J. Beyersmann, P. Gastmeier, H. Grundmann, S. Bärwolff, C. Geffers, M. Behnke, H. Rüden, and M. Schumacher. Use of Multistate Models to Assess Prolongation of Intensive Care Unit Stay Due to Nosocomial Infection. *Infection Control and Hospital Epidemiology*, 27:493-499, 2006.

D. Glidden. Robust innference for event probabilities with non-Markov data. *Biometrics*, 58:361–368, 2002.

H. Grundmann, S. Bärwolff, F. Schwab, A. Tami, M. Behnke, C. Geffers, E. Halle, U. Göbel, R. Schiller, D. Jonas, I. Klare, K. Weist, W. Witte, K. Beck-Beilecke, M. Schumacher, H. Rüden, and P. Gastmeier. How many infections are caused by patient-to-patient transmission in intensive care units? *Critical Care Medicine (in press)*, 2005.

G. Schulgen, A. Kropec, I. Kappstein, F. Daschner, and M. Schumacher. Estimation of extra hospital stay attributable to nosocomial infections: heterogeneity and timing of events. *Journal of Clinical Epidemiology*, 53:409–417, 2000.

G. Schulgen and M. Schumacher. Estimation of prolongation of hospital stay attributable to nosocomial infections. *Lifetime Data Analysis*, 2:219–240, 1996.

Matthias Wangler
mw@imbi.uni-freiburg.de

Balloon Plot

Graphical tool for displaying tabular data

by Nitin Jain and Gregory R. Warnes

Introduction

Numeric data is often summarized using rectangular tables. While these tables allow presentation of all of the relevant data, they do not lend themselves to rapid discovery of important patterns. The primary difficulty is that the visual impact of numeric values is not proportional to the scale of the numbers represented.

We have developed a new graphical tool, the “balloonplot”, which augments the numeric values in tables with colored circles with area proportional to the size of the corresponding table entry. This visually highlights the prominent features of data, while preserving the details conveyed by the numeric values themselves.

In this article, we describe the balloonplot, as implemented by the `balloonplot` function in the `gplots` package, and describe the features of our implementation. We then provide an example using the “Titanic” passenger survival data. We conclude with some observations on the balloonplot relative to the previously developed “mosaic plot”.

Function description

The `balloonplot` function accepts a table (to be displayed as found) or lists of vectors for x (column category), y (row category) and z (data value) from which a table will be constructed.

The `balloonplot` function creates a graphical table where each cell displays the appropriate numeric value plus a colored circle whose size reflects the relative magnitude of the corresponding component. The *area* of each circle is proportional to the frequency of data. (The circles are scaled so that the circle for largest value fills the available space in the cell.)

As a consequence, the largest values in the table are “spotlighted” by the biggest circles, while smaller values are displayed with smaller circles. Of course, circles can only have positive radius, so the radius of circles for cells with negative values are set to zero. (A warning is issued when this “truncation” occurs.)

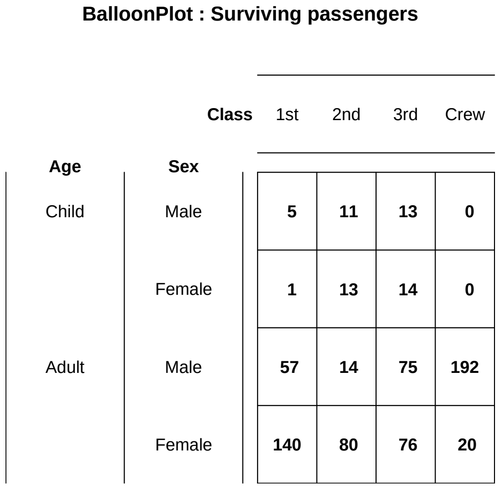
Of course, when labels are present on the table or provided to the function, the graphical table is appropriately labeled. In addition, options are provided to allow control of various visual features of the plot:

- rotation of the row and column headers
- balloon color and shape (globally or individually)
- number of displayed digits
- display of entries with zero values
- display of marginal totals
- display of cumulative histograms
- x- and y-axes group sorting
- formatting of row and column labels
- traditional graphics parameters (title, background, etc.)

Example using the `Titanic` data set

For illustration purposes, we use the Titanic data set from the datasets package. Titanic provides survival status for passengers on the tragic maiden voyage of the ocean liner “Titanic”, summarized according to economic status (class), sex, and age.

Typically, the number of surviving passengers are shown in a tabular form, such as shown in Figure 1.



BalloonPlot : Surviving passengers

	Class	1st	2nd	3rd	Crew
Age	Sex				
Child	Male	5	11	13	0
	Female	1	13	14	0
Adult	Male	57	14	75	192
	Female	140	80	76	20

Figure 1: Tabular representation of survived population by gender and age

Figure 1 was created by calling balloonplot with the balloon color set to match the background color and most options disabled. Note that one must actively focus on the individual cell values in order to see any pattern in the data.

Now, we redraw the table with light-blue circles (‘balloons’) superimposed over the numerical values (Figure 2). This is accomplished using the code:

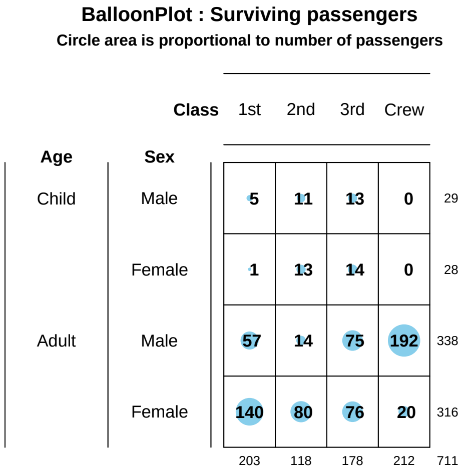
```
library(gplots)
data(Titanic)

# Convert to 1 entry per row format
dframe <- as.data.frame(Titanic)

# Select only surviving passengers
survived <- dframe[dframe$Survived=="Yes",]
attach(survived)

balloonplot(x=Class,
            y=list(Age, Sex),
            z=Freq,
            sort=TRUE,
            show.zeros=TRUE,
            cum.margins=FALSE,
            main=
            "BalloonPlot : Surviving passengers"
            )
title(main=list("Circle area is proportional to\
 number of passengers",
           cex=0.9),
      line=0.5)

detach(survived)
```



BalloonPlot : Surviving passengers
Circle area is proportional to number of passengers

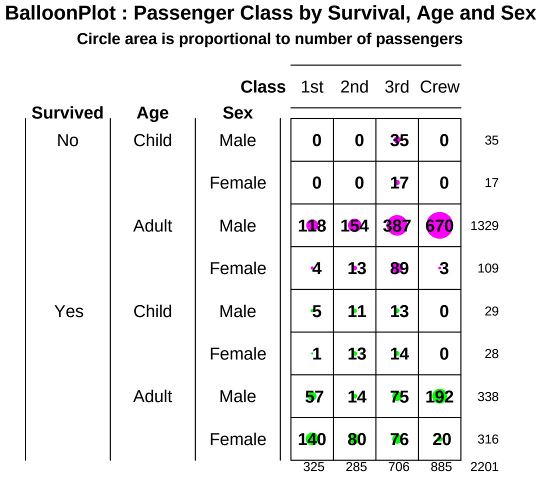
	Class	1st	2nd	3rd	Crew	
Age	Sex					
Child	Male	5	11	13	0	29
	Female	1	13	14	0	28
Adult	Male	57	14	75	192	338
	Female	140	80	76	20	316
		203	118	178	212	711

Figure 2: Balloon plot of surviving individuals by class, gender and age

With the addition of the blue “spotlights”, whose area is proportional to the magnitude of the data value, it is easy to see that only adult females and adult male crew members survived in large numbers. Also note the addition of row and column marginal totals.

Of course, the number of surviving passengers is only half of the story. We could create a similar plot showing the number of passengers who did not survive. Alternatively, we can simply add survival status as another variable to the display, setting the color of the circles to green for passengers who survived, and magenta for those who did not (Figure 3). This conveys considerably more information than Figures 1 and 2 without substantial loss of clarity. The large magenta circles make it clear that most passengers did not survive.

To further improve the display, we add a visual representation of the row and column sums (Figure 4). This is accomplished using light grey bars behind the row and column headers. The length of each bar is proportional to the corresponding sum, allowing rapid visual ascertainment of their relative sizes. We have also added appropriately colored markers adjacent to the headers under “Survived” to emphasize the meaning of each color.



BalloonPlot : Passenger Class by Survival, Age and Sex

Circle area is proportional to number of passengers

Survived	Age	Sex	1st	2nd	3rd	Crew	
No	Child	Male	0	0	35	0	35
		Female	0	0	17	0	17
	Adult	Male	118	154	387	670	1329
		Female	4	13	89	3	109
Yes	Child	Male	5	11	13	0	29
		Female	1	13	14	0	28
	Adult	Male	57	14	75	192	338
		Female	140	80	76	20	316
			325	285	706	885	2201

Figure 3: Balloon plot of Titanic passengers by gender, age and class. Green circles represent passengers who survived and magenta circles represent the passengers who did not survive.

```
attach(dframe)
colors <- ifelse( Survived=="Yes", "green",
                  "magenta")

balloonplot(x=Class,
            y=list(Survived, Age, Sex),
            z=Freq,
            sort=FALSE,
            dotcol=colors,
            show.zeros=TRUE,
            main="BalloonPlot : Passenger Class \
by Survival, Age and Sex")

points( x=1, y=8, pch=20, col="magenta")
points( x=1, y=4, pch=20, col="green")

title(main=list("Circle area is proportional to \
number of passengers", cex=0.9), line=0.5)

detach(dframe)
```

It is now easy to see several facts:

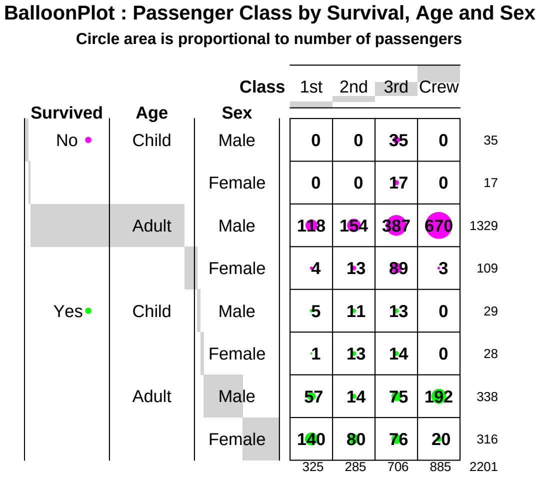
- A surprisingly large fraction (885/2201) of passengers were crew members

- Most passengers and crew were adult males

- Most adult males perished

- Most women survived, except in 3^{rd} class

- Only 3rd class children perished



BalloonPlot : Passenger Class by Survival, Age and Sex

Circle area is proportional to number of passengers

Survived	Age	Sex	1st	2nd	3rd	Crew	
No ●	Child	Male	0	0	35	0	35
		Female	0	0	17	0	17
	Adult	Male	118	154	387	670	1329
		Female	4	13	89	3	109
Yes●	Child	Male	5	11	13	0	29
		Female	1	13	14	0	28
	Adult	Male	57	14	75	192	338
		Female	140	80	76	20	316
			325	285	706	885	2201

Figure 4: Balloon plot of all the passengers of Titanic, stratified by survival, age, sex and class

Perhaps the most striking fact is that survival is lowest among 3^{rd} class passengers for all age and gender groups. It turns out that there is a well known reason for this difference in survival. Passengers in 1^{st} and 2^{nd} class, as well as crew members, had better access to the lifeboats. Since there were too few lifeboats for the number of passengers and crew, most women and children among the 1^{st} class, 2^{nd} class and crew found space in a lifeboat, while many of the later arriving 3^{rd} class women and children were too late: the lifeboats had already been filled and had moved away from the quickly sinking ship.

Discussion

Our goals in developing the balloonplot were twofold: First, to improve ability of viewers to quickly perceive trends. Second, to minimize the need for viewers to learn new idioms. With these goals in mind, we have restricted ourselves to simple modifications of the standard tabular display.

Other researchers have pursued more general approaches to the visual display of tabular data. (For a review of that work, see Hartigan and Kleiner (Hartigan and Kleiner, 1981) or Friendly (Friendly, 1992).) One of the most popular methods developed by these researchers is the `mosaic plot` (Snee, 1974).

We have previously experimented with mosaic plots. Unfortunately, we found that they do not lend themselves to rapid ascertainment of trends, particularly by untrained users. Even trained users find that they must pay careful attention in order to decode the visual information presented by the mosaic plot. In contrast, balloonplots lend themselves to very quick perception of important trends, even for users who have never encountered them before.

While there are, of course, tasks for which mosaic plot is preferable, we feel that the balloonplot serves admirably to allow high-levels patterns to be quickly perceived by untrained users.

Conclusion

Using the well worn Titanic data, we have shown how balloonplots help to convey important aspects of tabular data, without obscuring the exact numeric values. We hope that this new approach to visualizing tabular data will assist other statisticians in more effectively understanding and presenting tabular data.

We wish to thank *Ramon Alonso-Allende* allende@cnb.uam.es for the discussion on R-help which lead to the development of balloonplot, as well as for the code for displaying the row and column sums.

Bibliography

M. Friendly. Graphical methods for categorical data. *Proceedings of SAS SUGI 17 Conference*, 1992.

J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. New York: Springer-Verlag*, 1981.

R. D. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28:9–12, 1974.

Gregory R. Warnes, Pfizer Inc., USA
gregory.r.warnes@pfizer.com
Nitin Jain, Smith Hanley Inc, USA
nitin.jain@pfizer.com

Drawing pedigree diagrams with R and graphviz

by Jing Hua Zhao

Human genetic studies often involve data collected from families and graphical display of them is useful. The wide interest in such data over years has led to many software packages, both commercial and noncommercial. A recent account of these packages is available (Dudbridge et al., 2004), and a very flexible package Madeline (http://eyegene.ophthy.med.umich.edu/madeline/index.html) is now released under the GNU General Public License. A comprehensive list of many packages, including the package LINKAGE (Terwilliger and Ott, 1994) for human parametric linkage analysis and GAS (Genetic Analysis System, http://users.ox.ac.uk/~ayoung/gas.html) for some other analyses, can be seen at the linkage server at Rockefeller University (http://linkage.rockefeller.edu).

Here I describe two functions in R that are able to draw pedigree diagrams; the first being plot.pedigree in kinship developed for S-PLUS by Terry Therneau and Beth Atkinson and ported to R by the author, and the second pedtodot in gap based on David Duffy's gawk script (http://www2.qimr.edu.au/davidD/Course/pedtodot) that requires graphviz (http://www.graphviz.org). Both are easy to use and can draw many pedigree diagrams quickly to a single file, therefore can serve as alternatives to some programs that only offer interactive use.

Representation of pedigrees

The key elements to store pedigrees using a database is via the so-called family trios each containing individual's, father's and mother's IDs. Founders, namely individuals whose parents are not in the pedigree, are set to be zero or missing. Individual's gender (e.g. 1=male, 2=female) is included as auxiliary information, together with pedigree ID in order to maintain multiple pedigrees in a single database, each record of which indicates a node in the pedigree graph.

For instance, information for pedigree numbered 10081 in genetic analysis workshop 14 (GAW14, http://www.gaworkshop.org) is shown as follows.

```
pid id father mother sex affected
10081  1  2  3   2 2
10081  2  0  0   1 1
10081  3  0  0   2 2
10081  4  2  3   2 2
10081  5  2  3   2 1
10081  6  2  3   1 1
10081  7  2  3   2 1
10081  8  0  0   1 1
10081  9  8  4   1 1
10081 10  0  0   2 1
10081 11  2 10   2 1
10081 12  2 10   2 2
```