
> head(bxUsers)
  UserID                           Location  Age
1      1                nyc, new york, usa NULL
2      2          stockton, california, usa   18
3      3    moscow, yukon territory, russia NULL
4      4          porto, v.n.gaia, portugal   17
5      5 farnborough, hants, united kingdom NULL
6      6      santa monica, california, usa   61
```

For this assignment, I was not planning to track either age or geographical distribution of users, so I did not include it in the further analysis focusing on the remaining two data frames.

The information about the books is included in the bxBooks data frame, which has 8 variables

```
> #bxBooks dataframe
> str(bxBooks)
'data.frame':      271379 obs. of  8 variables:
 $ ISBN               : chr  "0195153448" "0002005018" "0060973129" "0374157065" ...
 $ Book.Title         : chr  "Classical Mythology" "Clara Callan" "Decision in Normandy" "Flu: Th
e Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It" ...
 $ Book.Author        : chr  "Mark P. O. Morford" "Richard Bruce Wright" "Carlo D'Este" "Gina Bar
i Kolata" ...
 $ Year.Of.Publication: int  2002 2001 1991 1999 1999 1991 2000 1993 1996 2002 ...
 $ Publisher          : chr  "Oxford University Press" "HarperFlamingo Canada" "HarperPerennial"
"Farrar Straus Giroux" ...
 $ Image.URL.S        : chr  "http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg" "http
://images.amazon.com/images/P/0002005018.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0060
973129.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0374157065.01.THUMBZZZ.jpg" ...
```

```
 $ Image.URL.M        : chr  "http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg" "http
://images.amazon.com/images/P/0002005018.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0060
973129.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0374157065.01.MZZZZZZZ.jpg" ...
 $ Image.URL.L        : chr  "http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg" "http
://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg" "http://images.amazon.com/images/P/0060
973129.01.LZZZZZZZ.jpg" "http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg" ...
> bxBooks <- bxBooks[-c(6:8)] #drop columns containing pictures URLs
> #delete periods in the column names
> colnames(bxBooks) <- gsub(".", "", colnames(bxBooks), fixed=T)
```

Three of the columns contains URL for small, medium and large images of the books, which I would not need for the association rules analysis. So, I dropped columns 6 to 8. As a result, the sample of the first few rows of the data is as follow:

```
> #Look at the first few rows
> head(bxBooks)
        ISBN                                                                                    BookTitle
1 0195153448                                                                                Classical Mythology
2 0002005018                                                                                       Clara Callan
3 0060973129                                                                               Decision in Normandy
4 0374157065 Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It
5 0393045218                                                                             The Mummies of Urumchi
6 0399135782                                                                             The Kitchen God's Wife
           BookAuthor YearOfPublication                Publisher
1   Mark P. O. Morford              2002    Oxford University Press
2 Richard Bruce Wright              2001      HarperFlamingo Canada
3         Carlo D'Este              1991           HarperPerennial
4     Gina Bari Kolata              1999       Farrar Straus Giroux
5      E. J. W. Barber              1999 W. W. Norton &amp;amp; Company
6             Amy Tan              1991           Putnam Pub Group
>
```

The BookTitle column required some additional text preprocessing (removing special characters, removing parenthesis the end of the titles, converting titles to the lower case) in order to facilitate comparisons for those books that have multiple editions included in the dataset (the same book title having different ISBNs):

```
> #Code source for clean-up:https://github.com/WinVector/zmPDSwR/blob/master/Bookdata/create_book
data.R
> #bxBooks dataframe
> Sys.setlocale('LC_ALL','C') # for non-English characters
[1] "C"
> # Clean up book titles (delete parenthesis,)
> bxBooks$BookTitle <- gsub("(", "#", bxBooks$BookTitle, fixed=T)
> bxBooks$BookTitle <- gsub("^#", "(", bxBooks$BookTitle)
> bxBooks$BookTitle <- gsub("#.*$", "", bxBooks$BookTitle)
> bxBooks$BookTitle <- sub("[[:space:]]+$","", bxBooks$BookTitle)  #save cleaned-up titles
> bxBooks$BookTitle<- tolower(bxBooks$BookTitle) #convert titles to the lower case
```

In the next step, merged the two data frames with pertinent information (bxBookRatings and bxBooks) into one data frame using ISBN as the key field:

```
> #merge by ISBN
> books_merged <-merge(bxBookRatings, bxBooks, by="ISBN")
```

The resulting data frame (books_merged) contains 1031176 observations of 7 variables (ISBN, UserID, BookRating, BookTitle, BookAuthor, YearOFPublication, and Publisher):

```
> #check the resulting data frame
> str(books_merged)
'data.frame':      1031176 obs. of  7 variables:
 $ ISBN            : chr  "0000913154" "0001010565" "0001010565" "0001046438" ...
 $ UserID          : int  171118 86123 209516 23902 196149 23902 206300 23902 244994 246671 ...
 $ BookRating      : int  8 0 0 9 0 6 0 9 0 0 ...
 $ BookTitle       : chr  "the way things work: an illustrated encyclopedia of technology" "mog'
s christmas" "mog's christmas" "liar" ...
 $ BookAuthor      : chr  "C. van Amerongen (translator)" "Judith Kerr" "Judith Kerr" "Stephen F
ry" ...
 $ YearOfPublication: int  1967 1992 1992 0 1992 1993 1999 1993 2000 2000 ...
 $ Publisher       : chr  "Simon &amp; Schuster" "Collins" "Collins" "Harpercollins Uk" ...
```

```
> head(books_merged)
        ISBN UserID BookRating                                                    BookTitle
1 0000913154 171118          8 the way things work: an illustrated encyclopedia of technology
2 0001010565  86123          0                                                mog's christmas
3 0001010565 209516          0                                                mog's christmas
4 0001046438  23902          9                                                           liar
5 0001046713 196149          0                                  twopence to cross the mersey
6 000104687X  23902          6                t.s. eliot reading 'the wasteland' and other poems
                    BookAuthor YearOfPublication                Publisher
1 C. van Amerongen (translator)             1967     Simon &amp; Schuster
2                   Judith Kerr             1992                  Collins
3                   Judith Kerr             1992                  Collins
4                   Stephen Fry                0         Harpercollins Uk
5               Helen Forrester             1992 HarperCollins Publishers
6                    T.S. Eliot             1993 HarperCollins Publishers
```

At this point, to decrease the size of the data frame, I tried to remove those records that represent books with a rating of zero. However, later when trying to find association rules, I ran into a situation when I was not able to find any meaningful connections even with very low minimum support levels despite a dramatically increased processing time. So, in the final version of the code, I kept all transactions for all books, including the zero-rated books.

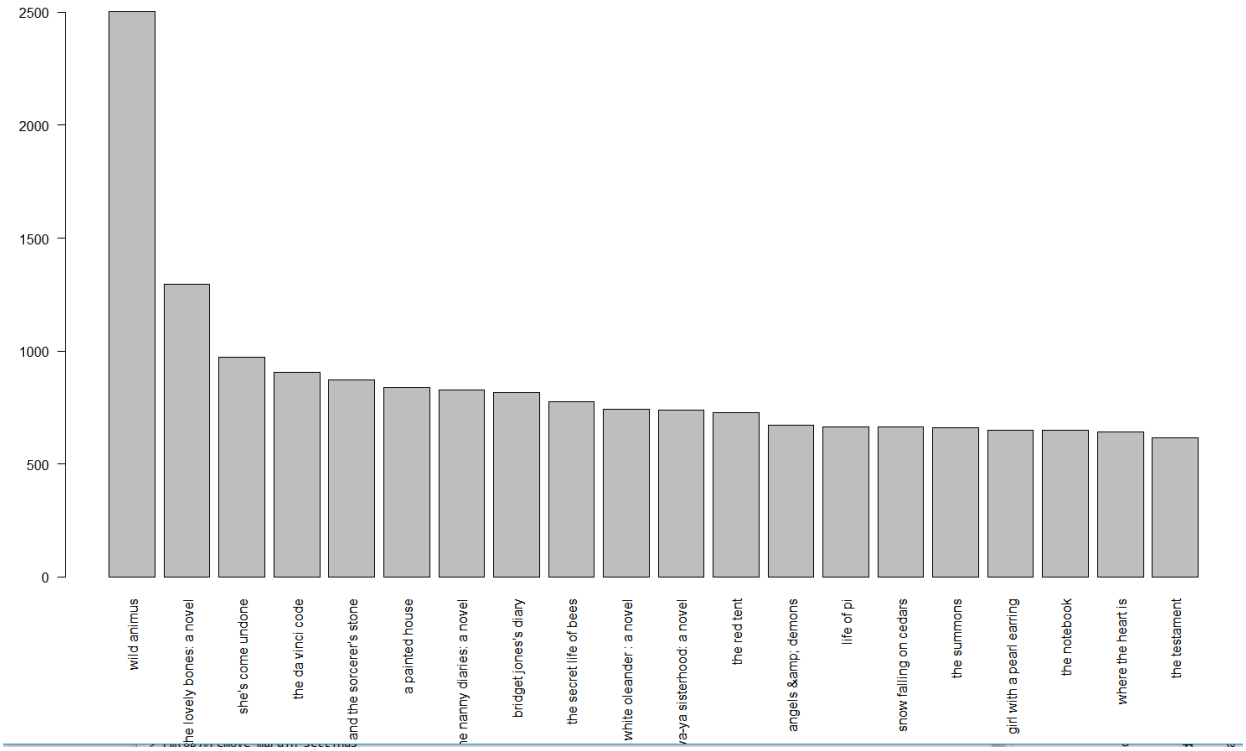
To better understand the data, I looked at the top 20 book titles:

```
> #create a histogram of top 20 book titles
> sorted_titles <- sort(table(books_merged$BookTitle), decreasing = TRUE)
> top20titles <- sorted_titles[1:20]
> top20titles
```

```
                                    wild animus                     the lovely bones: a novel
                                           2502                                           1295
                                she's come undone                             the da vinci code
                                            974                                            907
          harry potter and the sorcerer's stone                               a painted house
                                            871                                            839
                     the nanny diaries: a novel                          bridget jones's diary
                                            828                                            815
                           the secret life of bees                       white oleander : a novel
                                            774                                            743
divine secrets of the ya-ya sisterhood: a novel                                  the red tent
                                            740                                            727
                              angels &amp; demons                                    life of pi
                                            670                                            664
                            snow falling on cedars                                 the summons
                                            663                                            660
                         girl with a pearl earring                                the notebook
                                            651                                            651
                                where the heart is                              the testament
                                            643                                            617
```

The most frequent book in the dataset is Wild Animus by Rich Shapero, followed by The Lovely Bones by Alice Sebold. The top 20 books are presented on the histogram below:

```
> op <- par(mar=c(10,4,4,2))#set margins sizes for book title space
> barplot(top20titles,  las=2)#las=2 labels are perpendicular to axis
> rm(op)#remove margin settings
```

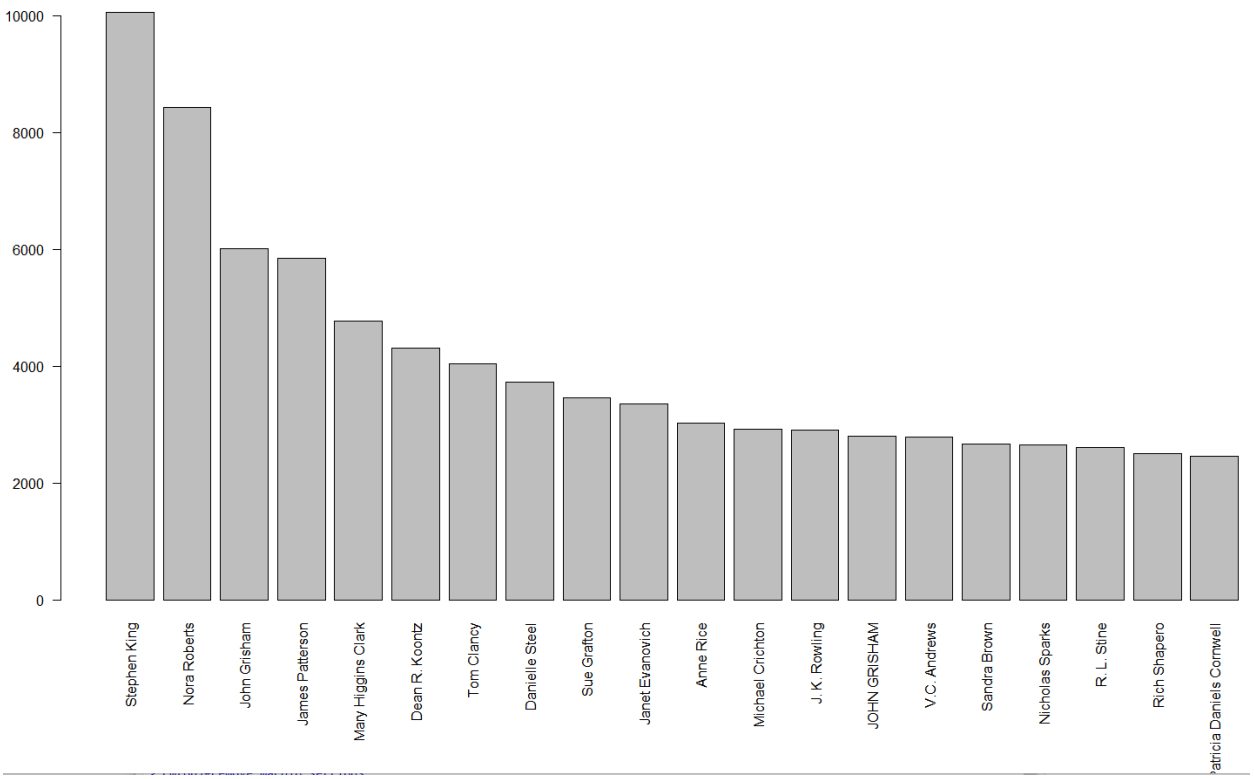


I used the same procedure to draw a histogram of top 20 authors whose books are include in the dataset:

```
> #create a histogram of top 20 authors
> sorted_authors <- sort(table(books_merged$BookAuthor), decreasing = TRUE)
> top20authors <- sorted_authors[1:20]
> top20authors
> top20authors
```

```
          Stephen King           Nora Roberts           John Grisham         James Patterson
                 10053                   8429                   6010                   5845
    Mary Higgins Clark          Dean R. Koontz            Tom Clancy         Danielle Steel
                  4777                   4313                   4036                   3726
          Sue Grafton         Janet Evanovich             Anne Rice       Michael Crichton
                  3457                   3350                   3030                   2921
        J. K. Rowling           JOHN GRISHAM           V.C. Andrews           Sandra Brown
                  2908                   2808                   2785                   2663
      Nicholas Sparks           R. L. Stine       Rich Shapero Patricia Daniels Cornwell
                  2650                   2606                   2502                   2461
> op <- par(mar=c(10,4,4,2))#set margins sizes for book title space
> barplot(top20authors,  las=2)#las=2 labels are perpendicular to axis
> rm(op)#remove margin settings
```

As the graph below demonstrates, the most popular author is Stephen King, followed by Nora Roberts and John Grisham:



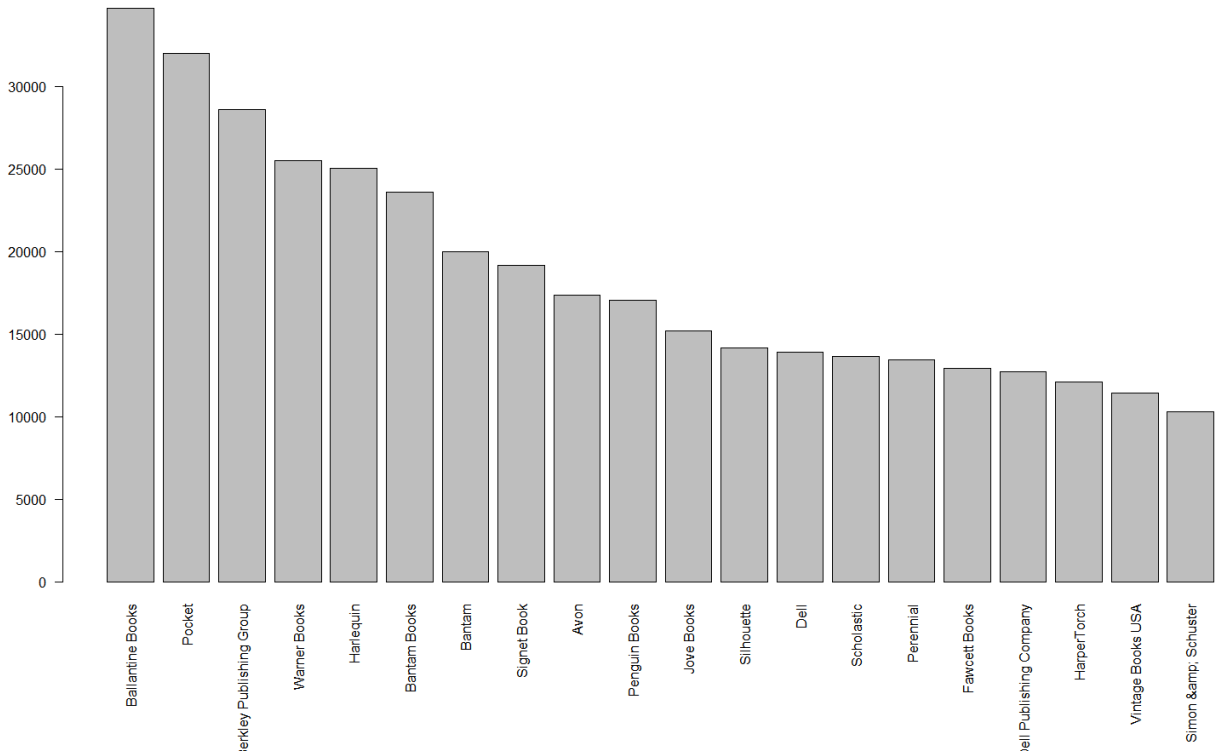
The data set includes books published by many different publishing houses, with the top three represented by Ballantine Books, Pocket and Berkley Publishing Group.

```
> #create a histogram of top 20 publishers
> sorted_publishers <- sort(table(books_merged$Publisher), decreasing = T)
> top20publishers <- sorted_publishers[1:20]
> top20publishers
```

Ballantine Books	Pocket	Berkley Publishing Group	Warner Books
34724	31989	28614	25506
Harlequin	Bantam Books	Bantam	Signet Book
25029	23600	20007	19155
Avon	Penguin Books	Jove Books	Silhouette
17352	17033	15178	14184
Dell	Scholastic	Perennial	Fawcett Books
13924	13662	13466	12905
Dell Publishing Company	HarperTorch	Vintage Books USA	Simon & Schuster
12733	12081	11427	10318

```
> op <- par(mar=c(10,4,4,2))#set margins sizes for book title space
> barplot(top20publishers,  las=2)#las=2 labels are perpendicular to axis
> rm(op)#remove margin settings
```

The rest of the top 20 publishing houses is presented below:



In order to proceed with the transaction analysis, I converted all variables to factors and saved the file in the tab-separated file format using the following code:

```
> #convert each attributes to a factor
> books_merged$ISBN <- as.factor(books_merged$ISBN)
> books_merged$UserID <- as.factor(books_merged$UserID)
> books_merged$BookRating <- as.factor(books_merged$BookRating)
> books_merged$BookTitle <- as.factor(books_merged$BookTitle)
> books_merged$BookAuthor <- as.factor(books_merged$BookAuthor)
> books_merged$YearOfPublication <- as.factor(books_merged$YearOfPublication)
> books_merged$Publisher <- as.factor(books_merged$Publisher)
>
> ###save the file in the tab-separated file format
> write.table(books_merged, file="books_merged.tsv", sep="\t", row.names = FALSE, col.names = TRU
E)
```

Next, I converted the file to the transaction class using the read.transaction() command from the arules package:

```
> ###convert data file to the transaction class
> bookbsk <- read.transactions('books_merged.tsv', cols=c("UserID", "BookTitle"), format = "singl
e", sep="\t", rm.duplicates = TRUE)
```

I used class(), colnames(), summary(), dim(), colnames(), rownames() and other commands to explore transactions objects (a matrix in the sparse format), that has 92108 transactions (rows, UserIDs) and 220447 columns (book titles):

```
> bookbsk #prints object type and dimentions
transactions in sparse format with
 92108 transactions (rows) and
 220447 items (columns)
```

Next, I looked at the sizes of transactions included in the matrix:

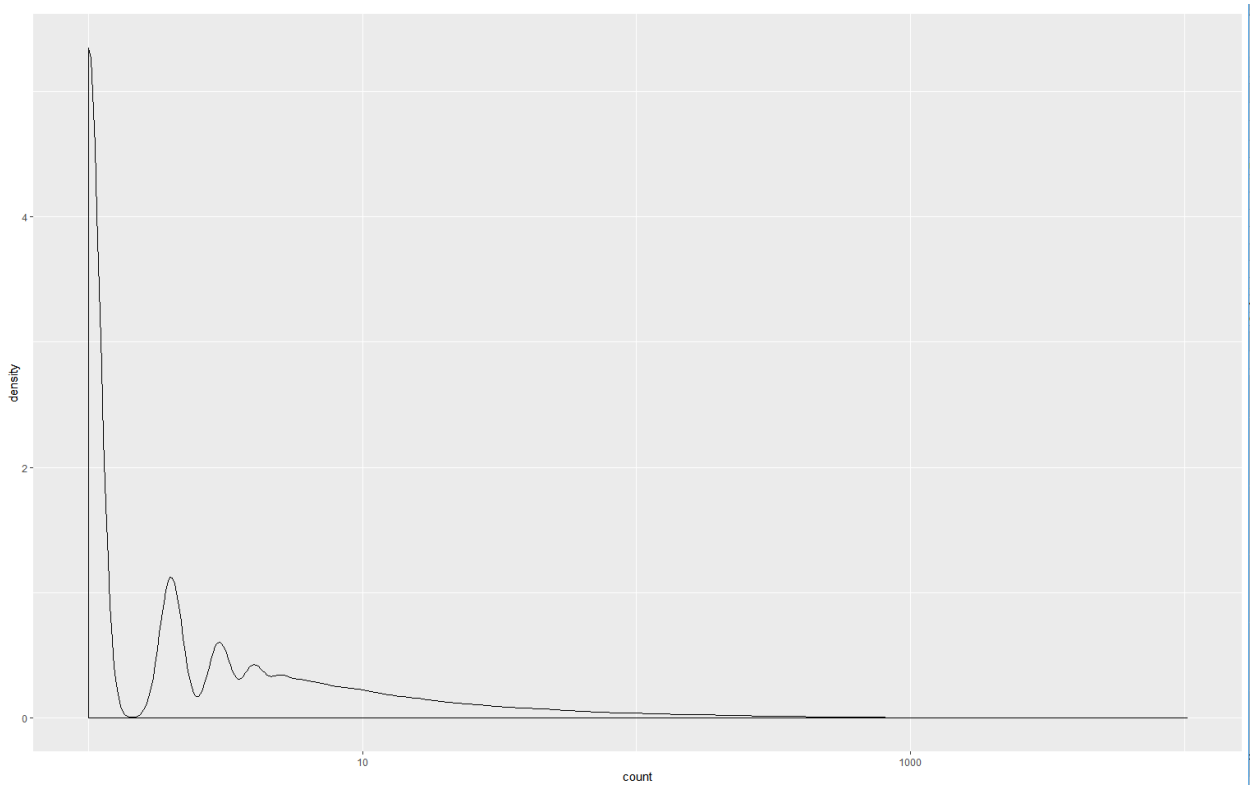
```
> #transaction sizes
> transactionsizes<-size(bookbsk)
> summary(transactionsizes) #distribution of transaction sizes
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     1.0     1.0    11.1     4.0 10253.0
```

More than half of users were interested in one book only, as a minimum, first quartile, and the median of transaction sizes are all equal to 1. The distribution of transaction sizes is considerably skewed to the right as the mean is 11.1 due to the maximum number of books equal to 10253.0, while the third quartile is only 4.0. So, there is a very limited number of users who were interested in a very large number of books. More precisely, 90% of users were interested in 13 books or less:

```
> quantile(transactionsizes, probs=seq(0, 1, 0.1)) #transaction size distribution in 10% incremen
ts
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
    1     1     1     1     1     1     2     3     5    13 10253
```

Below is the density distribution graph for transaction sizes:

```
> #plot the distribution
> ggplot(data.frame(count=transactionsizes)) + geom_density(aes(x=count)) + scale_x_log10()
```




```
> quantile(transactionsizes, probs = c(0.99, 1))
  99%  100%
  179 10253
```

Since finding association rules requires transaction having a length of at least two, in order to decrease the size of the sparse matrix, I filtered out the data for users interested in only one book using the following:

```
> ###Filter data for users interested in more than one book
> dim(bookbsk) #dimension before redaction
[1]  92108 220447
> bksize<- size(bookbsk)
> bookbsk_2up <- bookbsk[bksize>1] #saving only transactions with >1 book
> dim(bookbsk_2up) #dimension after dropping transactions with 1 book
[1]  40822 220447
```

As the output above shows, this allowed decreasing the size of the matrix from 92,108 observations (users) of 220,447 variables (books) to 40,822 observations (users) of 220,447 variables (books).

Next, I used apriori() algorithm to find the association rules between the transactions.

First, I used the recommended parameters – the minimum support of 0.005 and the minimum confidence equal to 0.70.

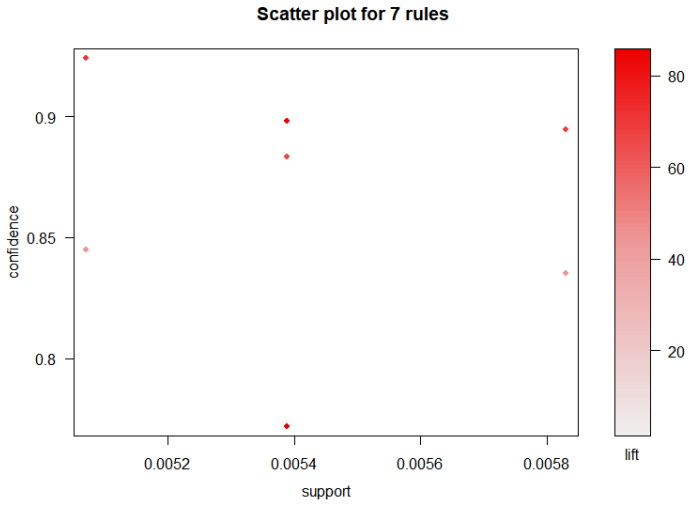
```
> rules1 <- apriori(bookbsk_2up, parameter = list(support = 0.005, confidence = 0.70))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.7    0.1    1 none FALSE            TRUE       5   0.005      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 204

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[216031 item(s), 40822 transaction(s)] done [0.98s].
sorting and recoding items ... [287 item(s)] done [0.05s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 done [0.01s].
writing ... [7 rule(s)] done [0.00s].
creating S4 object  ... done [0.03s].
> summary(rules1)
set of 7 rules

rule length distribution (lhs + rhs):sizes
3
7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      3       3       3       3       3       3

summary of quality measures:
    support           confidence          lift           count
 Min.   :0.005071   Min.   :0.7719   Min.   :44.39   Min.   :207.0
 1st Qu.:0.005230   1st Qu.:0.8400   1st Qu.:55.73   1st Qu.:213.5
 Median :0.005389   Median :0.8835   Median :67.39   Median :220.0
 Mean   :0.005424   Mean   :0.8646   Mean   :66.17   Mean   :221.4
 3rd Qu.:0.005610   3rd Qu.:0.8963   3rd Qu.:77.27   3rd Qu.:229.0
 Max.   :0.005830   Max.   :0.9241   Max.   :85.45   Max.   :238.0

mining info:
        data ntransactions support confidence
 bookbsk_2up         40822   0.005        0.7
```

It resulted in a set of 7 rules presented below in the decreasing order of confidence:

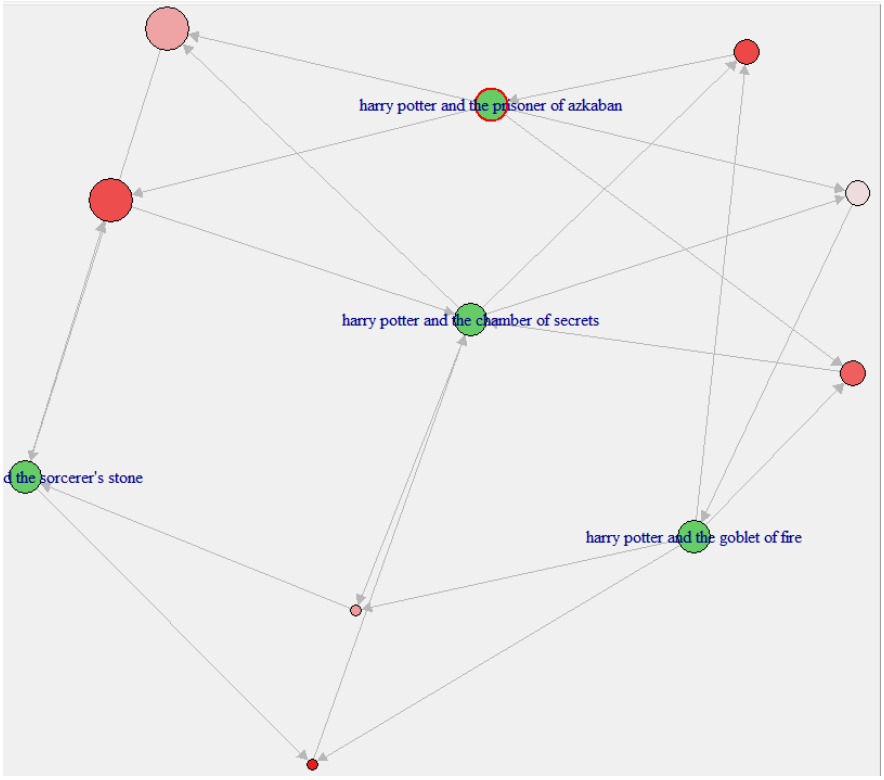
```
> inspect(sort(rules1, by="confidence", decreasing=TRUE))
    lhs                                                rhs                                                support confidence     lift count
[1] {harry potter and the goblet of fire,
     harry potter and the sorcerer's stone}     => {harry potter and the chamber of secrets}  0.005070795  0.9241071 69.60129   207
[2] {harry potter and the chamber of secrets,
     harry potter and the goblet of fire}       => {harry potter and the prisoner of azkaban} 0.005389251  0.8979592 85.44636   220
[3] {harry potter and the prisoner of azkaban,
     harry potter and the sorcerer's stone}     => {harry potter and the chamber of secrets}  0.005830190  0.8947368 67.38920   238
[4] {harry potter and the goblet of fire,
     harry potter and the prisoner of azkaban} => {harry potter and the chamber of secrets}  0.005389251  0.8835341 66.54544   220
[5] {harry potter and the chamber of secrets,
     harry potter and the goblet of fire}       => {harry potter and the sorcerer's stone}    0.005070795  0.8448980 44.90941   207
[6] {harry potter and the chamber of secrets,
     harry potter and the prisoner of azkaban} => {harry potter and the sorcerer's stone}    0.005830190  0.8350877 44.38796   238
[7] {harry potter and the chamber of secrets,
     harry potter and the prisoner of azkaban} => {harry potter and the goblet of fire}      0.005389251  0.7719298 84.93725   220
```

All these seven rules represent various editions of the books from the Harry Potter series, and show relatively high levels of confidence as shown on the scatter plot below:

```
> #visualizing rules
> plot(rules1)
```

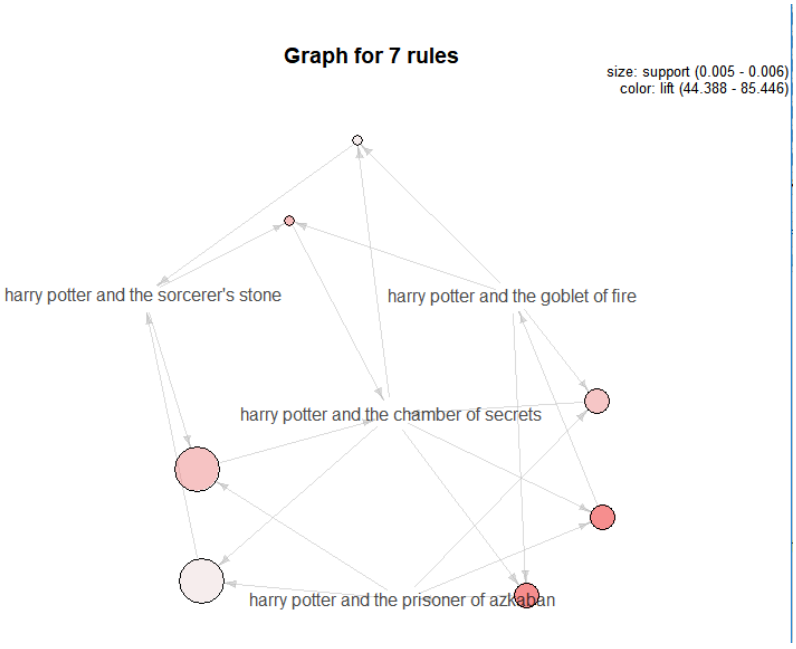


The following plot conveniently presents the relationships between the books:



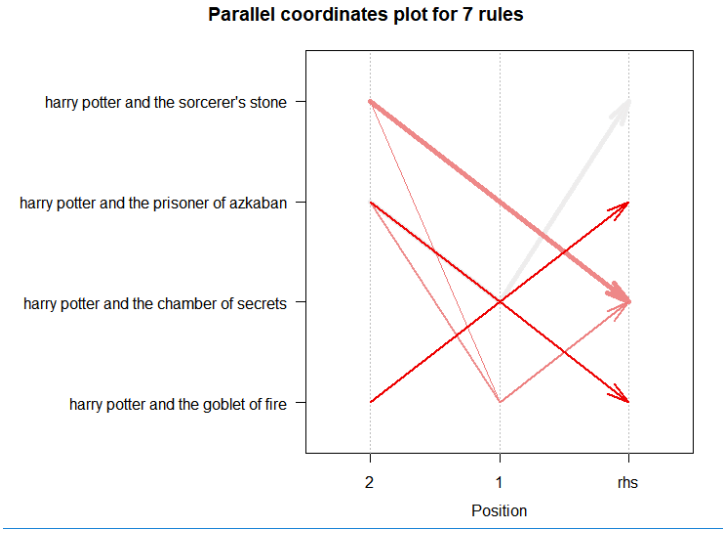
The support for these rules ranges between 0.005 and 0.006 and lift between 44.388 and 85.446 as seen on the following grasp:

```
> plot(rules1, method="graph", control=list(type="items"))
```



The same rules presented as parallel coordinates plot:

```
> plot(rules1, method="paracoord", control = list(reorder=TRUE))
```



While the above graphs nicely illustrate the relationships between transactions, from the practical point of view, this set of seven rules does not provide much interesting information. It is to be expected, that books in the same popular series would be related.

So, I tried to use different parameters in the appriori() command in order to find new association rules.

First, I kept the same confidence level of 0.70, but decreased the support to 0.001. It resulted in the 5108 association rules found.

```
> #Change parameteres using apriori()
> rules2 <- apriori(bookbsk_2up, parameter = list(support = 0.001, confidence = 0.70))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.7    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 40
```

```
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[216031 item(s), 40822 transaction(s)] done [0.92s].
sorting and recoding items ... [3172 item(s)] done [0.03s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.22s].
writing ... [5108 rule(s)] done [0.03s].
creating S4 object  ... done [0.03s].
> summary(rules2)
set of 5108 rules

rule length distribution (lhs + rhs):sizes
   2    3    4    5    6    7    8
  72 1266 2358 1043  299   62    8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   4.088   5.000   8.000

summary of quality measures:
    support           confidence          lift           count
 Min.   :0.001004   Min.   :0.7000   Min.   : 25.7   Min.   : 41.0
 1st Qu.:0.001053   1st Qu.:0.7500   1st Qu.: 67.4   1st Qu.: 43.0
 Median :0.001151   Median :0.8182   Median :138.4   Median : 47.0
 Mean   :0.001264   Mean   :0.8241   Mean   :158.3   Mean   : 51.6
 3rd Qu.:0.001298   3rd Qu.:0.8889   3rd Qu.:194.7   3rd Qu.: 53.0
 Max.   :0.005830   Max.   :1.0000   Max.   :610.7   Max.   :238.0

mining info:
        data ntransactions support confidence
 bookbsk_2up         40822   0.001        0.7
```

It is very computationally intensive to process too many rules, and limiting the maximum length would not help to decrease the size much, and 72 of them has size 2, 1266 rules have the size 3, 2358 rules have the side 4, and another 1043 rules have the size 5.

Next, I kept the support parameter at a low level of 0.001 and increased the minimum confidence level to 0.75:

```
> #Change parameteres using apriori()
> rules3 <- apriori(bookbsk_2up, parameter = list(support = 0.001, confidence = 0.75))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
       0.75    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 40
```

```
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[216031 item(s), 40822 transaction(s)] done [0.95s].
sorting and recoding items ... [3172 item(s)] done [0.03s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.22s].
writing ... [3907 rule(s)] done [0.02s].
creating S4 object  ... done [0.04s].
> summary(rules3)
set of 3907 rules

rule length distribution (lhs + rhs):sizes
   2    3    4    5    6    7    8
  41  810 1784  931  279   55    7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.000   4.000   4.202   5.000   8.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001004   Min.   :0.7500   Min.   : 36.36   Min.   : 41.00
 1st Qu.:0.001053   1st Qu.:0.7975   1st Qu.: 75.19   1st Qu.: 43.00
 Median :0.001151   Median :0.8462   Median :153.74   Median : 47.00
 Mean   :0.001282   Mean   :0.8550   Mean   :173.51   Mean   : 52.34
 3rd Qu.:0.001347   3rd Qu.:0.9111   3rd Qu.:211.97   3rd Qu.: 55.00
 Max.   :0.005830   Max.   :1.0000   Max.   :610.70   Max.   :238.00

mining info:
        data ntransactions support confidence
 bookbsk_2up         40822   0.001       0.75
```

As the output above shows, it resulted in 3907 association rules found.

Trying to decrease the number of rules to process to more manageable level while still allowing less obvious relationships to be detected, I increased the support level to 0.002 with the same confidence level of 0.75.

```
> #Change parameteres using apriori()
> rules4 <- apriori(bookbsk_2up, parameter = list(support = 0.002, confidence = 0.75))
```

It dramatically decreased the number of rules to 191:

```
> summary(rules4)
set of 191 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5
 11 100  66  14
```

However, after removing redundant rules using the following code, there were only 10 rules left:

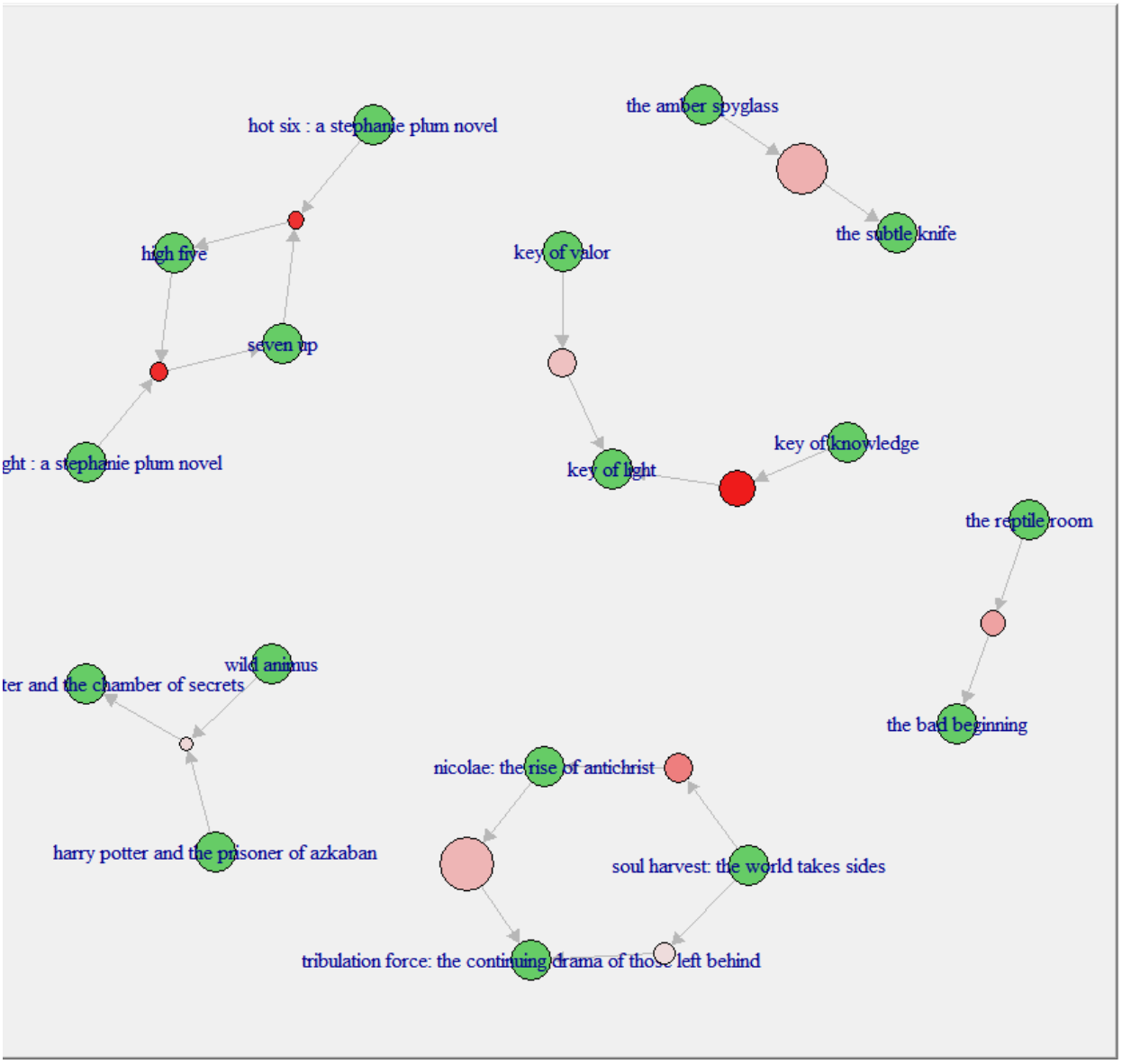
```
> #Remove redundant rules
> redundant <- which (colSums(is.subset(rules4, rules4))>1) #vector of redundant rules
> rules5<-rules4[-redundant] #remove redundant rules
> summary(rules5)
set of 10 rules

rule length distribution (lhs + rhs):sizes
2 3
7 3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    2.00    2.00    2.30    2.75    3.00

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.002058   Min.   :0.7542   Min.   : 57.0   Min.   : 84.00
 1st Qu.:0.002143   1st Qu.:0.7701   1st Qu.:144.6   1st Qu.: 87.50
 Median :0.002254   Median :0.7807   Median :158.6   Median : 92.00
 Mean   :0.002305   Mean   :0.7877   Mean   :167.8   Mean   : 94.10
 3rd Qu.:0.002395   3rd Qu.:0.8134   3rd Qu.:213.8   3rd Qu.: 97.75
 Max.   :0.002695   Max.   :0.8250   Max.   :253.2   Max.   :110.00

mining info:
        data ntransactions support confidence
 bookbsk_2up          40822   0.002       0.75
```

In the decreasing order of confidence these rules are as follow:

```
> inspect(sort(rules5, by="confidence", decreasing=TRUE)) #print rules in decreasing order of confidence
     lhs                                                rhs                                                                support confidence      lift count
[1]  {key of knowledge}                              => {key of light}                                                     0.002425163  0.8250000 253.21917    99
[2]  {hard eight : a stephanie plum novel,
      high five}                                     => {seven up}                                                         0.002131204  0.8207547 127.88110    87
[3]  {hot six : a stephanie plum novel,
      seven up}                                      => {high five}                                                        0.002106707  0.8190476 153.37230    86
[4]  {soul harvest: the world takes sides}           => {nicolae: the rise of antichrist}                                  0.002302680  0.7966102 229.00859    94
[5]  {the reptile room}                              => {the bad beginning}                                                0.002229190  0.7844828 144.25295    91
[6]  {the amber spyglass}                            => {the subtle knife}                                                 0.002645632  0.7769784 145.49455   108
[7]  {nicolae: the rise of antichrist}               => {tribulation force: the continuing drama of those left behind}     0.002694625  0.7746479 168.20572   110
[8]  {key of valor}                                  => {key of light}                                                     0.002278183  0.7685950 235.90667    93
[9]  {harry potter and the prisoner of azkaban,
      wild animus}                                   => {harry potter and the chamber of secrets}                          0.002057714  0.7567568  56.99691    84
[10] {soul harvest: the world takes sides}           => {tribulation force: the continuing drama of those left behind}     0.002180197  0.7542373 163.77380    89
>
```

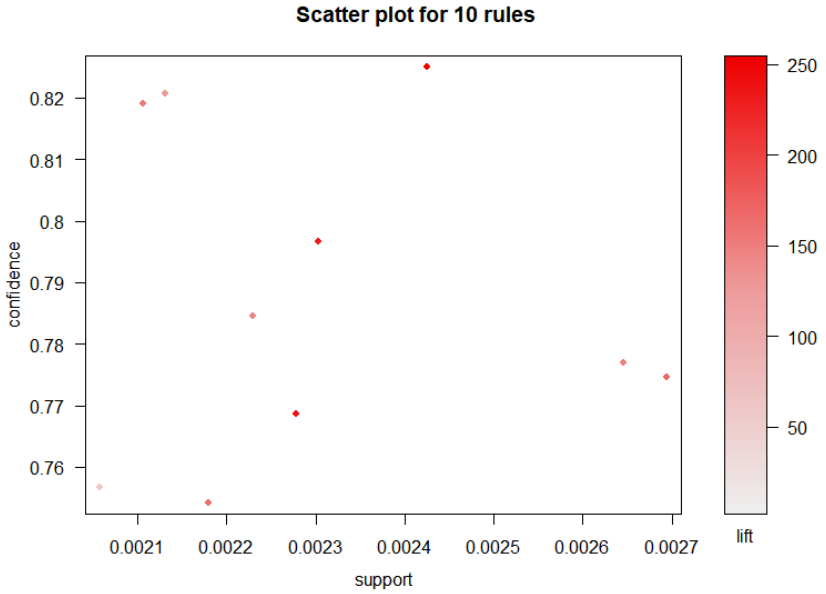
This time, the rules included books written not only by J.K. Rowling but they still mostly included books from the same series (with an exception of Wild Animus) on both left and right side of the rule.

```
> plot(rules5, method="graph", interactive=TRUE, shading="confidence")
```



Not surprisingly, the level of the confidence is relatively high for the majority of the rules:

```
> plot(rules5)
```



As I was interested in finding association rules that include books written by various authors in different styles that could be used, for example, for cross-promotion opportunities, I decided to go back to the algorithm using the minimum support of 0.001 and the minimum level of confidence of 0.75 which originally returned 3907 rules.

After removing the redundant rules, there were 563 rules left:

```
> #Remove redundant rules
> redundant1 <- which (colSums(is.subset(rules3, rules3))>1) #vector of redundant rules
> rules3a<-rules3[-redundant1] #remove redundant rules
> summary(rules3a)
set of 563 rules

rule length distribution (lhs + rhs):sizes
  2   3   4
 23 346 194

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.304   4.000   4.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001004   Min.   :0.7500   Min.   : 36.36   Min.   : 41.00
 1st Qu.:0.001053   1st Qu.:0.7636   1st Qu.: 60.02   1st Qu.: 43.00
 Median :0.001151   Median :0.7869   Median : 68.94   Median : 47.00
 Mean   :0.001215   Mean   :0.7963   Mean   :102.56   Mean   : 49.58
 3rd Qu.:0.001274   3rd Qu.:0.8192   3rd Qu.:140.06   3rd Qu.: 52.00
 Max.   :0.002695   Max.   :0.9767   Max.   :523.36   Max.   :110.00

mining info:
```

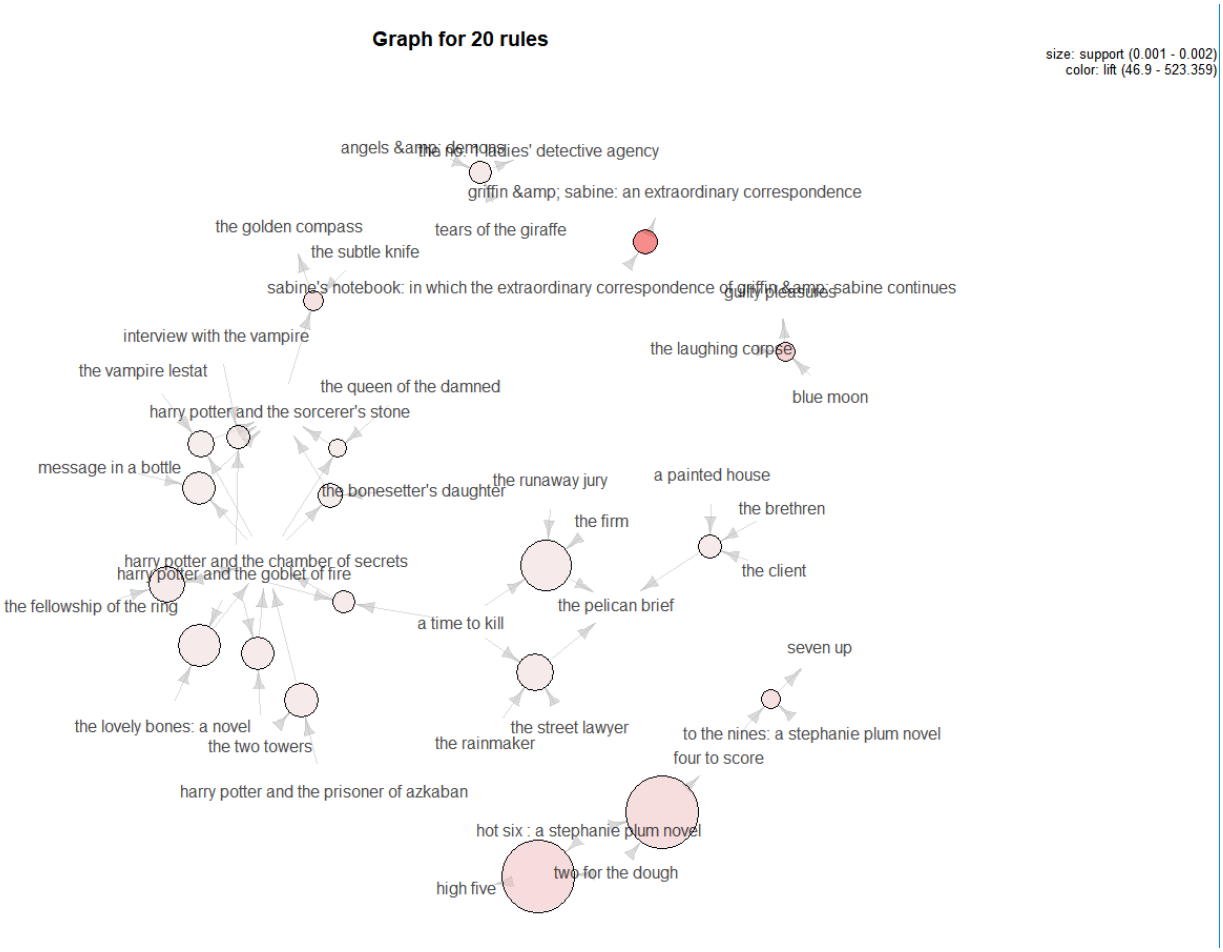
```
        data ntransactions support confidence
bookbsk_2up          40822   0.001       0.75
```

However, looking at the top 20 rules sorted by the level of confidence shows that they finally include not only books belonging to the same series:

```
> rules3b<- head(sort(rules3a, by="confidence", decreasing=TRUE), n=20) #first 20 rules in decreasing order of confidence
> inspect(rules3b)
     lhs                                                                                      rhs                                                 support confidence      lift count
[1]  {blue moon,
      the laughing corpse}                                                                 => {guilty pleasures}                               0.001028857  0.9767442 231.81774    42
[2]  {a painted house,
      the brethren,
      the client}                                                                          => {the pelican brief}                              0.001102347  0.9183673  70.20523    45
[3]  {harry potter and the sorcerer's stone,
      the subtle knife}                                                                    => {the golden compass}                             0.001053354  0.9148936 131.04487    43
[4]  {angels &amp; demons,
      tears of the giraffe}                                                                => {the no. 1 ladies' detective agency}             0.001077850  0.8979592  74.50506    44
[5]  {a time to kill,
      harry potter and the goblet of fire}                                                 => {harry potter and the chamber of secrets}        0.001077850  0.8979592  67.63190    44
[6]  {harry potter and the goblet of fire,
      the two towers}                                                                      => {harry potter and the chamber of secrets}        0.001273823  0.8965517  67.52589    52
[7]  {harry potter and the chamber of secrets,
      message in a bottle}                                                                 => {harry potter and the sorcerer's stone}          0.001273823  0.8965517  47.65499    52
[8]  {four to score,
      to the nines: a stephanie plum novel}                                                => {seven up}                                       0.001028857  0.8936170 139.23372    42
[9]  {harry potter and the chamber of secrets,
      the queen of the damned}                                                             => {harry potter and the sorcerer's stone}          0.001004360  0.8913043  47.37608    41
[10] {a time to kill,
      the firm,
      the runaway jury}                                                                    => {the pelican brief}                              0.001592279  0.8904110  68.06808    65
[11] {a time to kill,
      the rainmaker,
      the street lawyer}                                                                   => {the pelican brief}                              0.001347313  0.8870968  67.81473    55
[12] {harry potter and the chamber of secrets,
      the vampire lestat}                                                                  => {harry potter and the sorcerer's stone}          0.001151340  0.8867925  47.13625    47
[13] {harry potter and the goblet of fire,
      the fellowship of the ring}                                                          => {harry potter and the chamber of secrets}        0.001322816  0.8852459  66.67437    54
[14] {sabine's notebook: in which the extraordinary correspondence of griffin &amp; sabine continues} => {griffin &amp; sabine: an extraordinary correspondence} 0.001126843  0.8846154 523.35897    46
[15] {harry potter and the chamber of secrets,
      the bonesetter's daughter}                                                           => {harry potter and the sorcerer's stone}          0.001126843  0.8846154  47.02053    46
[16] {harry potter and the prisoner of azkaban,
      the two towers}                                                                      => {harry potter and the chamber of secrets}        0.001298320  0.8833333  66.53032    53
[17] {harry potter and the goblet of fire,
      interview with the vampire}                                                          => {harry potter and the sorcerer's stone}          0.001102347  0.8823529  46.90028    45
[18] {harry potter and the goblet of fire,
      the lovely bones: a novel}                                                           => {harry potter and the chamber of secrets}        0.001445299  0.8805970  66.32423    59
[19] {hot six : a stephanie plum novel,
      two for the dough}                                                                   => {four to score}                                  0.001984224  0.8804348 152.94089    81
[20] {hot six : a stephanie plum novel,
      two for the dough}                                                                   => {high five}                                      0.001984224  0.8804348 164.86747    81
```

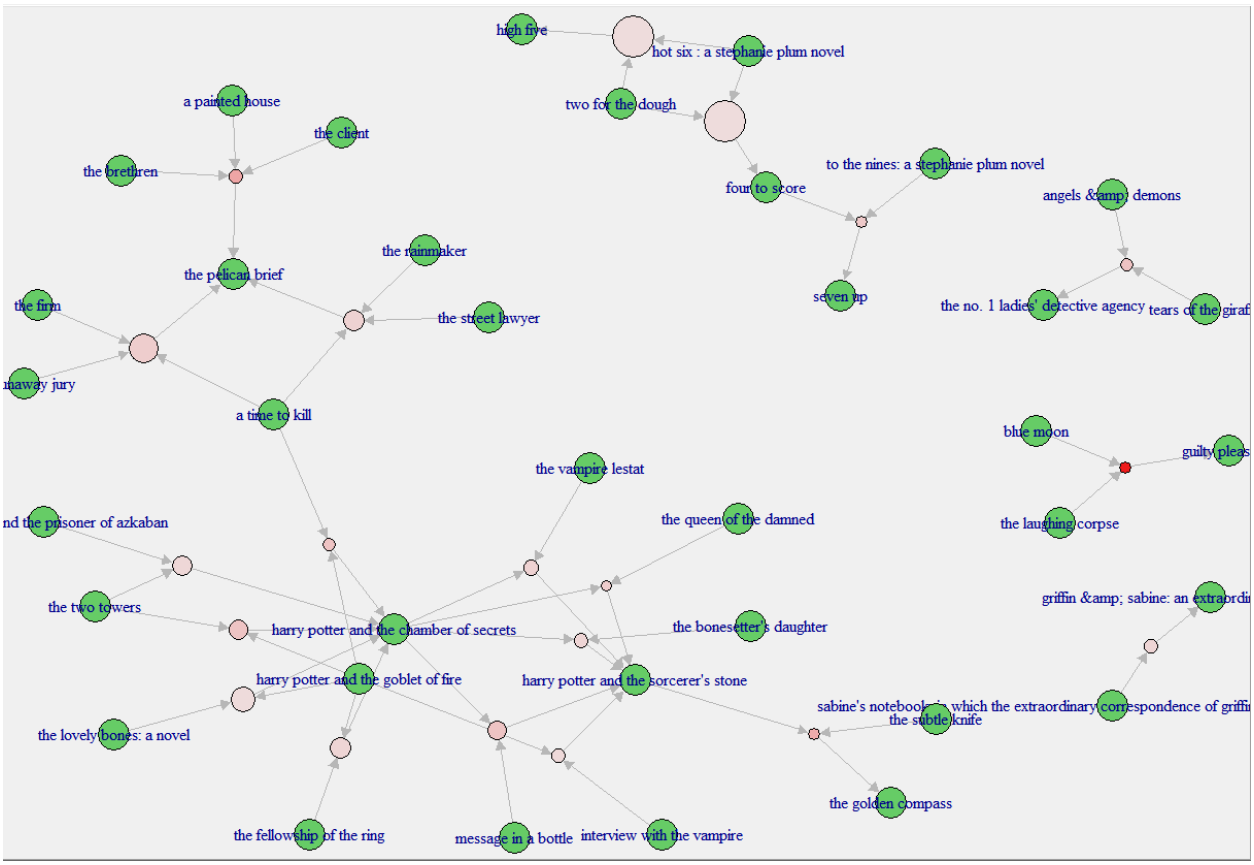
A larger number of rules is more convenient to inspect visually

```
> plot(rules3b, method="graph", control=list(type="items"))
```



Or as an interactive plot:

```
> plot(rules3b, method="graph", interactive=TRUE, shading="confidence")
```



This plot shows that users who like fantasy usually read more than one book (series) from the same genre. For example, there are rules showing that the Harry Potter fans also enjoy the Lord of the Rings novels, some fantasy adventures such as the Golden Compass, along with some books about Vampires and other supernatural drama (e.g., The Lovely Bones). This information can be used in cross-promoting new fantasy books to the fans of this genre and to organizing print advertisement materials.

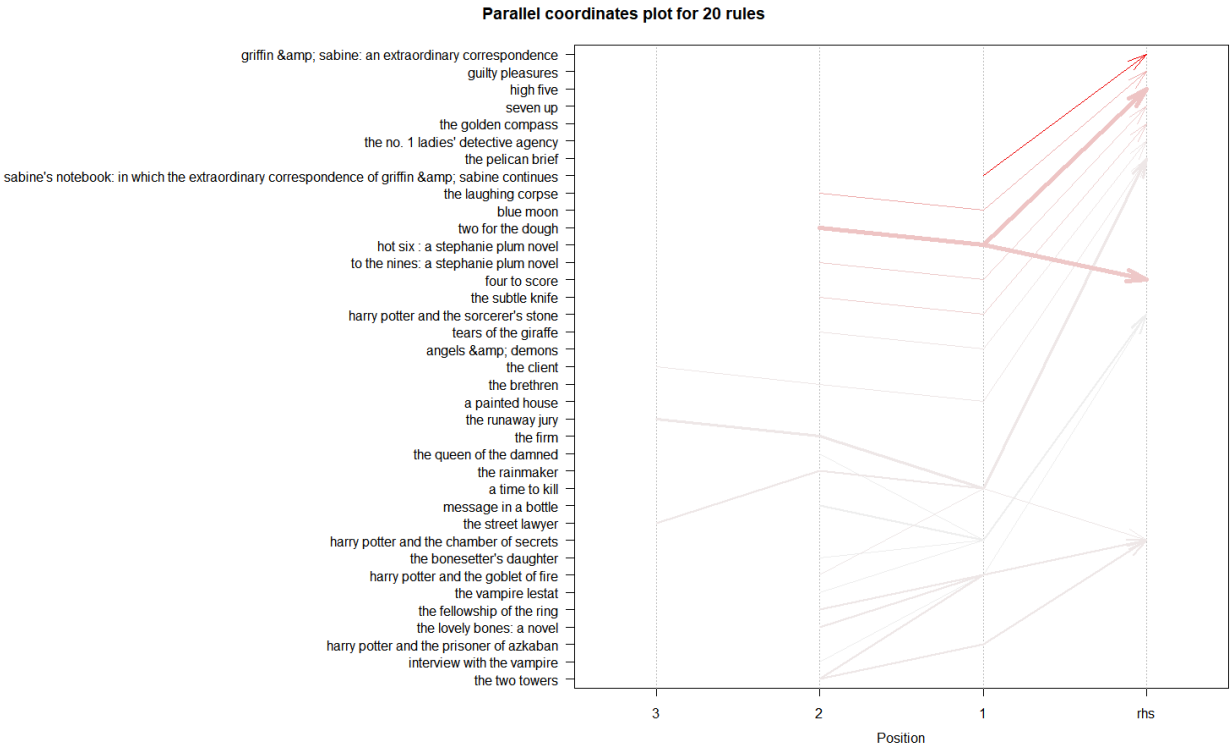
Another interesting finding is a relationship between the cluster of fantasy books and a cluster of fiction literature geared toward a little more mature audience, such as A Time to Kill, The Pelican Brief, the Firm, The Client etc. In addition, many of these books were later turned

into movies. One possible explanation could be that the same users purchase books both for themselves and for their children.

```
> plot(rules3b, method="paracoord", control = list(reorder=TRUE))
```

A parallel coordinates plot can be used to establish pairs of books for different audiences that should be promoted together:

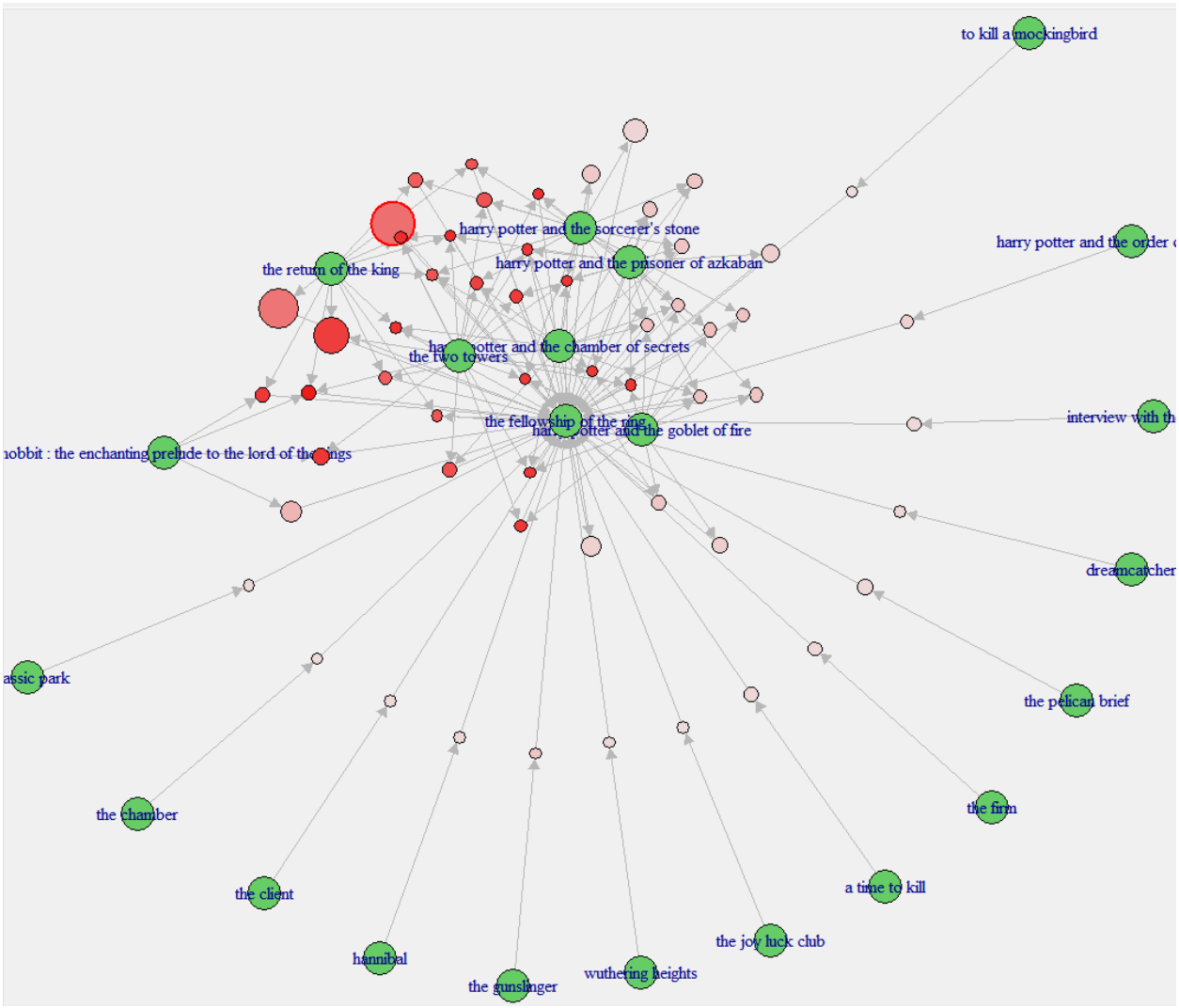
```
> plot(rules3b, method="paracoord", control = list(reorder=TRUE))
```



However, using apriori() algorithm we can find what exactly users were interested before or after reading a particular book and create a more targeted advertisement. Using the Fellowship of the Ring as an example:

```
> #what were users interested before the fellowship of the ring
> rules6 <- apriori(data=bookbsk_2up, parameter=list(supp=0.001, conf=0.08), appearance = list(de
fault="lhs", rhs="the fellowship of the ring"), control=list(verbose=FALSE))
> summary(rules6)
set of 56 rules
```

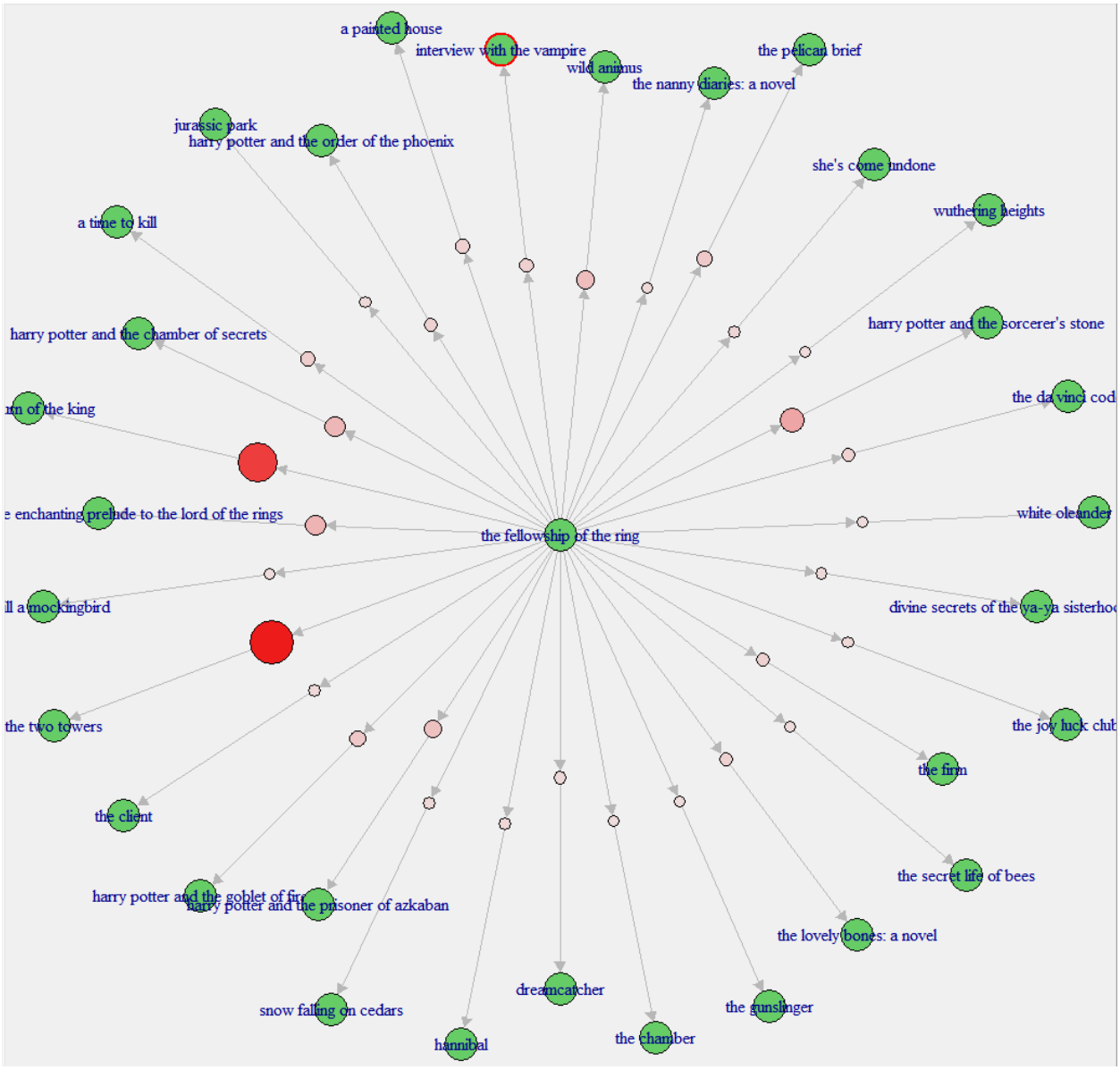
```
rule length distribution (lhs + rhs):sizes
 2  3  4  5
21 17 15  3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       3       3       4       5

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001004   Min.   :0.08255   Min.   : 8.892   Min.   : 41.00
 1st Qu.:0.001072   1st Qu.:0.15112   1st Qu.:16.277   1st Qu.: 43.75
 Median :0.001237   Median :0.25444   Median :27.406   Median : 50.50
 Mean   :0.001409   Mean   :0.45468   Mean   :48.973   Mean   : 57.54
 3rd Qu.:0.001378   3rd Qu.:0.81750   3rd Qu.:88.053   3rd Qu.: 56.25
 Max.   :0.004385   Max.   :0.91667   Max.   :98.734   Max.   :179.00

mining info:
        data ntransactions support confidence
 bookbsk_2up         40822   0.001       0.08
```



The list includes such diverse books as To Kill a Mockingbird, The Dreamcatcher and Jurassic Park so that this trend might be undetected without the association rules analysis.

```
> #what were users interested afther the fellowship of the ring
> rules7 <- apriori(data=bookbsk_2up, parameter=list(supp=0.001, conf=0.08), appearance = list(de
fault="rhs", lhs="the fellowship of the ring"), control=list(verbose=FALSE))
> summary(rules7)
set of 31 rules

rule length distribution (lhs + rhs):sizes
 2
31

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001004   Min.   :0.1082   Min.   : 3.860   Min.   : 41.00
 1st Qu.:0.001078   1st Qu.:0.1161   1st Qu.: 7.527   1st Qu.: 44.00
 Median :0.001225   Median :0.1319   Median :10.384   Median : 50.00
 Mean   :0.001478   Mean   :0.1592   Mean   :15.006   Mean   : 60.35
 3rd Qu.:0.001470   3rd Qu.:0.1583   3rd Qu.:14.582   3rd Qu.: 60.00
 Max.   :0.004385   Max.   :0.4723   Max.   :68.857   Max.   :179.00

mining info:
        data ntransactions support confidence
 bookbsk_2up         40822   0.001       0.08
```

The list of the books that users enjoy after The Fellowship of the Ring is equally diverse, as demonstrated by the graph below:



Overall, apriori() algorithm was very useful in detecting trends in book preferences that could not be noticed otherwise. However, it requires substantial data and computational resources and the results are highly sensitive to the chosen parameters (see table below). As a result, finding hidden rules is a process that might require several iterations and parameter adjustments.

Apriori() parameters and the number of rules detected

Support = 0.005 confidence = 0.70	support = 0.001 confidence = 0.70	support = 0.001 confidence = 0.75	support = 0.002 confidence = 0.75
7 rules	5108 rules	3907	191