

UpSetR: A novel R package for visualizing intersecting sets and their attributes

Harvard Medical School, Department of Biomedical Informatics
 Jake Conway & Nils Gehlenborg, PhD



Abstract

Understanding the relationships and interactions between data is imperative in bioinformatics. With copious amounts of biological data being generated today, there is high demand for tools that can visualize large numbers of sets and their intersections. Here we present *UpSetR*, a novel R package that visualizes large numbers of sets and intersections via a matrix-based technique. Along with visualizing data on a set-based level, *UpSetR* allows users to explore and visualize their data on an element-based level through the implementation of queries and attribute plots. Through its seamless integration with *ggplot2* and the ability to apply virtually any query to the data, *UpSetR* is an extremely powerful tool for data exploration and producing publication quality visualizations.

Approach

- UpSetR* was implemented using *ggplot2*, and arranged using a grid layout (100 x 100)
- Built in intersection and element queries to target points of interest.
 - User can create their own query to operate on rows of data

```
Myfunc <- function(row, release, rating){
  data <- (row["ReleaseDate"] %in% release) &
  (row["AvgRating"] > rating)
}
```

- Built in histogram and scatter plot functions to display element based data.
 - User can create their own plot to display the data

```
myplot <- function(mydata,x,y){
  (ggplot(data = mydata, aes_string(x=x, y=y, colour = "color")) +
  geom_point() + scale_color_identity())
  + theme(plot.margin = unit(c(0,0,0,0), "cm"))
}
```

Related Works and Future Features

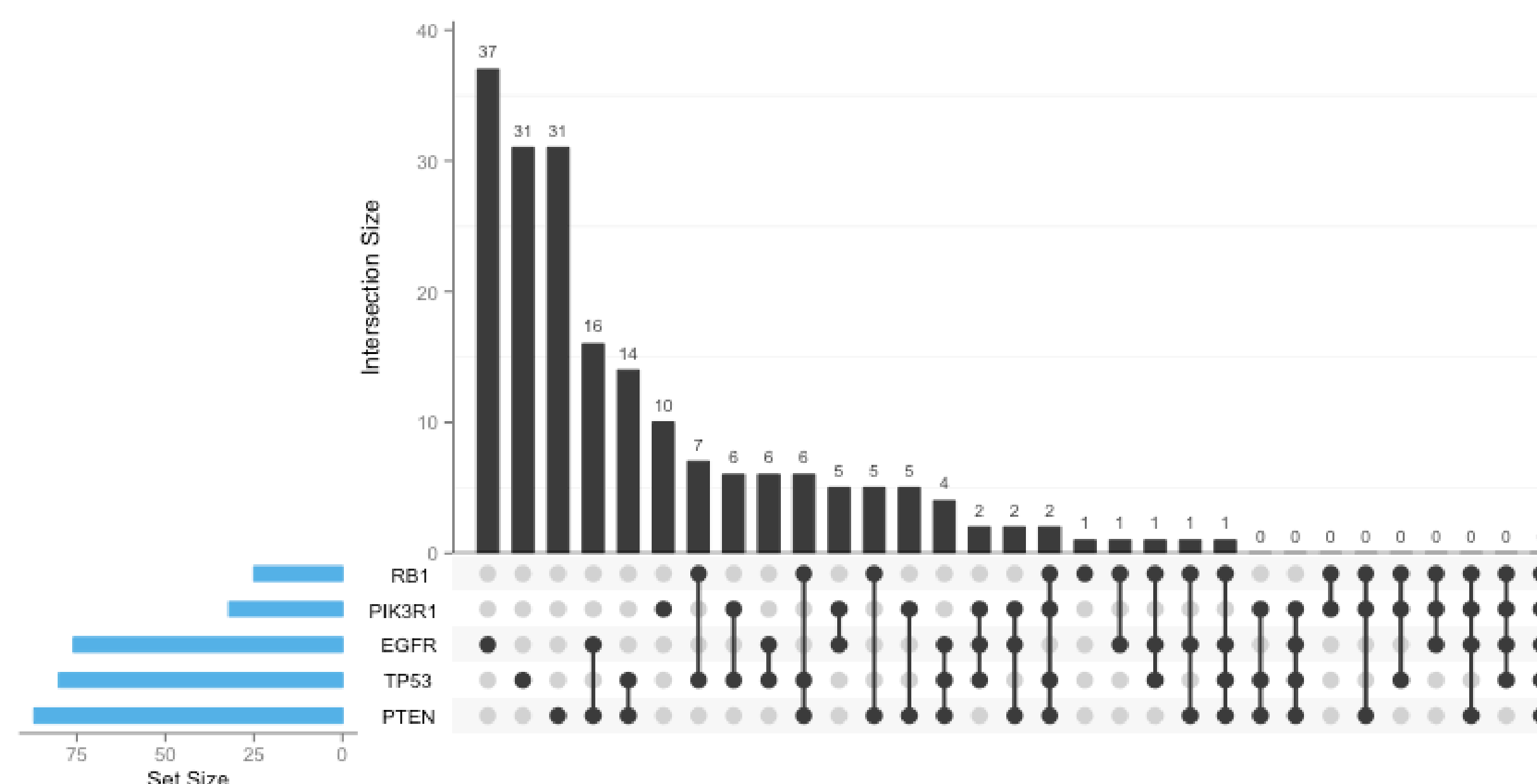
- There is a Shiny application online that supports matrix portion of *UpSetR*
 - <https://upsetr.shinyapps.io/UpSetR-shiny>
- In a future version of the application, queries and attribute plots will be introduced.
- Currently working on a Shiny app linked to Ensembl gene database.
- Implementation of summary statistics for sets, intersections, and attributes in a future release
- UpSetR* source code can be found at:
 - <https://github.com/hms-dbmi/UpSetR>
- An interactive version of the original UpSet and its paper can be found at: <http://vcg.github.io/upset/about/>
- Implementation of *UpSetR* inspired by:
 - A. Lex & N. Gehlenborg. Points of view: Sets and intersections. *Nature Methods* 11, 779 (2014)

UpSetR

An example of a correctly formatted data set for *UpSetR*.

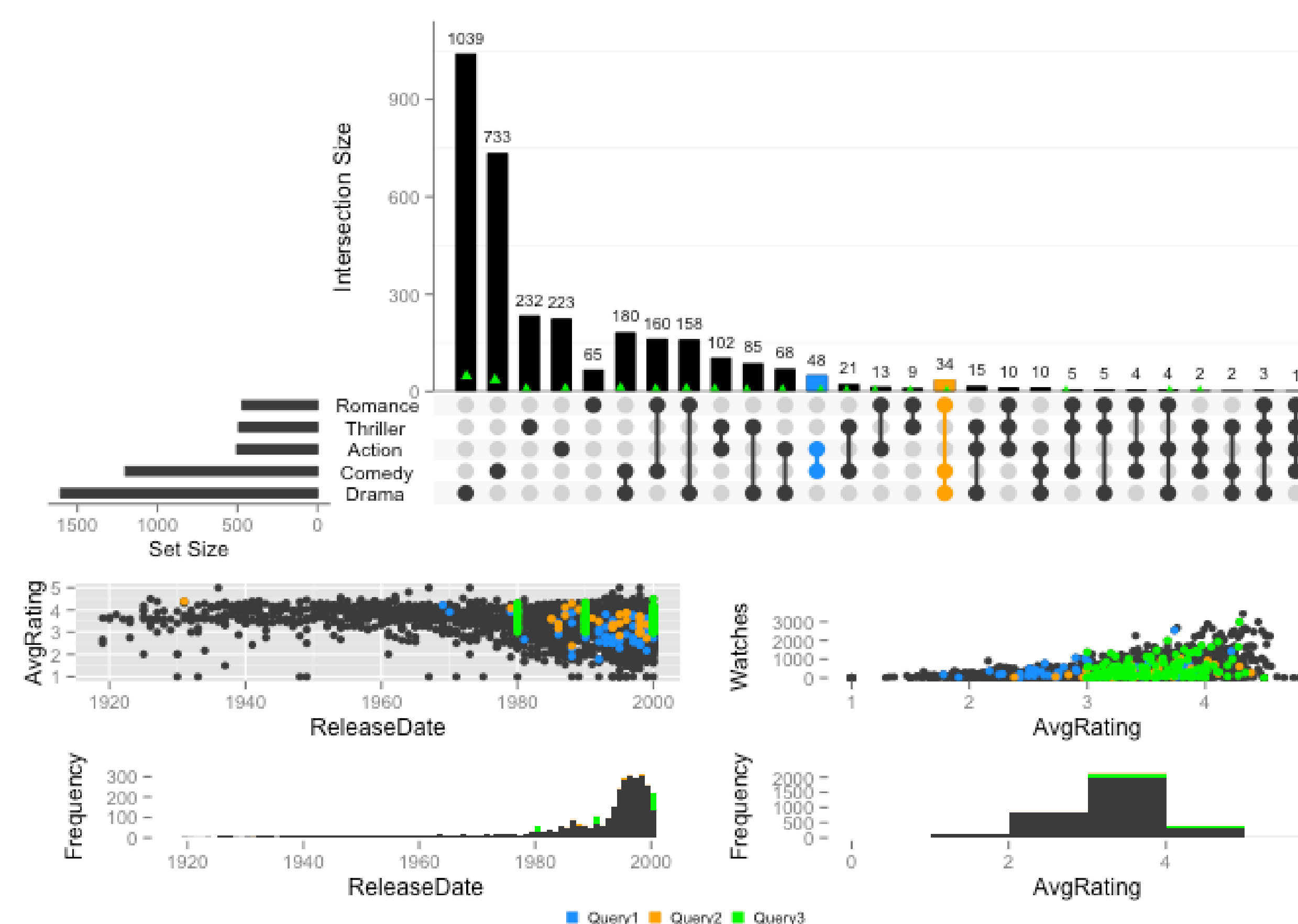
	Name	ReleaseDate	Action	Comedy	Drama	Romance	Thriller	AvgRating	Watches
1	Toy Story (1995)	1995	0	1	0	0	0	4.15	2077
2	Jumanji (1995)	1995	0	0	0	0	0	3.20	701
3	Grumpier Old Men (1995)	1995	0	1	0	1	0	3.02	478
4	Waiting to Exhale (1995)	1995	0	1	1	0	0	2.73	170
5	Father of the Bride Part II (1995)	1995	0	1	0	0	0	3.01	296
6	Heat (1995)	1995	1	0	0	0	1	3.88	940
7	Sabrina (1995)	1995	0	1	0	1	0	3.41	458
8	Tom and Huck (1995)	1995	0	0	0	0	0	3.01	68
9	Sudden Death (1995)	1995	1	0	0	0	0	2.66	102
10	GoldenEye (1995)	1995	1	0	0	0	1	3.54	888

UpSetR visualizing the intersections of the 5 most common gene mutations in a glioblastoma multiforme cohort.



```
upset(mutations, sets = c("PTEN", "TP53", "EGFR", "PIK3R1", "RB1"), sets.bar.color = "#56B4E9", order.matrix = "freq", empty.intersections = "on")
```

UpSetR visualizing the intersections of the 5 most common movie genres, along with 3 queries and 4 attribute plots displaying the relationships of elements in each query.



```
upset(movies, main.bar.color = "black", queries = list(list(query = intersects, params = list("Action", "Comedy"), color = "dodgerblue", active = T), list(query = intersects, params = list("Romance", "Comedy", "Drama"), color = "orange", active = T), list(query = myfunc, params = list(c(1980,1990, 2000), 3), color = "green")), attribute.plots = list(gridrows = 55, plots = list(list(plot = myplot, x = "ReleaseDate", y = "AvgRating", queries = T), list(plot = scatter_plot, x = "AvgRating", y = "Watches", queries = T), list(plot = histogram, x = "ReleaseDate", queries = T), list(plot = histogram, x = "AvgRating", queries = T))))
```

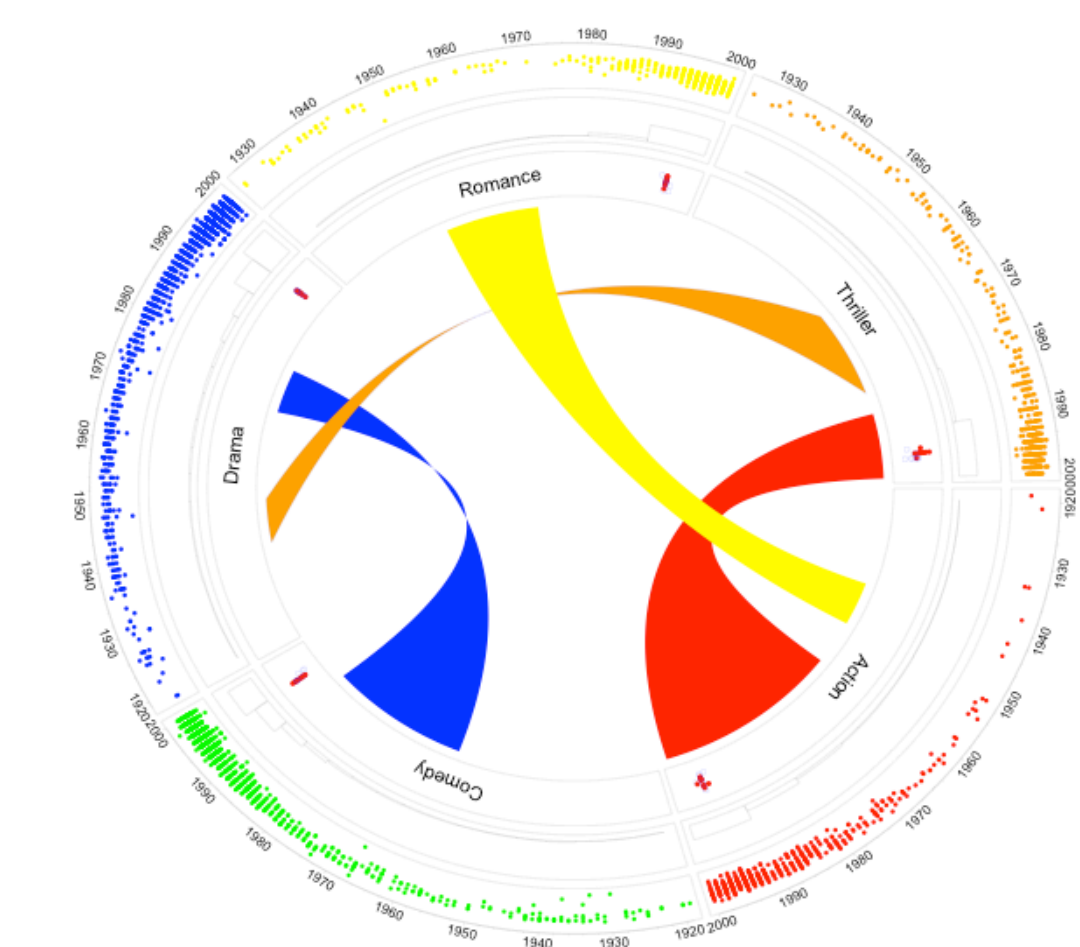
Other Set Visualization R Packages

venneuler :



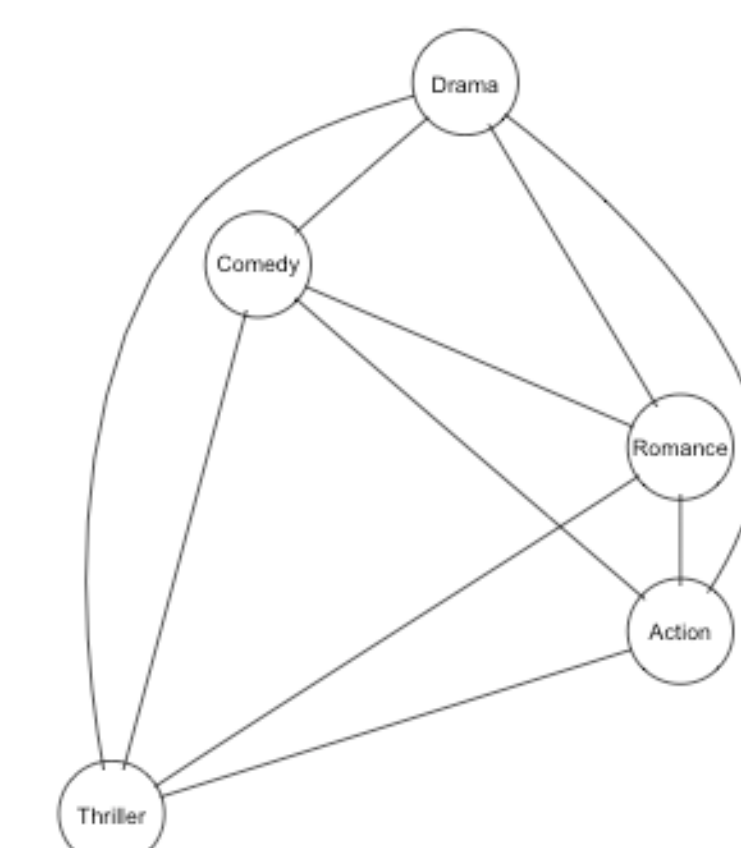
Hanbo, C. et al. VennDiagram: A Package for the Generation of Highly-customizable Venn and Euler Diagrams in R. *BMC Bioinformatics* 12(1): 35, 2011.

circlize :



Gu, Z. et al. circlize implements and enhances circular visualization in R. *Bioinformatics* 30(19):2811-2812, 2014.

igraph/Rgraphviz :



<https://cran.r-project.org/web/packages/igraph/igraph.pdf>

Shortcomings Compared to UpSetR

- Becomes difficult to interpret when large number of intersections present. In *UpSetR* each intersection is represented as a column in the matrix.
- Need to know what the intersection sizes are prior to plotting.
- Need to enter intersections manually. *UpSetR* only needs a correctly formatted data set.
- Circlize* can only display up to 2 way interactions.
- Igraph/Rgraphviz* only show which sets interact. Can't display intersections.
- Don't allow for queries or additional plots of element data.