References

Anderson, James P. (1980), "Computer security threat monitoring and surveillance," Technical report, James P. Anderson Co., Fort Washington, PA, April 1980.

Computer Immune Systems,
www.cs.unm.edu/~forrest/
(accessed June 23, 1998)

COAST project,
www.cs.purdue.edu/coast/
(accessed June 23, 1998)

DuMouchel, W. and Schonlau, M. (1998), "A comparison of test statistics for computer intrusion detection based on principal components regression of transition probabilities," In *Proceedings of the 30th Symposium on the Interface: Computing Science and Statistics*, (to appear).

Emerald,
phlox.csl.sri.com/emerald/
(accessed June 23, 1998)

Intrusion Detection for Large Networks,
seclab.cs.ucdavis.edu/arpa/
(accessed June 23, 1998)

Netranger,
www.wheelgroup.com/netrangr/1netrang.html
(accessed June 23, 1998)

Martin Theus
AT&T Labs-Research
theus@research.att.com

Matthias Schonlau
AT&T Labs-Research and
National Institute of
Statistical Sciences
schonlau@research.att.com

NEW SOFTWARE TOOLS

Mosaic Displays in S-PLUS: A General Implementation and a Case Study.

By John W. Emerson

Introduction

Hartigan and Kleiner (1981) introduced the mosaic as a graphical method for displaying counts in a contingency table. Later, they defined a mosaic as "a graphical display of cross-classified data in which each count is represented by a rectangle of area proportional to the count" (Hartigan and Kleiner 1984). Mosaics have been implemented in SAS (see Friendly 1992) as a graphical tool for fitting log-linear models. Interactive mosaic plots (see Theus 1997a, b) have been implemented in Java. A third implementation is available in MANET, a data-visualization software package specifically for the Macintosh. No general implementation has been available in S-PLUS, one of the most popular statistical packages.

The implementation presented in this article, while lacking the modelling features of Friendly's SAS implementation, provides a simply specified function for mosaics displaying the joint distribution of any number of categorical variables. As an illustration, this article examines patterns in television viewer data. A four-way table of 825 $(5 \times 11 \times 5 \times 3)$ cells represents Nielsen television ratings (number of viewers) broken down by day, time, network, and switching behavior (changing channels, turning the television off, or staying with the current channel) for the week starting November 6, 1995. Simple patterns in the data appearing in the mosaic support intuitive explanations of viewer behavior.

The Data

Nielsen Media Research maintains a sample of over 5,000 households nationwide, installing a Nielsen People Meter (NPM) for each television set in the household. The sample is designed to reflect the demographic composition of viewers nationwide, and uses 1990 Census data to achieve the desired result. Nielsen summarizes the stream of minute-by-minute measurements to provide quarter-hour viewing measurements (defined as the channel being watched at the midpoint of each

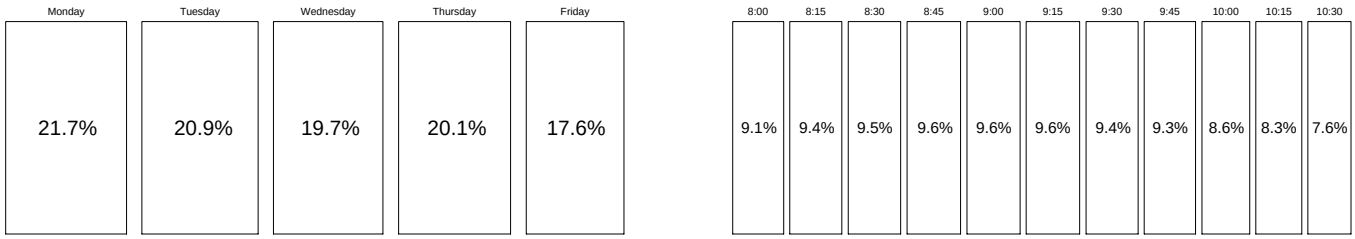


Figure 1. (a) Mosaic of the week's aggregate audience by day (lefthand panel); and **(b)** mosaic of the week's aggregate audience by time (righthand panel).

quarter-hour block) for each viewer in the sample. (Details are presented in Nielsen's *National Reference Supplement* 1995.)

A "TV guide" of the prime-time programming of the four major networks (ABC, CBS, NBC, and FOX) for the weekdays starting Monday, November 6, 1995 appears in Figure 2. During any quarter-hour, the individual is observed watching a major network channel, a non-network channel, or not watching television. At 10:00 however, FOX ends its network programming, so Nielsen does not record individuals watching FOX after 10:00. I confine this study to a subset consisting of 6307 East coast viewers in 2328 households.

Creating Mosaics in S-PLUS

Friendly (1994) describes the complete algorithm used to construct a mosaic for a general four-way table, alternatively dividing horizontal and vertical strips of area into tiles of area proportional to the counts in the remaining sub-contingency table. Without repeating the description of a general mosaic display, I note the important features of my S-PLUS implementation, which help explore various aspects of any cross-classified data set:

- Any number of categorical variables may be included in the mosaic, though in practice even a five-way table may be sufficiently complicated to defy explanation.
- Empty cells of the contingency table are represented (where possible) by a dashed line segment.
- The order in which the variables are represented may be specified, allowing simple exploration of any marginal or conditional frequencies on any subset of variables without physically manipulating the raw contingency table itself.
- The direction (horizontal or vertical) used in dividing the mosaic by each variable may be specified, allowing more flexibility than the traditional alternating divisions.
- Shading of the tiles resulting from the inclusion of the final variable in the mosaic may be specified, if desired. The amount of space separating the tiles at each level of the mosaic may also be customized.

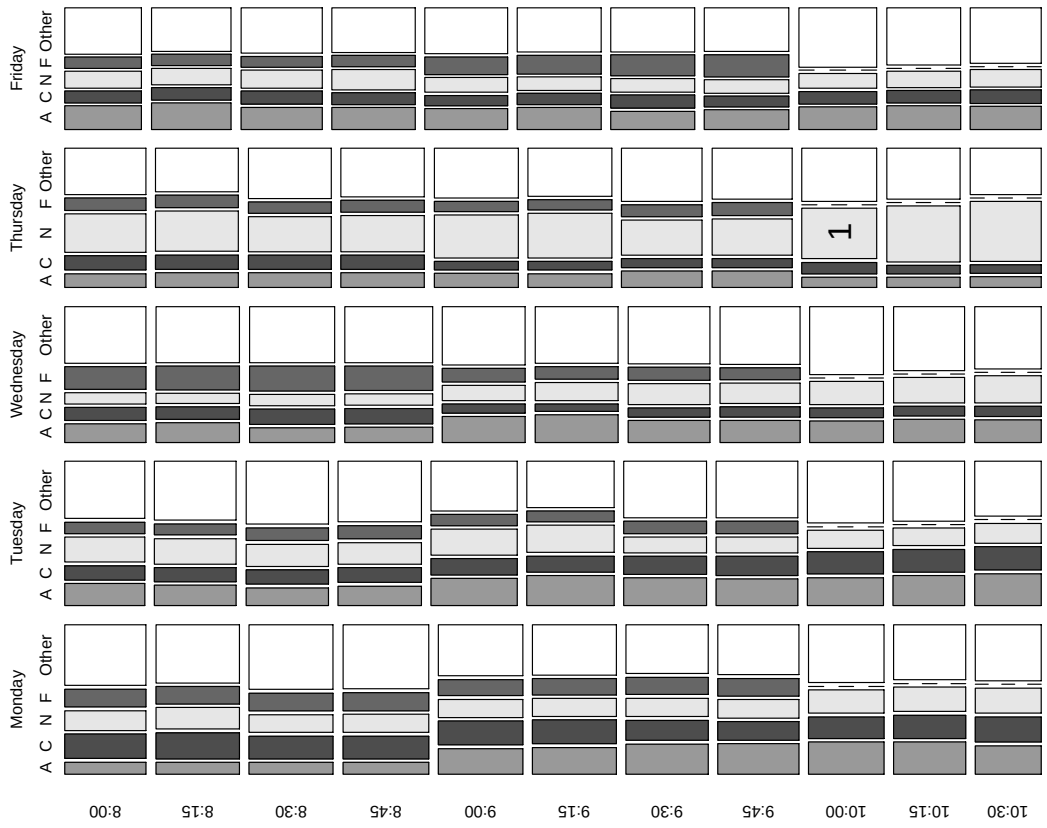
The documentation and S-PLUS code are available – details are provided at the end of the article. The basic algorithm, an efficient recursive procedure, proceeds as follows:

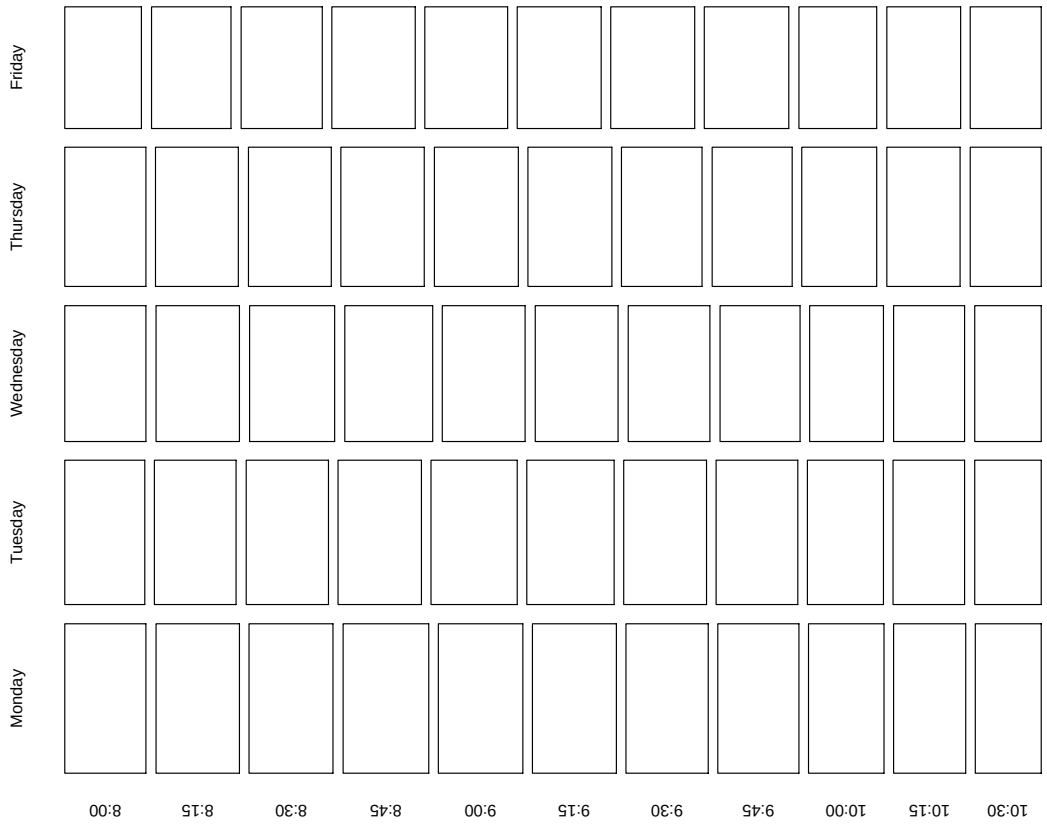
1. Initialize the parameters and the graphics device – the lower left and upper right corners of the plot area are (x_1, y_1) and (x_2, y_2). The term "parameters" refers to a collection of counts from the contingency table, labels, and values associated with features discussed above.
2. Call the recursive function
   mosaic.cell$((x_1, y_1), (x_2, y_2),$ all parameters$)$.
3. Recursive function
   mosaic.cell$((a_1, b_1), (a_2, b_2),$
   selected parameters for the current tile$)$:
   (a) Divide the current tile, given by (a_1, b_1) and (a_2, b_2), into sub-tiles, taking into account the spacing and split direction arguments of the parameters.
   (b) Add labels if the current variable is one of the first two divisions of the axis.
   (c) If this division corresponds to the last variable of the contingency table, draw the sub-tiles. Otherwise, call *mosaic.cell*() once for each of the current sub-tiles, with the appropriate sub-tile coordinates and subsets of the current parameters.

Results: Television Viewer Behavior

Simple mosaics dividing the week's aggregate audience by day and time are presented in Figures 1a and b, respectively. Though they serve the same purpose as histograms, their tile areas are more difficult to compare than the tile heights in histograms. The advantage of mosaics does not appear until at least two categorical

Figure 2. TV Guide, 11/6/95 – 11/10/95.


Friday
Thursday
Wednesday
Tuesday
Monday
A C N F Other
1
8:00
8:15
8:30
8:45
9:00
9:15
9:30
9:45
10:00
10:15
10:30


Friday
Thursday
Wednesday
Tuesday
Monday
8:00
8:15
8:30
8:45
9:00
9:15
9:30
9:45
10:00
10:15
10:30

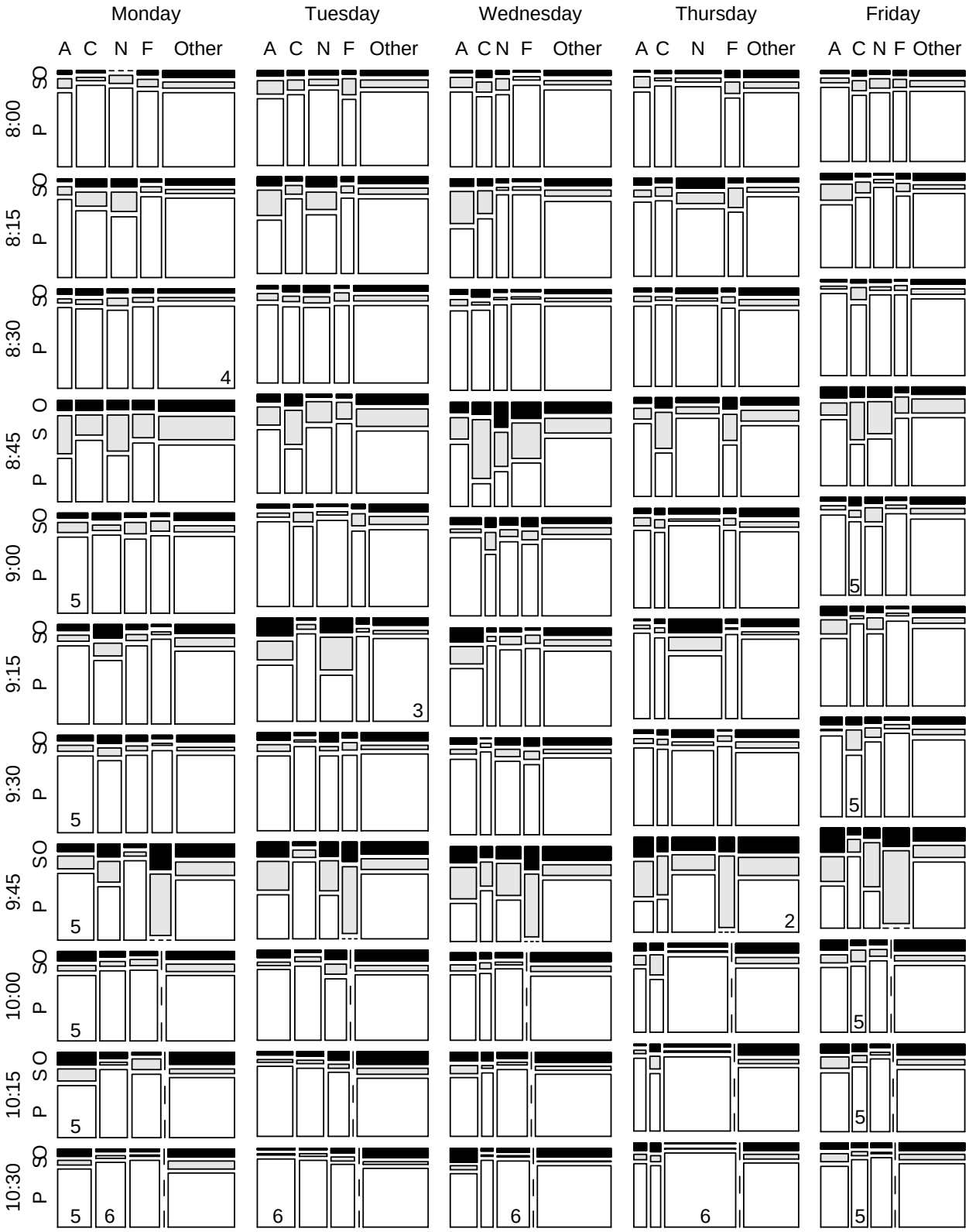


Figure 5. Mosaic of network shares and audience transitions. P = persistent, S = switch, O = off. Numbered tiles are discussed in Section 4.

Thursday	Transition			
9:45 Network	Off	Persist	Switch	9:45 Network Total
ABC	54	86	99	239
CBS	21	47	56	124
NBC	80	349	94	523
FOX	31	0	156	187
CABLE	135	443	152	730
Transition Total	321	925	557	1803

Table 1. Thursday 9:45 Contingency Table.

variables are included. These one-way mosaics show that the aggregate audience is smaller later in the week (Figure 1a) and later in the evening (Figure 1b). The mosaic corresponding to the two-way table of the aggregate audience, divided first by day and then by time, appears in Figure 3 just for clarity of exposition – in this example, interesting analysis begins with the addition of specific network counts by day and time.

As we add the network variable (to simplify exposition, the term "network" will include the aggregate cable, or non-network, alternative) and the transition categories to the mosaic (Figures 4 and 5, respectively), several points illustrate the use of these mosaics in studying television viewer behavior. The following numbers are marked in the relevant tiles in the mosaics.

1. When the network variable is added to the two-way mosaic in Figure 3 to form a three-way contingency table, the resulting mosaic tiles at each day and time represent the network ratings, or share of the viewing audience (Figure 4). For example, on Thursday at 10:00, 685 of 1692 viewers watching television were tuned into NBC's hit *E.R.*, so the NBC rectangle occupies 40.4% of the area in Thursday's 10:00 tile.
2. Figure 5 includes an additional variable with three categories: among the viewers watching a certain network (at time t on day d), some turn the TV off and do not watch anything at time $t+1$ (represented by the black tiles); others switch networks at time $t+1$ (shaded tiles), while the remaining viewers watch the same network, or *persist* (unshaded tiles). For example, consider the NBC viewers in the Thursday 9:45 tile who watch the end of *Caroline in the City*: 523 of 1803 viewers watching television then tuned into the end of *Caroline in the City* – the NBC tile is 30% of the area of the Thursday 9:45 tile. Of the 523 viewers, only 80 turned the television off at 10:00 (black tile), 94 switched to a different network at 10:00 (shaded tile), and the remaining 349 watched the beginning of *E.R.* on NBC (persisting in their viewing of NBC, the unshaded area). Table 1 presents the two-way contingency table for the viewers watching television at 9:45 classified by network choice and viewing behavior after the quarter-hour. Note that there can be no viewers persisting in watching FOX from the 9:45 quarter-hour – these FOX viewers must either turn the TV off or switch channels. This empty cell corresponds to the empty transition tile in the FOX 9:45 tile. Similarly, all FOX tiles after 10:00 are empty.
3. A quick study of the TV schedule in Figure 2 and the mosaic in Figure 5 shows that viewer persistence is higher when there is show continuity. For example, on Tuesday night after the 9:15 quarter-hour, CBS and FOX have continuations of longer shows (both are movies) while ABC and NBC start new shows at 9:30 (competing half-hour comedies). This tile shows a striking example of high persistence with show continuity and lower persistence going into new programming: ABC and NBC have lower persistence rates of roughly 60% and 50%, while CBS and FOX enjoy high persistence rates of close to 90% each. Note the uniformly high degree of switching at 8:45 and 9:45 in Figure 5.
4. It is also evident from the mosaic that persistence during the odd quarter-hour transitions (that is, always during a show) is fairly uniform between the networks, and usually high compared to other transitions. The 8:30 frame on Monday, for example, shows uniformly high flow of viewers persisting into 8:45.
5. These mosaics provide insight into different sources of viewer persistence. The primary trend appears to be higher persistence during shows (and lower persistence at end of shows), but more specific elements of persistence are also evident in the mosaics. First, consider *Monday Night Football* on ABC after 9:00. There is unusually

low persistence given the show continuity, particularly after 10:00, because sports and news programs fail to maintain the audience as effectively as other programs. CBS's *Ice Wars* figure skating event on Friday also has a slightly lower persistence rate given the show continuity.

6. Finally, consider dramas such as *Chicago Hope* (Monday at 10:00 on CBS), *NYPD Blue* (Tuesday at 10:00 on ABC), *Law & Order* (Wednesday at 10:00 on NBC), and *E.R.* (Thursday at 10:00 on NBC). All have particularly high persistence into the final quarter-hour – viewers watching the later parts of these popular dramas tend to finish watching rather than turning away before the climax.

It should be noted that although these mosaics focus attention on network persistence, viewer persistence in the other alternatives must also be addressed. Persistence in the aggregate non-network category is understandably high, since only switches from a non-network alternative into a major network and back again are observed (no switching between non-network alternatives can be studied). A detailed study of overall rates of switching would require a richer data set. Viewers not watching television also persist in not watching television, though these counts are not included in this study.

Mosaics are a promising method for displaying multivariate categorical data, and it is hoped that this S-PLUS implementation will be useful to the statistical community.

Acknowledgements

The authors wishes acknowledge the guidance of John Hartigan in refining these mosaics, and Ron Shachar (Yale School of Management), and Greg Kasparian and David Poltrack of CBS for their help in obtaining the data for this study.

Additional Resources

Additional resources associated with this article – both the software and the data – are available at the Web site `www.stat.yale.edu/~emerson/JCGS/`.

References

Friendly, M. (1992), "User's guide for MOSAICS: A SAS/IML program for mosaic displays," Technical Report 206, Department of Psychology, York University.

See also `www.math.yorku.ca/SCS/friendly.html`

Friendly, M. (1994), "Mosaic displays for multi-way contingency tables," *Journal of the American Statistical Association*, **89**, 190–200.

Hartigan, J. A., and Kleiner, B. (1981), "Mosaics for contingency tables," *Proceedings of the 13th Symposium on the Interface between Computer Science and Statistics*.

Hartigan, J.A., and Kleiner, B. (1984), "A mosaic of television ratings," *The American Statistician*, **38**, 32–35.

Nielsen Media Research (1995), *National Reference Supplement*, A.C. Nielsen.

Theus, M. (1997a), "Visualization of categorical data," in *Advances in Statistical Software 6*, 47–55, Lucius & Lucius.

Theus, M. and Lauer, St. R. W. (1997b), "Visualizing log-linear models," submitted to *Journal of Computational and Graphical Statistics*.

The Web site `www.research.att.com/~theus/Mondrian/Mondrian.html` has more information on these tools.

John W. Emerson
Yale University
`emerson@stat.yale.edu`