Getting Sankey with Bar Charts

Shane Rosanbalm, Rho, Inc., Chapel Hill, NC

ABSTRACT

In this paper we present a SAS® macro for depicting change over time in a stacked bar chart with Sankey-style overlays.

Imagine a clinical trial in which subject disease severity is tracked over time. The disease severity has valid values of 0, 1, 2, and 3. The time points are baseline, 12 months, 30 months, and 60 months.

A straightforward way to represent this data would be with a vertically-oriented stacked bar chart. Visit would be used as the x-axis variable. Disease severity would be used to form the groups in the stacked bars. The y-axis measure would be percent of subjects in each group.
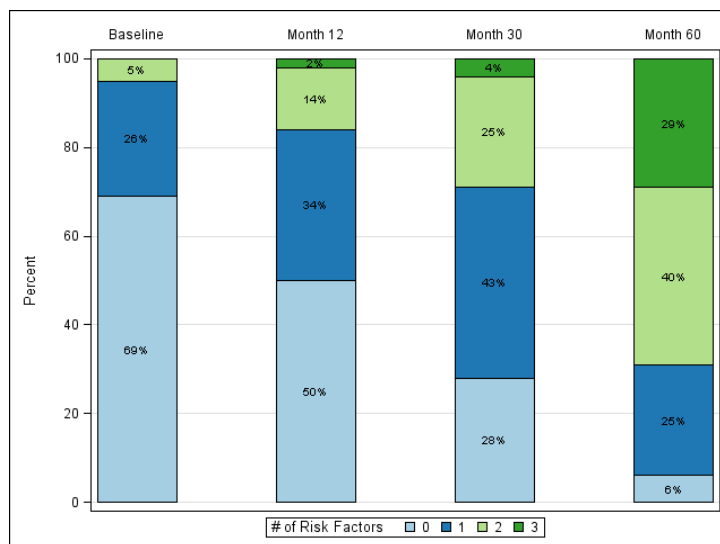

Figure 1. A Longitudinal Bar Chart

This type of data visualization allows us to see the change within each group over time. However, the data visualization does not allow us to see which group of subjects is driving these changes. For instance, if the number of subjects at severity level 1 were to increase from baseline to month 12, how do we know whether the new subjects are coming from group 0, 2, or 3?

Sankey diagrams provide a visual depiction of the magnitude of flow between nodes in a network. If we think of the groups in a stacked bar chart as these nodes, then Sankey-style overlays can be used to show how the subjects flow from one severity level to another over time. In this paper we will present just such a data visualization.

SANKEY DIAGRAMS

Sankey diagrams provide a visual depiction of the magnitude of flow between nodes in a network. Consider the following example.
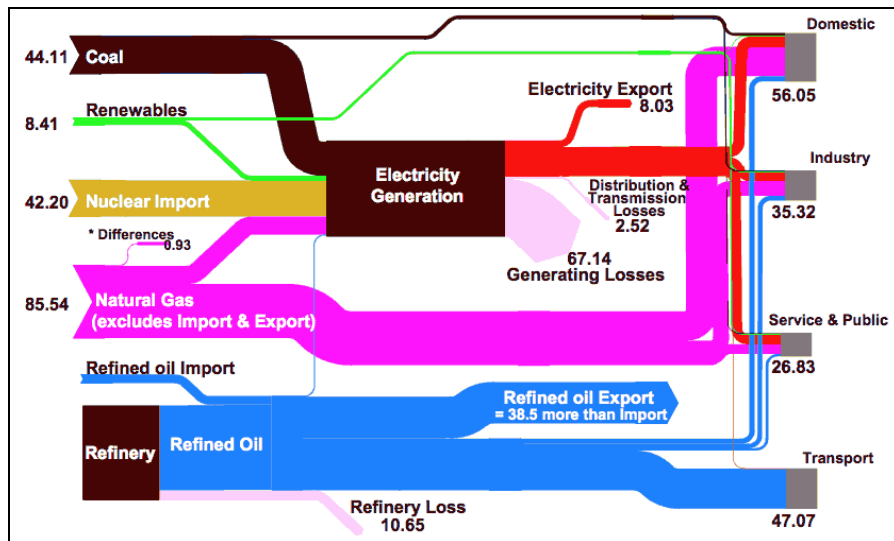

Figure 2. A Sankey Diagram of Energy Supply and Consumption

In this Sankey diagram the nodes and links represent energy. The nodes on the left represent energy sources and the nodes on the right represent energy consumers. The nodes and links are drawn in proportion to the amount of energy. For instance, there are two links flowing from the coal node. The upper link is very thin whereas the lower link is much thicker. This indicates that most coal is used for electricity generation, with a very small amounts used directly by consumers.

Sankey diagrams can be used to represent quantities other than energy. Consider the following example taken from the world of auto racing.
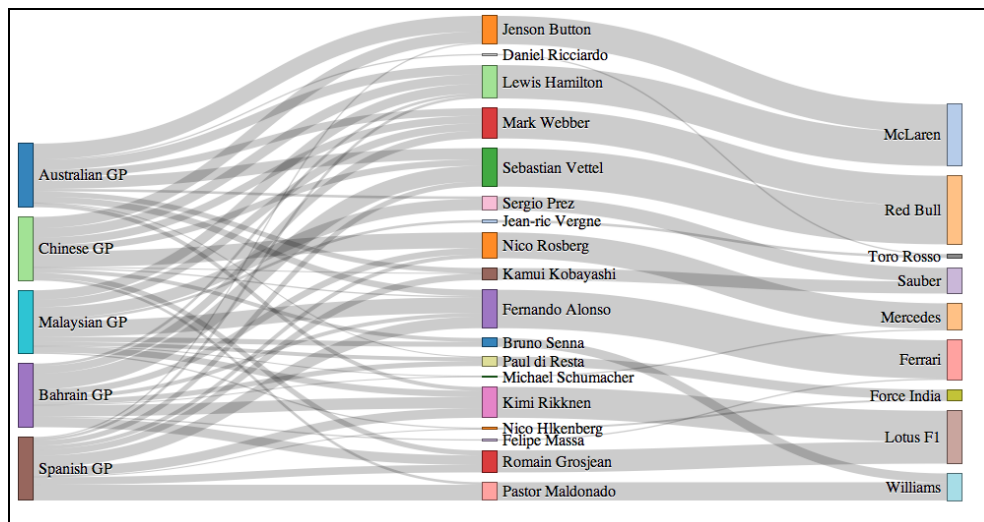

Figure 3. A Sankey Diagram of Points Distribution in Auto Racing

In this Sankey diagram the nodes and links represent driver points in a racing series. Several things are immediately obvious based on this diagram. We can see from the sizes of the nodes on the left that each race awards the same number of points. We can see from the sizes of the nodes on the right that Red Bull is the most successful team. We can see that Lewis Hamilton has earned points in every race, and yet he does not have the most points.

LONGITUDINAL BAR CHARTS

Leaving Sankey diagrams behind for a moment, consider the following bar chart. This chart depicts data that is collected on the number of risk factors that subjects exhibit at each of several study visits. You might collect data on risk factors when direct measurement of the disease of interest is invasive or cost-prohibitive.
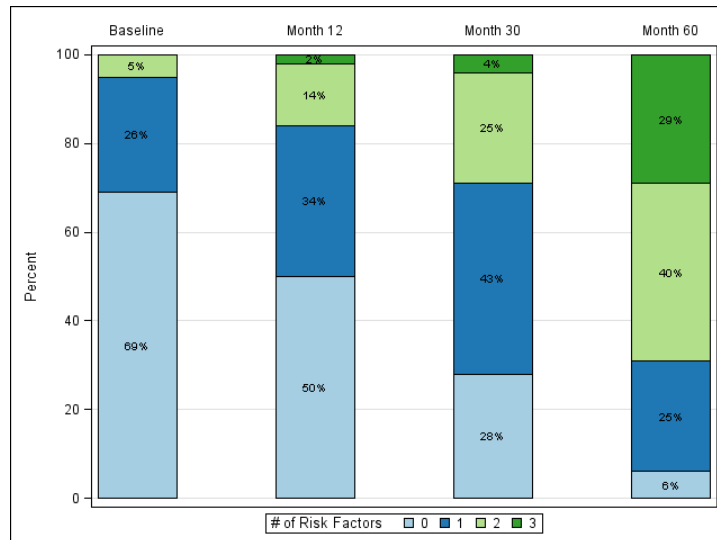

Figure 4. A Longitudinal Bar Chart

Notice how the number of risk factors steadily increases over time. For instance, at Baseline we have 0% of subjects with 3 risk factors, whereas at Month 12 we have 2% of subjects with 3 risk factors. It is sometimes of interest to ask questions of the form: What group spawned the 2% of subjects with 3 risk factors at Month 12?

A common guess is that these subjects are from the group with 2 risk factors at Baseline. The implicit assumption behind this guess is that a subject's risk profile is relatively stable over time. But, how do we know this to be the case? Perhaps the number of risk factors is highly variable over time.

This paper presents a macro designed to create longitudinal bar charts with Sankey-style overlays. The Sankey-style overlays allow the reviewer to more easily answer questions of the form: What group spawned these 2% of subjects with 3 risk factors at Month 12?
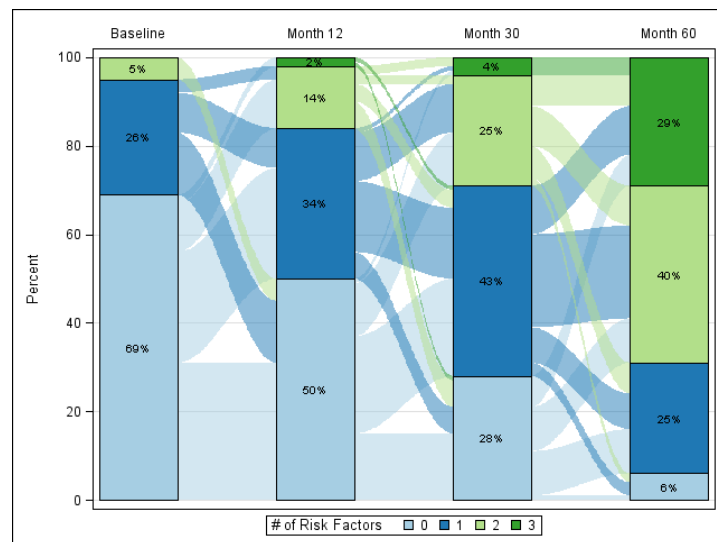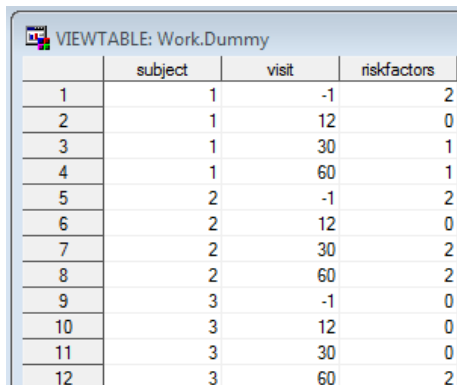

Figure 5. A Bar Chart with Sankey-style Overlays

In this Sankey diagram the nodes and links represent subjects. We can see that the 2% of subjects with 3 risk factors at Month 12 all had 0 risk factors at Baseline. We can also see that the risk profile is highly unstable over time.

SOME DATA TO WORK WITH

The dataset used to generate the Sankey bar charts in this paper is a so-called vertical dataset.

	subject	visit	riskfactors
1	1	-1	2
2	1	12	0
3	1	30	1
4	1	60	1
5	2	-1	2
6	2	12	0
7	2	30	2
8	2	60	2
9	3	-1	0
10	3	12	0
11	3	30	0
12	3	60	2

VIEWTABLE: Work.Dummy

Display 1. A Vertical Dataset

This term comes from the fact that the multiple outcomes for each subject are stored on separate records (i.e., vertically). Some datasets are so-called horizontal datasets, in which there is one record per subject and separate variables for each visit. Given the current prominence of the CDISC ADaM standards within the pharma industry, the Sankey bar chart macros assume a vertical dataset as the source.

Appendix 1 contains the data step code used to generate the above dataset.

SANKEY BAR CHARTS

Sankey bar charts are produced using a set of three SAS macros. There is an outer container macro, %SankeyBarChart, which calls to two helper macros.

- The first of these helper macros is called %RawToSankey. This macro converts a vertical dataset (i.e., one record per subject and visit) into two summary datasets. The first summary dataset is for the Sankey nodes (i.e., the bar segments) and the second summary dataset is for the Sankey links (i.e., the connectors).

- The second of these helper macros is called %Sankey. This macro uses the above summary datasets to produce the Sankey bar chart by way of the SGPLOT procedure.

THE FIRST HELPER MACRO

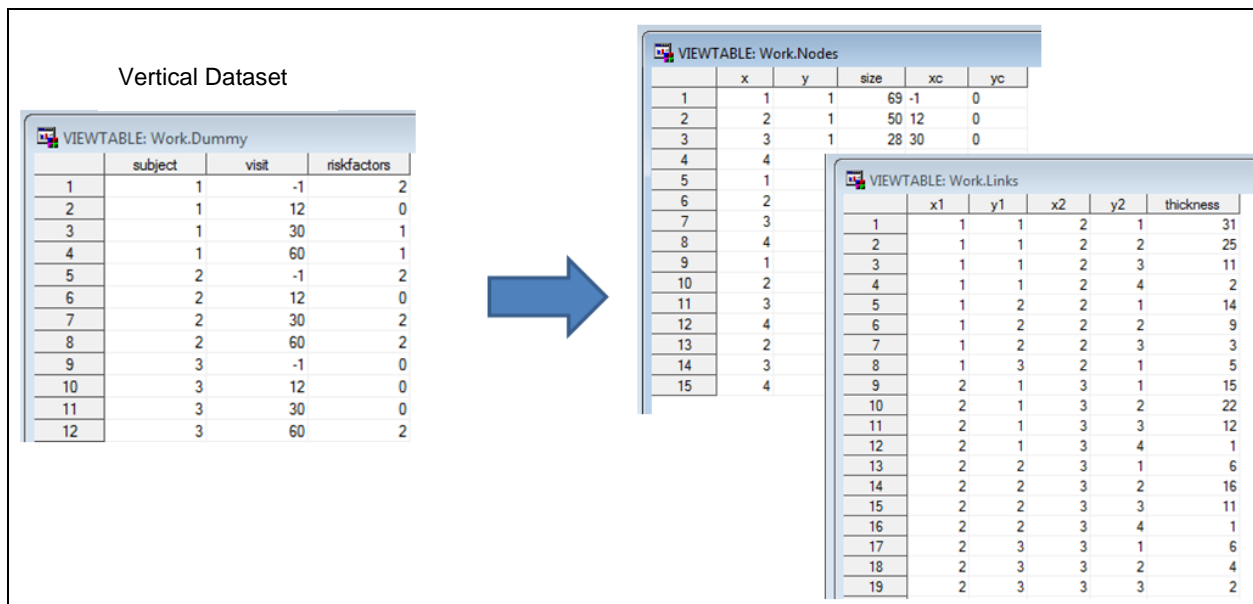The %RawToSankey helper macro has 4 required parameters and 2 optional parameters.

Parameter	Description	Required?
data	Vertical dataset to be converted	Yes
subject	Subject identifier	Yes
yvar	Categorical y-axis variable	Yes
xvar	Categorical x-axis variable	Yes
yvarord	Sort order for y-axis values E.g., yvarord=%str(red rum, george) (default is equivalent to ORDER=DATA)	No
xvarord	Sort order for x-axis values E.g., xvarord=%str(pink plum, fred) (default is equivalent to ORDER=DATA)	No

Table 1. Parameters for %RawToSankey Macro

Using the aforementioned vertical dataset, a typical call to %RawToSankey might appear as follows.

```
%rawtosankey
    (data=dummy
    ,subject=subject
    ,yvar=riskfactors
    ,xvar=visit
    ,yvarord=%str(0, 1, 2, 3)
    ,xvarord=%str(-1, 12, 30, 60)
    );
```

The macro takes the vertical dataset at left and creates two summary dataset: *nodes* and *links*.

Display 2. Converting a Vertical Dataset into Sankey-ready Datasets *Nodes* and *Links*

THE SECOND HELPER MACRO

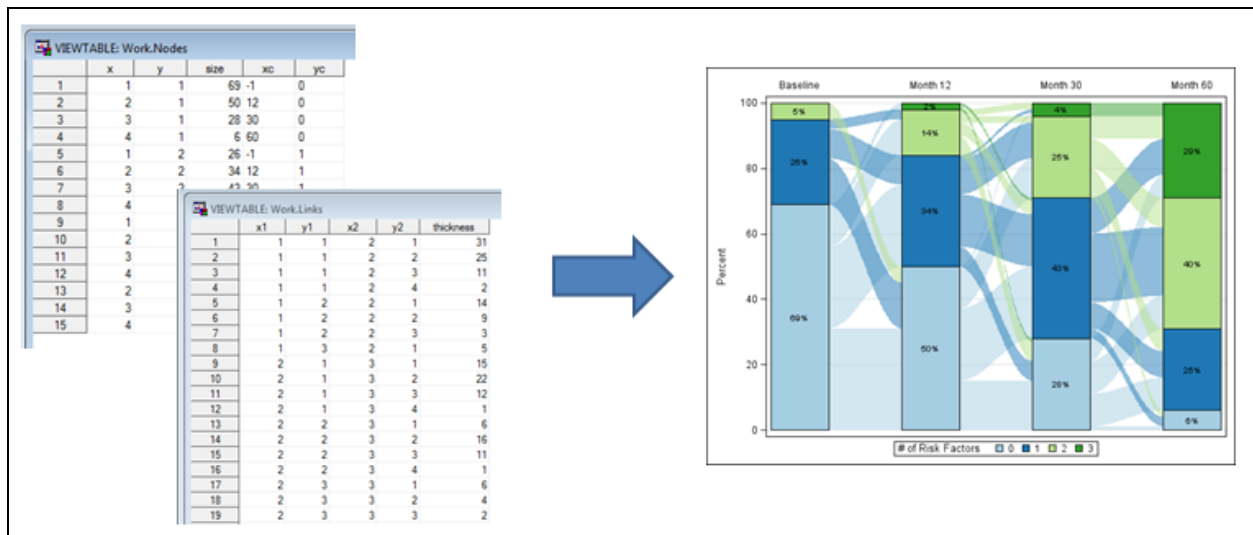The %Sankey helper macro has no required parameters and 6 optional parameters.

Parameter	Description	Required?
colorlist	A space-separated list of colors: one per y-value E.g., colorlist=red vlio cxb2df8a (default is qualitative Brewer palette)	No
barwidth	Width of bars Valid values are from 0-1 (default is 0.25)	No
xfmt	Format for x-axis	No
legendtitle	Text for legend title	No
interpol	Method of interpolating between bars Valid values are cosine, linear (default is cosine)	No
percents	Show percents inside each bar Valid values are yes, no (default is yes)	No

Table 2. Parameters for %Sankey Macro

Continuing with the aforementioned vertical dataset, a typical call to %Sankey might appear as follows.

```
%sankey
    (barwidth=0.45
    ,xfmt=xfmt.
    ,legendtitle=%str(# of Risk Factors)
    );
```

The macro utilizes the aforementioned *nodes* and *links* datasets to produce the chart by way of the SGPLOT procedure.

Display 3. Converting Sankey-ready Datasets *Nodes* and *Links* into a Sankey Bar Chart

THE CONTAINER MACRO

The %SankeyBarChart macro is nothing more than a container for the two helper macros. As such, the parameter list for the container macro is nothing more than the sum of the parameter lists for the helper macros.

Parameter	Description	Required?	
data	Vertical dataset to be converted	Yes	
subject	Subject identifier	Yes	
yvar	Categorical y-axis variable	Yes	Data manipulation
xvar	Categorical x-axis variable	Yes	
yvarord	Sort order for y-axis values E.g., yvarord=%str(red rum, george) (default is equivalent to ORDER=DATA)	No	
xvarord	Sort order for x-axis values E.g., xvarord=%str(pink plum, fred) (default is equivalent to ORDER=DATA)	No	
colorlist	A space-separated list of colors: one per y-value E.g., colorlist=red vlio cxb2df8a (default is qualitative Brewer palette)	No	
barwidth	Width of bars Valid values are from 0-1 (default is 0.25)	No	
xfmt	Format for x-axis	No	Formatting
legendtitle	Text for legend title	No	
interpol	Method of interpolating between bars Valid values are cosine, linear (default is cosine)	No	
percents	Show percents inside each bar Valid values are yes, no (default is yes)	No	

Table 3. Parameters for %SankeyBarChart Macro

A typical call to %SankeyBarChart might appear as follows.

```
%sankeybarchart
   (data=dummy
   ,subject=subject
   ,yvar=riskfactors          Data manipulation
   ,xvar=visit
   ,yvarord=%str(0, 1, 2, 3)
   ,xvarord=%str(-1, 12, 30, 60)
   ,barwidth=0.45
   ,xfmt=xfmt.                 Formatting
   ,legendtitle=%str(# of Risk Factors)
   );
```
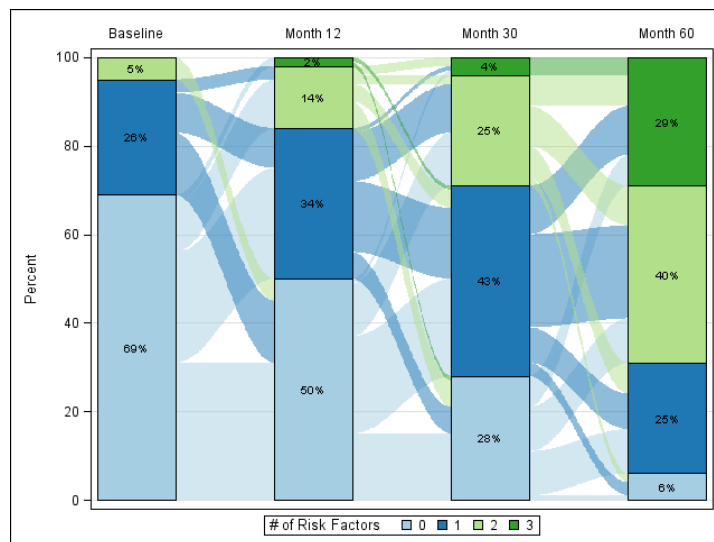
Figure 6. A Bar Chart with Sankey-style Overlays

CONCLUSION

Sankey bar charts are enhancements of longitudinal bar charts that add Sankey-style overlays between the bars at adjacent time points. These overlays illustrate which groups are driving changes in the bars over time, providing deeper insight into the data being visualized. The SAS macros presented in this paper can be used to assist in the creation of Sankey bar charts.

MACRO SOURCE CODE

The Sankey bar chart macro source code is available for direct download at graphics.rhoworld.com/tools/sankeybarchart. Alternatively, send email requests to graphics@rhoworld.com.

RECOMMENDED READING

- Graphically Speaking

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Shane Rosanbalm
Enterprise: Rho, Inc
Address: 6330 Quadrangle Drive
City, State ZIP: Chapel Hill, NC 27517
Work Phone: 919-595-6273
E-mail: shane_rosanbalm@rhoworld.com
Web: graphics.rhoworld.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX 1: SOME DATA TO WORK WITH

The dataset used to generate the charts in this paper was created with the RANNOR function.

```
data dummy;
    do subject = 1 to 100;
        do visit = -1, 12, 30, 60;
            random = rannor(1) + 0.5 + (visit+1)/30;
            riskfactors = min(3,floor(abs(random)));
            output;
        end;
    end;
run;
```

APPENDIX 2: SOME NOTES AND TECHNICAL DETAILS

1. The %RawToSankey macro enumerates the values of the XVAR and YVAR variables. This enumeration greatly simplifies the internally macro juggling between the datasets NODES and LINKS. The default enumeration values are produced using ORDER=DATA principles. If the XVAR and YVAR values in your dataset are not presorted in the order in which you would have them displayed you have two choices. First, you could sort your dataset. Second, you could use the optional data manipulation parameters XVARORD and YVARORD. The macro supports both character and numeric variables for XVAR and YVAR. Because character values are supported, the value lists for XVARORD and YVARORD must be provided in a delimited list. The macro is written to require a comma-separated list. This design decision necessitates the use of macro quoting functions (joy!) in specifying the parameters XVARORD and YVARORD. If you don't like using macro quoting functions, sort your dataset.

2. The %RawToSankey macro subsets the vertical dataset down to complete cases. That is, only subjects with data at all visits are allowed to appear in the final chart. This is why SUBJECT is included as a required parameter in the data manipulation.

3. The %Sankey macro uses SGPLOT to produce the chart. The bar segments are drawn with HIGHLOW, the connectors with BAND, and the annotations with SCATTER. Why HIGHLOW instead of VBAR you ask? The BAND plot requires a linear x-axis. The VBAR plot requires a discrete x-axis in order to allow for adjustments in the bar widths. This is not compatible with the BAND plot. Adjustable bar widths is considered an important feature of the macro, hence HIGHLOW is used to produce the bar segments.

4. The %Sankey macro produces one HIGHLOW statement per bar segment and one BAND statement per connection. This approach is used because of the need to change colors for one bar segment/connection to the next. The X/X1 variables in the NODES/LINKS datasets have a 1-to-1 correspondence with the COLORLIST, so writing separate statements for each bar segment/connection actually provided a straightforward way to color each bar segment/connection correctly.

5. These macros were written using SAS 9.3. Several new features in SAS 9.4 might have sent the %Sankey portion of the code in entirely different directions. Alas, we write our macros with the tools we have, not the tools we wish we had. Keep this in mind when, in the course of reviewing the source code you find yourself asking questions like, "Why did he do it THAT way!?"