Customized Proportional Venn Diagrams from SAS® System

Shiqun (Stan) Li, Minimax Information Services, Belle Mead, NJ

ABSTRACT

Within our SAS community, there has been a high demanding for a SAS procedure or SAS macro that can generate proportional Venn diagrams[1]. In SAS Global Forum 2009[2], for the first time we presented an algorithm and a SAS macro that can create proportional Venn diagrams. The macro uses ANNOTATE functions to draw the Venn diagrams. In this presentation, we will introduce another SAS macro, which utilizes PROC GPLOT instead to generate customized proportional Venn diagrams. The Venn diagrams can have enhanced patterns, customized colors, pop-up and drill-down properties. Several Venn diagram examples will be exhibited to demonstrate the usages of our SAS macro. This paper is prepared for an intermediate and advanced audience.

Key Words: Proportional Venn diagram, SAS/Graph, SAS/Macro, Data Visualization

BACKGROUND

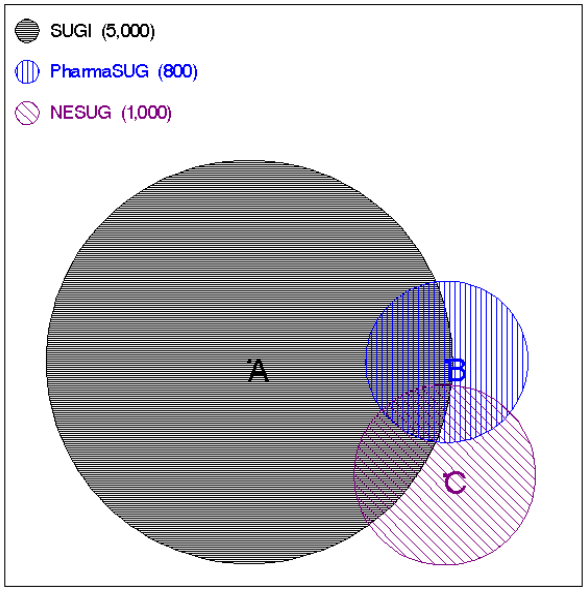
A Venn diagram usually has three circles that represent three subsets for the same population. The three circles are supposed to be arranged with these two characteristics: The areas of the circles are proportional to the relative sizes of the subsets (such as number of counts or percentages); the overlap areas of the circles stand for the sharing amount between the coresponding sets. The intersected areas, or the set relationships, are particularly interested in many applications. This kind of Venn diagrams were first proposed for representing logical positions by John Venn in 1880.

For the first time, we presented an algorithm and a SAS macro that can create proportional Venn diagrams in SAS Global Forum 2009[2]. In the macro, we use SAS ANNONTATE functions to produce the diagrams.

VENN DIAGRAMS AND PROPERTIES

At this point, we would like to introduce an example of proportional Venn diagrams. Suppose we are interested in the number of attendees of our three SAS user conferences: SUGI (Set A), PharmaSUG (Set B) and NESUG (Set C). All these three sets are subsets of SAS users around the world. Assume, for a certain year, there were 5,000 attendees for SUGI, 800 people for PharmaSUG and 1,000 persons for NESUG. In addition, we estimate 400 people went to both SUGI and PharmaSUG, 300 joined both SUGI and NESUG, 200 attended both PharmaSUG and NESUG. That is, |A|=5,000, |B|=800, |C|=1,000, |AB|=400, |AC|=300 and |BC|=200. Please note, the numbers in this example are not to be accurate. The numbers are for the example purpose only.

Given these set relationships, we can visually display the sets of the three conference attendees as a proportional Venn diagram as below:



In this diagram, circular area of A , representing |A| which is the number of SUGI attendees, is the largest, much larger than |B| and |C|. The intersected area of A and B in the graph is proportional to the size of AB. That is, |AB| (those who attended both SUGI and PharmaSUG) is proportionally represented by the joined area of A and B. Similarly, the intersection areas of A and C is proportional to the number of attendees to both SUGI and NESUG; and the overlay area of B and C stands for the number of people who went to both PharmaSUG and NESUG. From this diagram above, it is easy to visually demonstrate the relative sizes and the overlay relationships of these three conferences.

This kind of Venn diagrams can be an excellent tool in epidemiology, healthcare and marketing. In epidemiology, if we let the 3 subsets A, B and C represent 3 kinds of diseases respectively, the graph will present the relationships of the 3 diseases. We will demonstrate some examples on this later.

Venn diagram of circular shape is a very interesting topic. Proportional Venn diagrams and the properties were well discussed in several previous studies[4, 5, 6]. Here we will only point out the important properties that will be used in our algorithm.

Property 1:
Given two cicles with radii r_1 and r_2, assume there is distance of d units between these two circles. Then the overlay area of these two circles will be:
$A=½\ r_1^2(\alpha-\sin(\alpha)) + ½\ r_2^2\ (\beta-\sin(\beta))$
Where

$$\alpha=2\arccos[(d^2 + r_1^2 - r_2^2)/(2dr_1)]$$
$$\beta =2\arccos[(d^2 + r_2^2 - r_1^2)/(2dr_2)]$$

Proporty 2:
Suposse the distance of three circles A_1, A_2 and A_3 are d_{12}, d_{13} and d_{23}, then the centers of the three circles can be as followings: A_1 (0,0), A_2 (d_{12}, 0) and A_3 (x_3, y_3)
Where $x_3=(d_{12}^2+ d_{13}^2- d_{23}^2)/(2d_{12})$ and $y_3=\text{sqrt}(d_{13}^2-x_3^2)$

ALGORITHM

We here present an algorithm to create an area-proportional Venn diagram of 3 sets: A_1, A_2 and A_3. First, assume $|A_1|= a_1$, $|A_2|= a_2$, $|A_3|= a_3$, $|A_1A_2|= a_{12}$, $|A_1A_3|= a_{13}$, $|A_2A_3|=a_{23}$ and $|A_1A_2A_3|=a_{123}$. Without loss the generality, we further assume A_1, A_2 and A_3 are subsets of a population P, and with $0<=a_i, a_{ij}, a_{ijk}<1$ (where$1<=i<j<k<=3$). Based on these assumptions, we propose the following SAS algorithm to draw a proportional Venn diagram of circle shapes:

a) Calculate the radius of the circle for each set: $r_i=\text{sqrt}[a_i/(\pi)]$, where $|A_i|=a_i$.
b) Draw the first circle A_1 with center (0, 0) and radius r_1.
c) Compute the distance d_{12} for A_1 and A_2, given r_1, r_2 and $a_{12}=|A_1A_2|$.
d) Draw the second circle A_2 with center (d_{12}, 0) and radius r_2.
e) Estimate the distance d_{13} for A_1 and A_3, and distance d_{23} for A_2 and A_3.
f) Find the center (x_3, y_3) of the third circle A_3, where $x_3=(d_{12}^2+ d_{13}^2- d_{23}^2)/(2d_{12})$ and $y_3=\text{sqrt}(d_{13}^2-x_3^2)$.
g) Draw the third circle A_3 with center (x_3, y_3) and radius r_3.
h) Add the legend and others as required to the diagram.

In this algorithm, the key points are:

1) How to find the distances among the three circles.
2) How to draw the circles, given the centers and the radii.

1) SAS Macro for finding the distances

From Property 1, it is easy to see that the distance is hidden in a non-linear equation. There are many existed techniques for sloving such equations[7, 8]. We choose the numerical technique of bisection to slove for the distance d in the equation in Property 1, given the intersected area A, the radii r_1 and r_2. Of course, Newton-Raphson iteration method will work well too. But due to the monotonity of the function $A(d)=½\ r_1^2(\alpha-\sin(\alpha)) + ½\ r_2^2\ (\beta-\sin(\beta))$, we prefer to use the bisection method instead. The bisection algorithm can be coded in SAS as this:

```
%macro bisection(r1,r2,A, root ) ;
T=1E-6;
_a=abs(&r1-&r2)+T; _b=(&r1+&r2)-T;
```

```
DO while(1);
       _c=(_a+_b)*0.5;
       F_a=%func(&r1,&r2,_a, &A.);
       F_c=%func(&r1,&r2,_c, &A.);
       if abs(F_a) le T then do; &root.=_a; leave; end;
       if abs(F_c) le T then do; &root.=_c; leave; end;
       if F_a*F_c lt 0 then _b=_c;
       else _a=_c;
       if _b-_a<T then do;
         &root.=_c; leave;
       end;
END;
%mend bisection;
```

Where the self-defined function FUNC is the formula in Property 1, as following in our SAS language:

```
%macro func(r1, r2, d, A);
  &r1*&r1*arcos((&d*&d+&r1*&r1-&r2*&r2)/(2*&r1*&d)) +
  &r2*&r2*arcos((&d*&d+&r2*&r2-&r1*&r1)/
  (2*&d*&r2))-0.5*sqrt((&r1+&r2-&d)*(&r1+&d-&r2)*(&r2+&d-&r1)*(&r1+&r2+&d))-&A;
%mend func;
```

2) SAS Macro for drawing the circles

There are several ways to create a circle from SAS System. In our paper presented in SAS Global Forum 2009[2], we used SAS/ANNOTATE to draw the circles. The syntax for making a circle with a dot as center point can be like this:

```
function='pie'; x=&x.; y=&y.; size=&size; color="&Color."; line=0;
style="&PieStyle."; rotate=360; output;  *** a circle;
function='pie'; size=.3; style='solid'; output;  *** a dot at center;
```

Where: `PieStyle=PnNa`: n=line density, a=angle of lines/patterns.
Here, users can customize the color, line density and the line angle inside each circle.

In addition to the ANNOTATE technique, we can use PROC GPLOT to draw a circle too. The advantage of the GPLOT approach is easy to be customized. The disadvantage is that there will require more lines of codes.

Here is an example for creating a circle with center=(0,0) and radius=1:

```
goption reset=all hsize=7in vsize=7in border;
symbol1 interpol=m3n135 cv=blue co=black;
symbol2 i=none value=dot h=.75 c=blue;
axis1 order=(-1 to 1) major=none minor=none label=none value=none;
proc gplot data=circle;
  plot y*x=set/ href=0 vref=0 haxis=axis1 vaxis=axis1 nolegend;
run; quit;
```

where the dataset circle is generated by the code below:

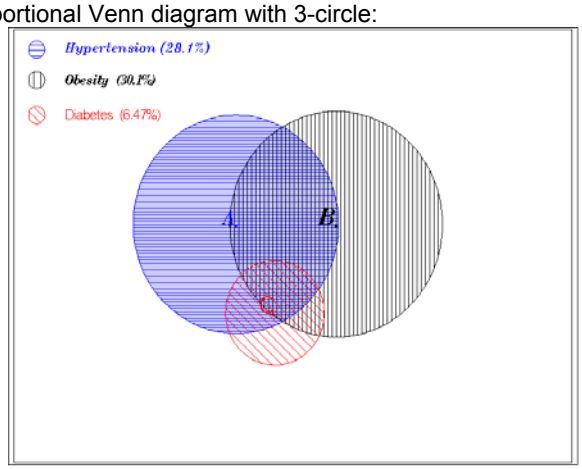
```
data circle;
do seq=0 to 99;
  set=1;
  x=cos(2*3.14159*seq/100);
  y=sin(2*3.14159*seq/100);
  output;
end;

*** center;
set=3; x=0; y=0; output;
run;
```

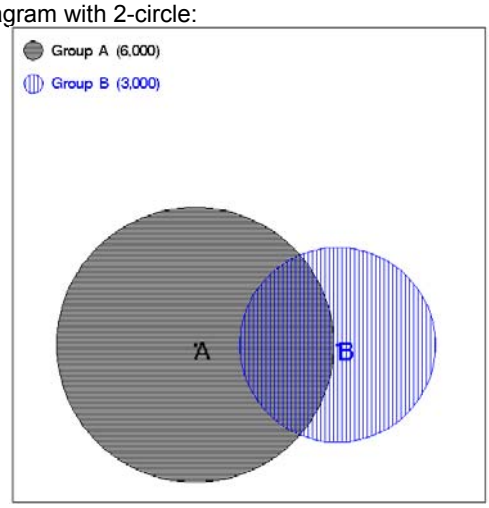
APPLICATION EXAMPLES

In this part, we will demonstrate several graphs that we generated using our Venn diagram macros. The data in the graphs mostly comes from NHANES database, a National Health and Nutrition Examination Survey database which can be found from this link: http://www.cdc.gov/nchs/nhanes.htm. Please note the numbers in the graphs are intended for demonstration purposes only.

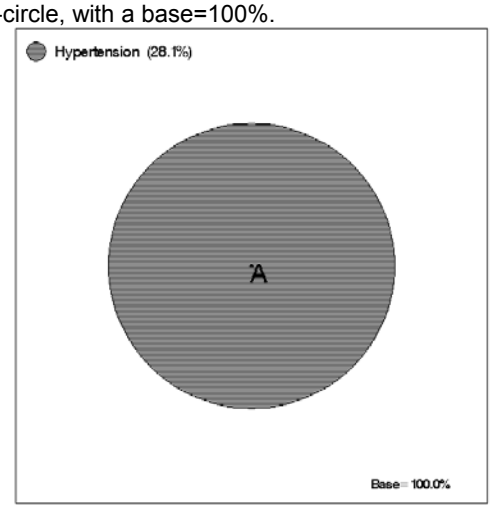
An example of typical proportional Venn diagram with 3-circle:



Another example of a Venn diagram with 2-circle:

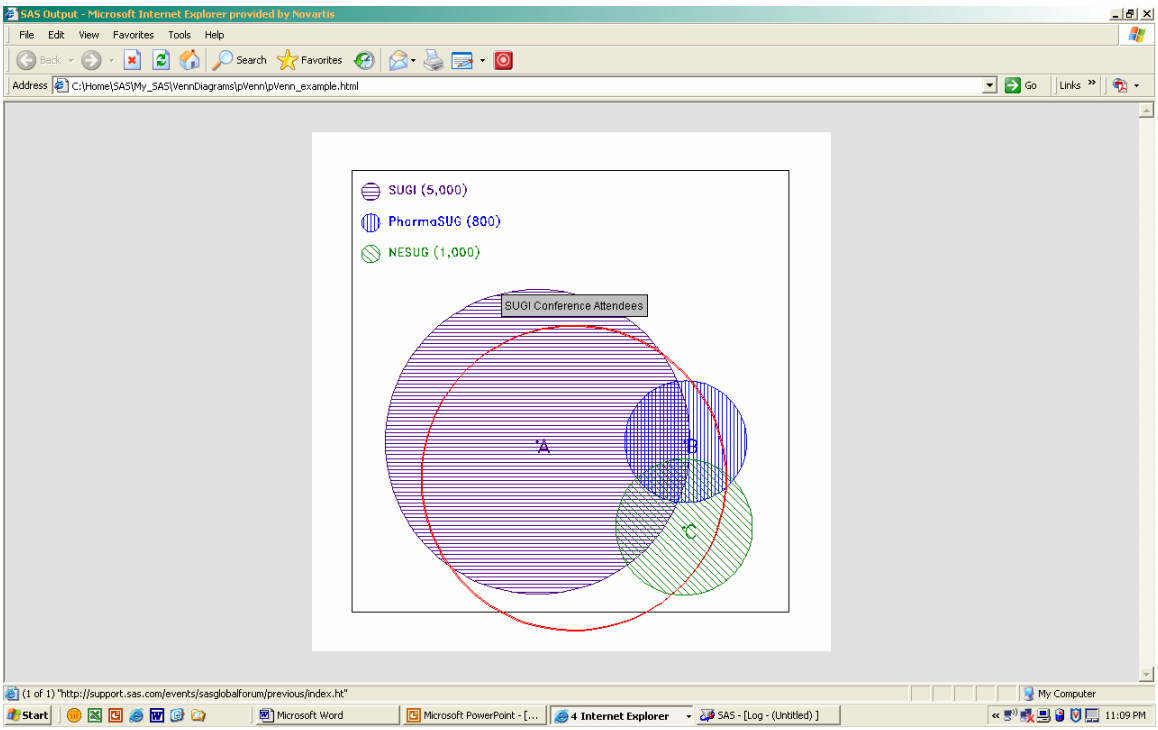


A Venn diagram with only one-circle, with a base=100%.



The graph below is a dynamic proportional Venn diagram. This interactive graph will have pop-up and drill-down functions. When we can move the mouse to different areas of the circles, additional information associated with this circle will be displayed in a pop-up box. Furthermore, if we click at a circle, it will take us to a pre-generated report, graph or link.

We use ANNOTATE method to generate the dynamic display below. For the techniques on how to create an interactive SAS graph, please refer to the papers [3, 10].



LIMITATIONS AND DISCUSSIONS

The intersected area of all 3 circles is called triangle circular. Given 3 sets A, B and C, the values |A|, |B|, |C|, |AB|, |AC| and |BC| will be enough to determine the corresponding locations of the 3 circles representing the 3 sets. The intersection area |ABC| is redundant. Therefore, it must be noted the overlay triangle circular may not be exactly proportional to the intersection of all three sets A, B and C. This is a limited nature for Venn diagrams of circle shape. There should have a footnote to the graph to notify the readers if there is a significant difference in this triangle circular.

It is also necessary to point out there are some certain situations that cannot be represented by 3-ring proportional Venn diagrams. A case like this is: Subset A: Male=50%, Subset B: Female=50%, Subset C: certain disease with prevalence=25% (15% for Male and 10% for Female). For this circumstance, it is impossible to use a Venn diagram with 3 circles to display the relationships of the data. This is very obvious in geometry. When this type of situations are encountered, the macro will issue a warning message and stop further procedures. This problem could be avoided by choosing a different geometry shape like rectangle for the Venn diagram.

To this point, we only discuss Venn diagram with up to 3 rings. If there are more than 3 sets to be displayed, a Venn diagram of circle shape may not be a good choice [5]. For most situations, it is impossible to position the location for the 4^{th} circle for the 4^{th} set. Other approaches with alternative shape will be required for cases with more than 3 sets. But fortunately, a 2-ring or 3-ring Venn diagram can serve most of our interests. If there is a need to demonstrate 4 sets at the same time, we could produce multiple 3-ring diagrams and display them in the same page.

CONCLUSION

There are a couple ways to create a proportional Venn diagram. One method is to use SAS/ANNOTATE; the other technique is to employ the PROC GPLOT as presented in this paper. Either method has its advantages and disadvantages. The way with Annotate may be easier to code. But the GPLOT technique is more flexible, although it may be more lines of codes to write. Due to the lengthy of the SAS codes, we can only attach a small sample code in the appendix. For more details of our macros, welcome to contact the author at the following contact information.

CONTACT INFORMATION

Your comments and questions are always valued and encouraged. Please contact the author at:

Shiqun (Stan) Li
Minimax Information Services
(908)240-8229
shiqun@gmail.com

TRADEMARKS

REFERENCES

1. SAS Surveys/Ballots: On “Provide a Procedure to Generate a Venn Diagram”:
   http://support.sas.com/techsup/feedback/sasware_ballot01/ballot01.survey.htm
   http://www2.sas.com/proceedings/sugi26/front/2001_SASwareBallot.pdf
   http://ftp.sas.com/techsup/download/sasware/02results.html
   http://ftp.sas.com/techsup/download/sasware/ballot2003.pdf

2. Shiqun Stan Li: Using SAS® to Create Proportional Venn Diagrams,
   http://support.sas.com/resources/papers/proceedings09/217-2009.pdf

3. Shiqun Stan Li and Wei Zhou: Using SAS® to Create Graphs with Pop-up Functions,
   http://www.lexjansen.com/pharmasug/2009/cc/cc12.pdf

4. Chow, S. and Ruskey, F., *Drawing Area-Proportional Venn and Euler Diagrams*, 11th International Symposium on Graph Drawing, Perugia, Italy, Lecture Notes in Computer Science, 2912 (2003) 466-477.

5. Stirling Chow and Peter Rodgers: Extended Abstract: Constructing Area-Proportional Venn and Euler Diagrams with Three Circles, http://www.cs.kent.ac.uk/pubs/2005/2354/content.pdf

6. Fewell, M. P., Area of Common Overlap of Three Circles. http://nla.gov.au/anbd.bib-an000041411907

7. J. Stoer and R. Bulirsch, Introduction to Numerical Analysis. Springer 2002

8. Alfio Quarteroni, Riccardo Sacco and Fausto Saleri, Numerical Mathematics. Springer 2006

9. SAMPLE: Generating the Venn Diagram, ftp.sas.com/techsup/download/sample/graph/other-venn.html

10. Mike Zdeb, Pop-Ups, Drill-Downs, and Animation, SUGI 29
    http://www2.sas.com/proceedings/sugi29/090-29.pdf

Appendix

Here is a small sample code to produce a proportional Venn diagram. We assume the centers and the radii of the circles are calculated already.

```
/*************************************************************************
 This is a sample code to generate a Proportional Venn diagram with 3-circle,
 given the locations and sizes of the circles.
*************************************************************************/
```

```
%macro points(x0=, y0=, r=, set=, points=50);
%* to generate the points at the circles;
%* center (x0, y0), radius=r, set=set number (1,2 or 3), #points at
circle=points;

*** points around the circle;
set=&set.;
do seq=0 to &points.-1;
  x=&r.*cos(2*3.14159*seq/&points.)+&x0.;
  y=&r.*sin(2*3.14159*seq/&points.)+&y0.;
  output;
end;

*** the center;
set=&set.+3; x=&x0.; y=&y0.; output;
%mend points;

data circles;
  %points(x0=-0.2, y0=0,    r=.5, set=1); * Circle A;
  %points(x0=0.25, y0=0,    r=.3, set=2); * Circle B;
  %points(x0=0.2,  y0=-0.4, r=.2, set=3); * Circle C;
run;
proc sort data=circles;
  by set seq;
run;

goption reset=all hsize=7in vsize=7in ;
*** can adjust the patterns here;
symbol1 interpol=m3n45 cv=black co=black;
symbol2 interpol=m3n135 cv=blue co=blue;
symbol3 interpol=m3n0 cv=purple co=purple;
symbol4 i=none value=dot h=.75 c=black;
symbol5 i=none value=dot h=.75 c=blue;
symbol6 i=none value=dot h=.75 c=purple;
axis1 order=(-1 to 1) major=none minor=none label=none value=none;

ods rtf file='C:\My_SAS\NESUG2009\pVenn_NESUG.rtf';
ods listing close;
proc gplot data=circles;
  plot y*x=set/ haxis=axis1 vaxis=axis1 nolegend;
run; quit;

ods listing;
ods rtf close;
```