

## Sankey Diagrams in SAS® Visual Analytics

Varsha Chawla and Renato Luppi, SAS Institute Inc.

### ABSTRACT

Before the Internet era, you might not have come across many Sankey diagrams. These diagrams, which contain nodes and links (paths) that cross, intertwine, and have different widths, were named after Captain Sankey. He first created this type of diagram to visualize steam engine efficiency. Sankey diagrams used to have very specialized applications such as mapping out energy, gas, heat, or water distribution and flow, or cost budget flow. These days, it's become very common to display the flow of web traffic or customer actions and reactions through Sankey diagrams. Sankey diagrams in SAS® Visual Analytics enable users to easily create meaningful visualizations that represent the flow of data from one event or value to another. In this paper, we consider the components that make up a Sankey diagram: 1. nodes; 2. links; 3. drop off links; and 4. a transaction. In addition, we look at a practical example of how Sankey diagrams can be used to evaluate web traffic and influence the design of a website. We use SAS Visual Analytics to demonstrate the best way to build a Sankey diagram.

### INTRODUCTION

Sankey diagrams have been around since the late 1800s. Captain Matthew Henry Phineas Riall Sankey used a Sankey diagram to visualize steam engine efficiency in 1898. Since then, the diagrams have been named after him. The very first Sankey diagram looked like this:

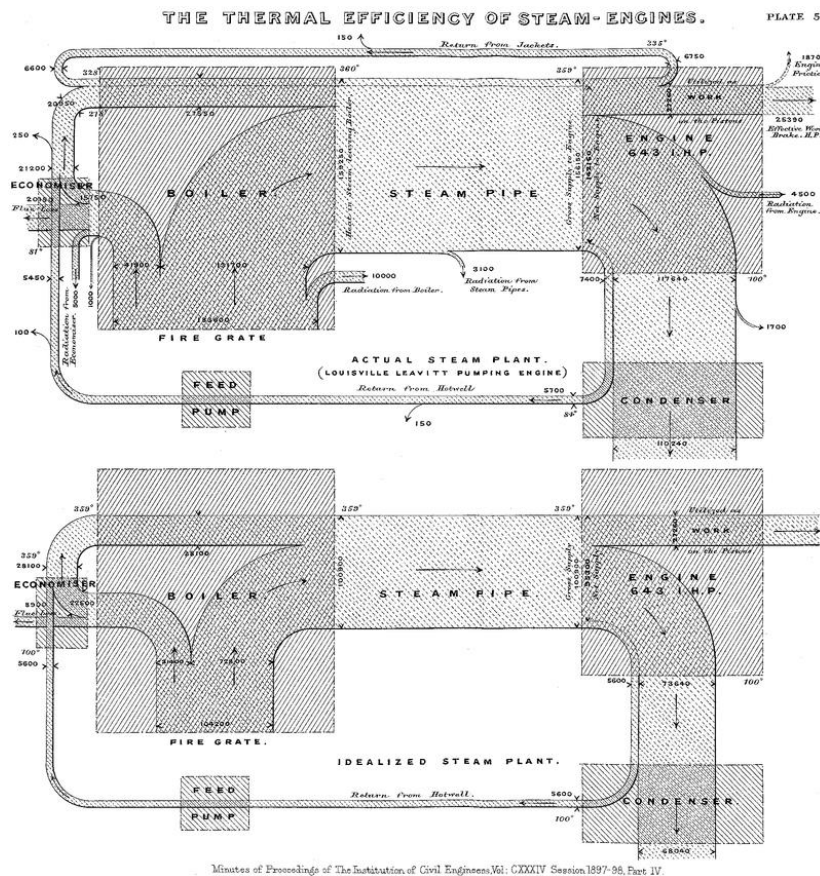


Figure 1. First Sankey Diagram

You might be wondering how Sankey diagrams made their way into modern graphic design and software, such as SAS Visual Analytics. These days, Sankey diagrams actually have many modern business and industry applications. You can use Sankey diagrams to visualize processes such as the following:

- Oil and petroleum flow and production – This can be done for any refined liquids or chemicals.
- Financials and the flow of money from one recipient or account to another.
- Mining, refining, and manufacturing.
- Electricity, coal, natural gas, and coal flow.
- Website and online traffic flow.

We will explore the last example in this paper.

## SANKEY DIAGRAMS IN SAS VISUAL ANALYTICS – THE BASICS

SAS Visual Analytics offers a Sankey diagram that is specifically focused on path analysis. This means that you need data that has some type of flow or sequence to it from one event to another. This type of Sankey diagram is one of the most powerful ones because it enables you to visualize a series of paths and home in on drop off and popular paths. Although Sankey diagrams that you see online might have paths that merge and split, the Sankey diagram for path analysis in SAS Visual Analytics follows a very direct one-way pattern. The Sankey diagram can only be built as an exploration in SAS Visual Analytics. To get started, simply select the Sankey icon from the visualization toolbar in SAS® Visual Analytics Explorer:



Display 1. Sankey Diagram Icon in SAS® Visual Analytics Explorer

### MINIMUM DATA REQUIREMENTS

The minimum data requirements for a Sankey Diagram are the Event, Sequence Order, and Transaction Identifier.

**Event** – This variable always implies some sort of state or action that was performed and will determine the nodes in the diagram. Depending on what you're analyzing, this will vary, but will always need to be a categorical variable in SAS Visual Analytics.

**Sequence Order** – This variable has to be in the form of a datetime stamp or a numerical measure. It is used to determine the order of events within a transaction.

**Transaction Identifier** – This variable uniquely identifies a sequence of events (path). It can be numerical or character based.

The table below shows an example of a single transaction (column TransID) with seven events (column Task) that occurred in the sequence given by column Time.

TransID	Task	Time
25635	Login	1/15/2015 7:35:09AM
25635	Balance	1/15/2015 7:35:15AM
25635	Deposit	1/15/2015 7:35:25AM
25635	Deposit	1/15/2015 7:35:43AM
25635	Deposit	1/15/2015 7:36:07AM
25635	Cash	1/15/2015 7:36:40AM
25635	Logoff	1/15/2015 7:36:51AM

Display 2. Sample Table with Minimum Data Requirements to Build a Sankey Diagram for Path Analysis in SAS Visual Analytics

**Roles** x

**Sankey Diagram**

Use Automatic Chart

Data source: SANKEY\_PATH1

Event

Sequence Order

Transaction Identifier

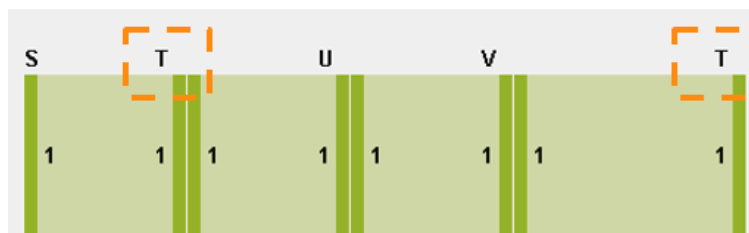
Weight

**Display 3. Minimum Data Requirements Assigned to Sankey Diagram Roles**

There is a fourth data element called Weight, which is optional and will be discussed later in this paper.

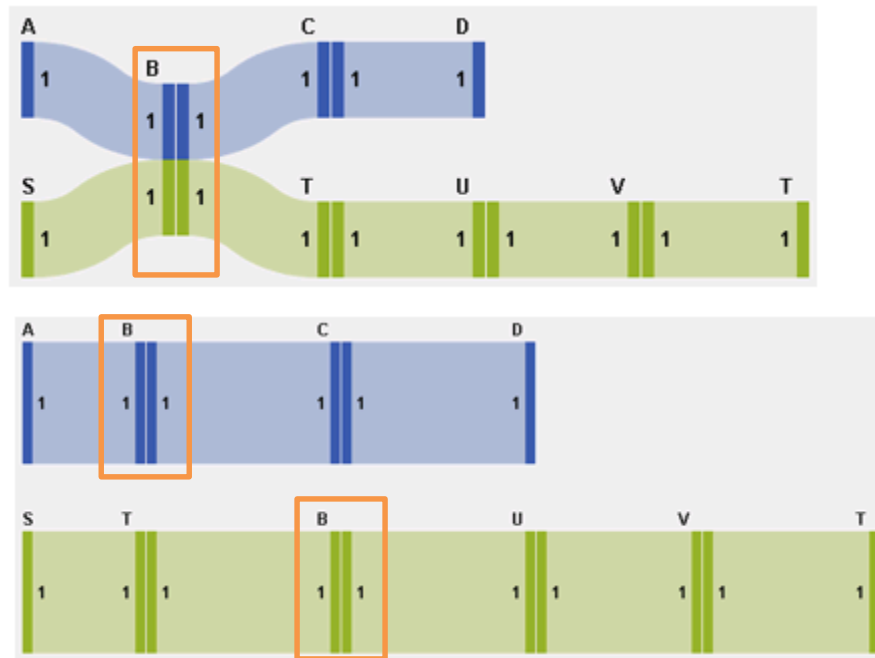
As mentioned before, the Sankey diagram in SAS Visual Analytics is specifically for path analysis. The Sankey diagram in SAS Visual Analytics has unique characteristics when compared to other Sankey diagrams:

- Forward flow – After a path has been determined, it either keeps moving in one direction or it stops. There is no backward movement. As a consequence, the same event can be repeated in multiple nodes.



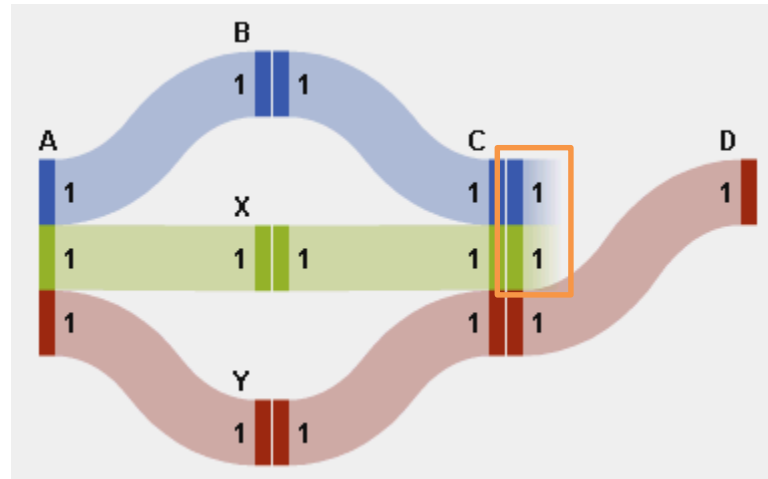
**Display 4. Same Event Repeated in Multiple Nodes**

- Events are displayed only once per position. By position we mean the order in which events occur in transactions (paths). As a consequence, events from multiple paths are merged together in the same node if the events occur in the same order (sequence) in the paths.



**Display 5. Nodes Merged and Not Merged, Depending on the Sequence the Events Occur**

- Drop off is the end of a path. It occurs only when there is at least one additional path that continues beyond the node, as shown in Display 6. If the red path didn't continue from C to D, the drop off wouldn't be marked for blue and green paths.



**Display 6. Drop Off at Event C for Blue and Green Paths**

## THE PROPERTIES TAB

Let's look at some more examples of some very simple Sankey diagrams for path analysis to get a solid understanding of what options SAS Visual Analytics provides. We are going to focus on the **Properties** tab in SAS Visual Analytics to see what options we have for viewing and clarifying the diagram in SAS® Visual Analytics Explorer.

Name: Visualization 1

Title:

☒ Generate graph title

**Properties**

☒ Show labels

☒ Show nodes

☐ Show outline

**Layout**

☐ Vertical layout

**Path Analytics**

Link color: Drop off

Link width: ☒ Frequency ☐ Weight

Minimum path length: 1

Maximum path length: -

Minimum frequency: 1

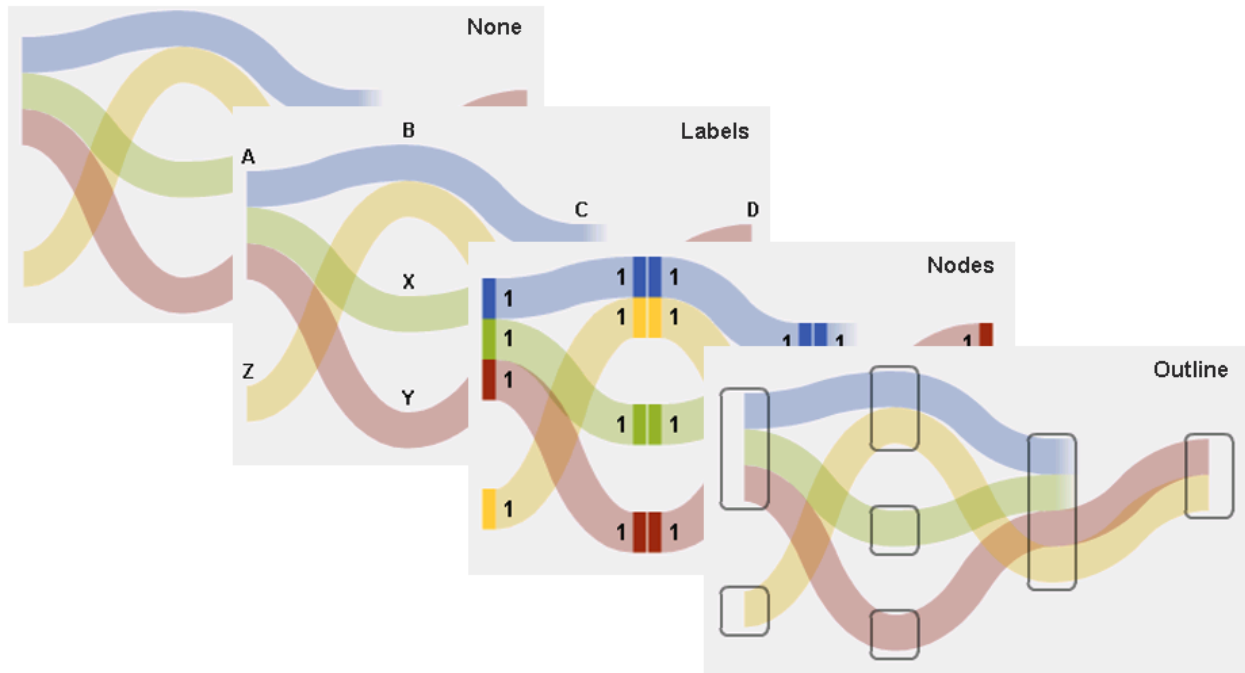
Maximum frequency: -

Path ranking: Top 200

☐ Compress

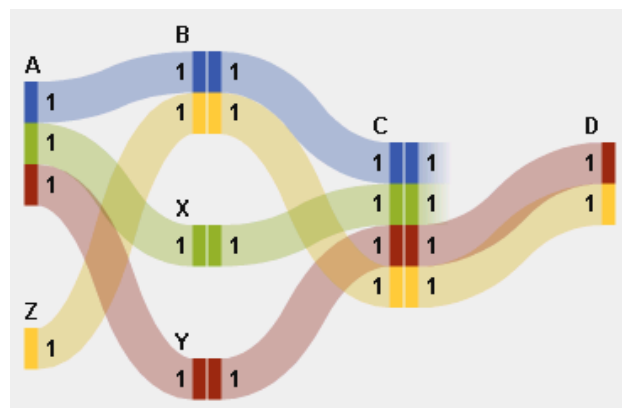
**Display 7. Properties Tab for Sankey Diagram in SAS Visual Analytics Explorer**

Labels, nodes and outline can be turned on and off independently, and they control the amount of information displayed on the diagram.



**Display 8. Labels, Nodes and Outline Controlling the Amount of Information Displayed**

Turning nodes on not only highlights the nodes, but also displays numeric information about the width of the paths. Showing labels and nodes together is considered a best practice because they clearly identify the events (letters) and quantify the paths (numbers). See Display 9 below. For simple diagrams, outlines provide a quick way to get a count of the number of events.



**Display 9. Labels and Nodes Are Visible in the Sankey Diagram**

If you check the **Vertical Layout** option on the **Properties** tab, the workspace flips the diagram for you. This is useful in cases where your paths aren't very long or are similar in length and need to be compared vertically.

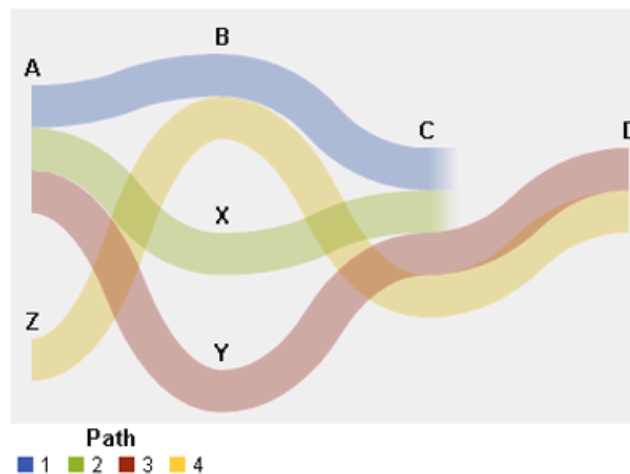


**Display 10. Horizontal and Vertical Layout for the Same Sankey Diagram**

Link color is another property that can help clarify your diagram. There are three options available:

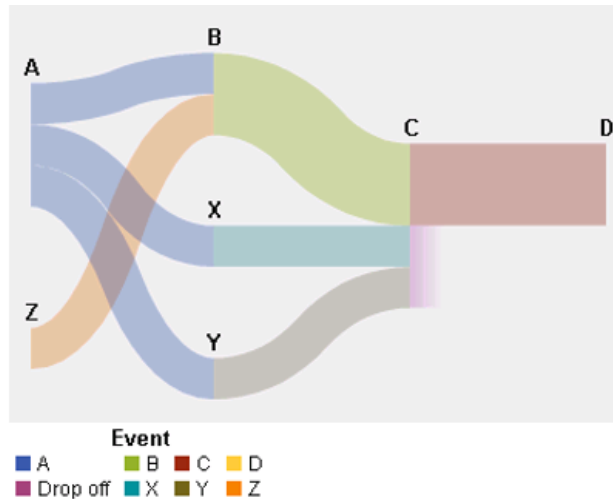
- Path
- Event
- Drop off

Link color for a path is useful when you have a minimal number of paths and want to follow each one separately. Note that if you are looking at 200 paths, for example, the colors will be difficult to distinguish from each other. In Display 11 below, however, each path is easier to follow when the path color option is selected:



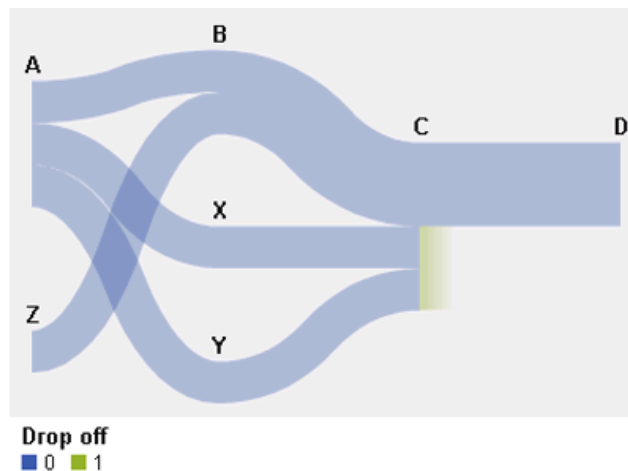
**Display 11. Path Is Selected as the Link Color**

When an event is selected as the link color, each segment coming out of an event is given a different color, as shown in Display 12:



**Display 12. Event Is Selected as the Link Color**

Finally, when you select the drop off option, the drop off links are displayed in a different color. This option makes it easier to see where drop off most frequently occurs: after specific events or after a certain length or depth within a path. Here is what drop off looks like:

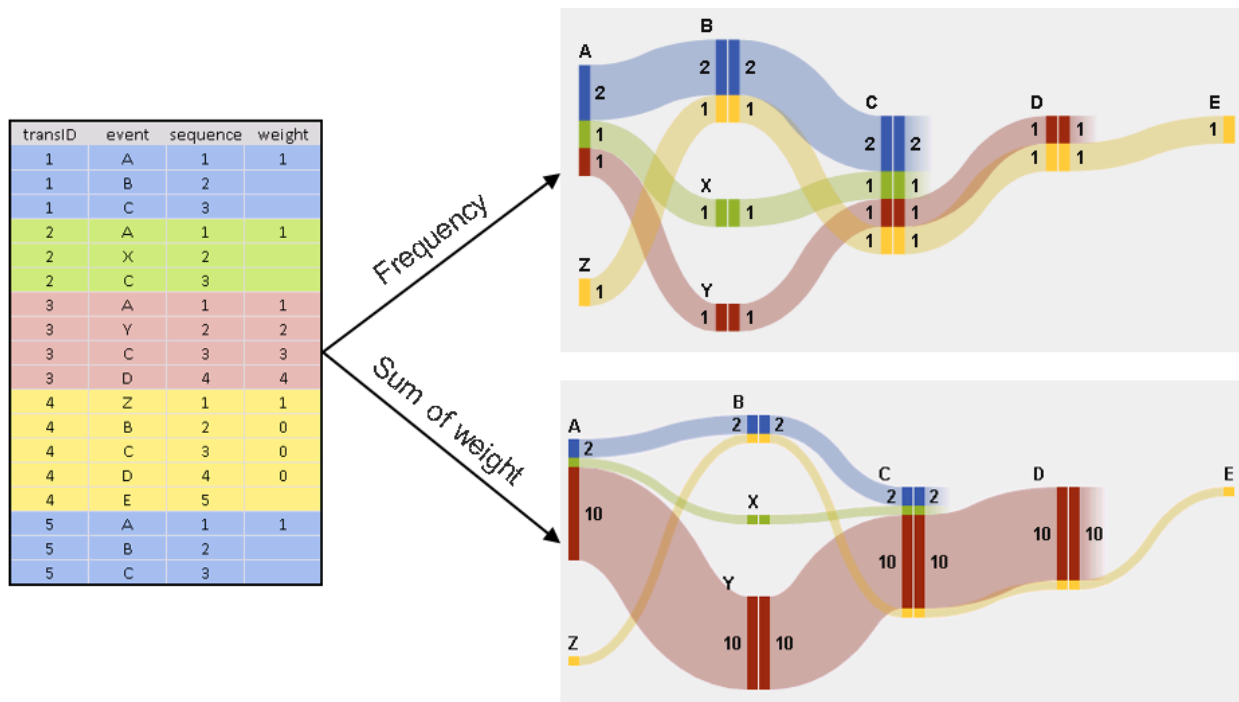


**Display 13. Drop Off Is Selected as the Link Color**

## UNDERSTANDING WEIGHT

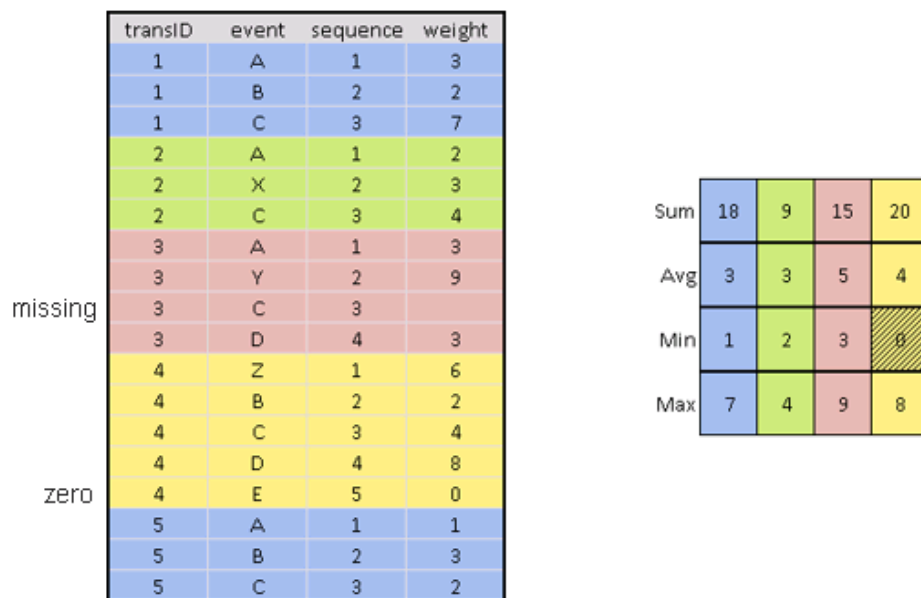
Link width in a Sankey diagram can be determined by frequency or weight. Frequency is the most common and is equivalent to weight equal to 1, while weight is optional and uses a data item as a multiplier factor to the default frequency count. The default frequency count of each link is 1. With frequency, the width of each path simply corresponds to the sum of the frequencies. With weight, four different types of aggregations can be applied: sum, average, minimum, and maximum. Here is a visual example of this:



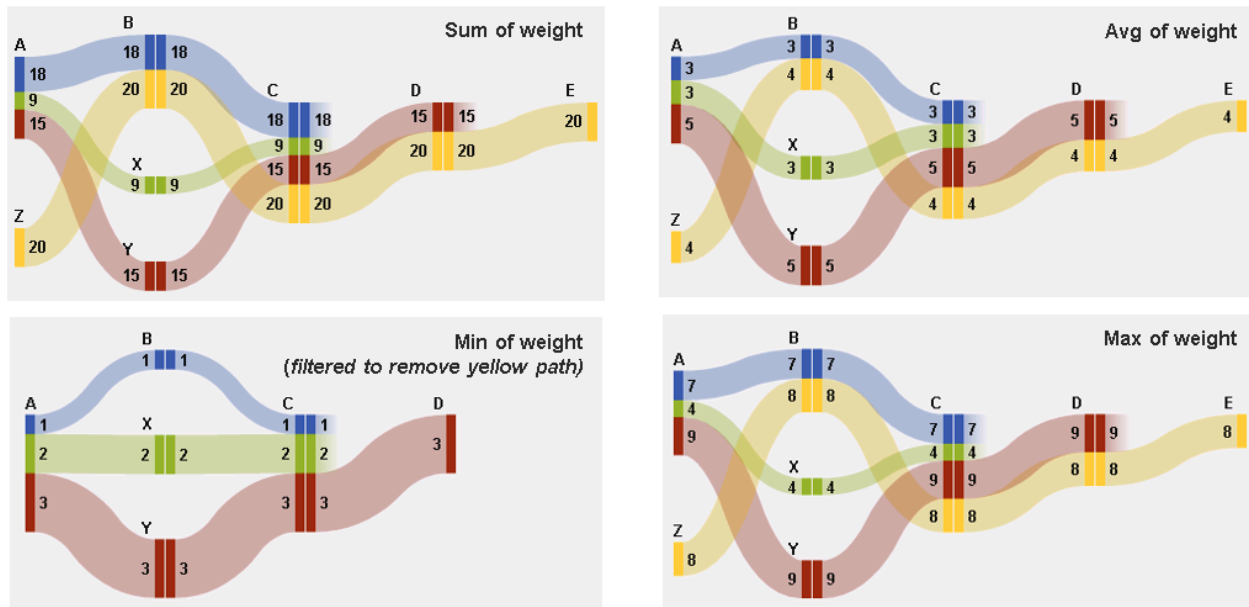


**Display 14. Difference between Frequency and Weight on the Link Width**

It's important to keep in mind that missing values and zeros are not the same. Replacing one by the other will affect the average and possibly minimum aggregations. The next two displays show you what the sum, average, minimum, and maximum aggregations do to the path width and how they are affected by missing and zero. Observe that a path of width zero cannot be displayed, and this is exactly what happens in the yellow path for aggregation minimum. Compare that with the red path where instead of zero there is a missing value and how it affects the average. Each path is defined by a different color here:

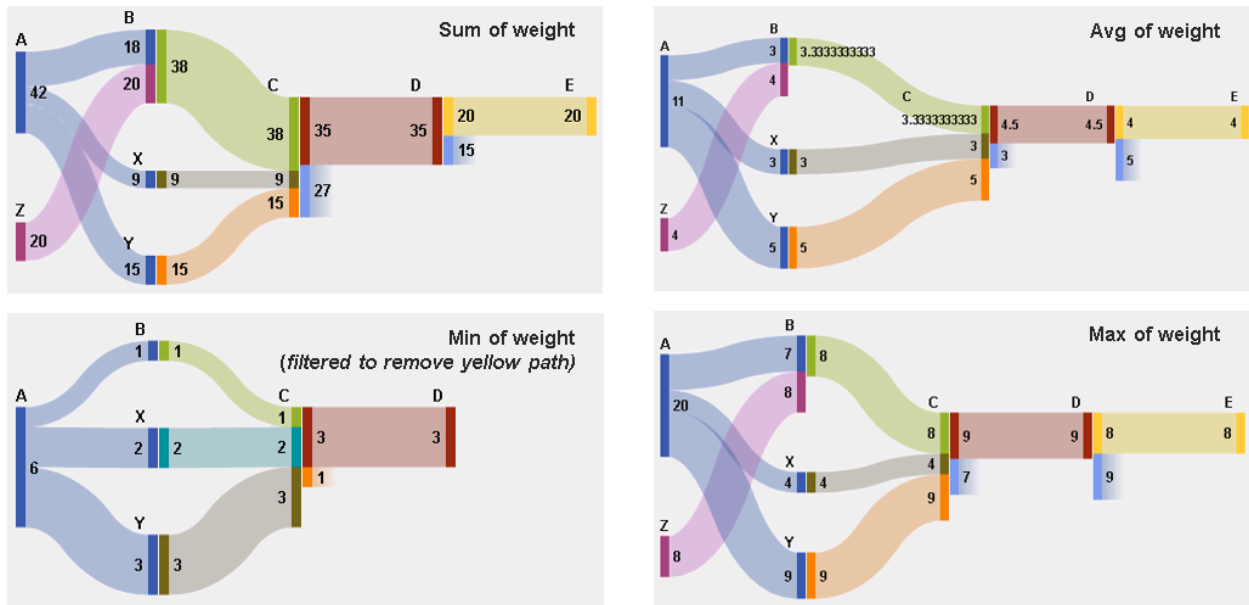


**Display 15. Data Table That Is Used to Create Each Path in the Next Display**



Display 16. Path Width Based on Different Aggregations of Weight and Colored by Path

Sankey diagrams are colored by path in the previous example. If they are colored by event, the results are different as shown in Display 17. Aggregations for Sankey diagrams colored by event are not so straightforward to understand. For example, the average between events B and C is 3.333. We get to this value because in that link we have two transactions with average 3 and one transaction with average 4. The average therefore is  $(3+3+4)/3$ , which is 3.333.



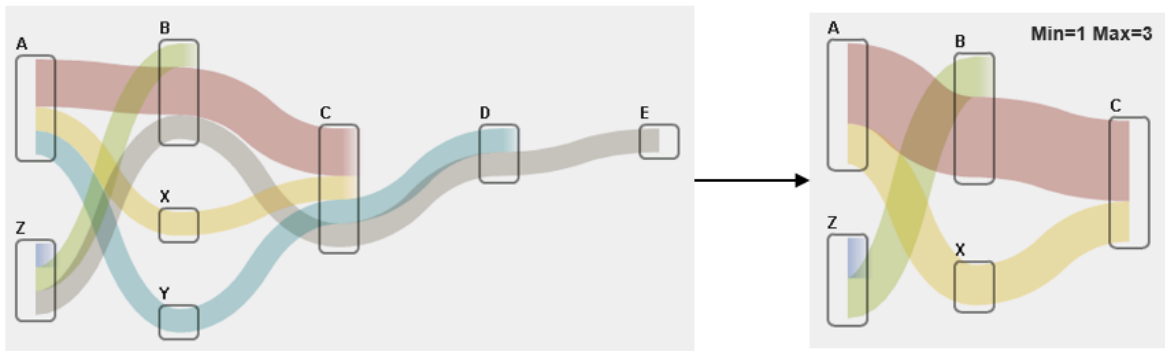
Display 17. Path Width Based on Different Aggregations of Weight and Colored by Event

Other options on the **Properties** tab have to do with the minimum and maximum path length as well as frequency. The numbers selected here reflect the number of events. Both the minimum and maximum can help the user home in on paths that meet these criteria. This is especially useful when you are working with a high number of paths.

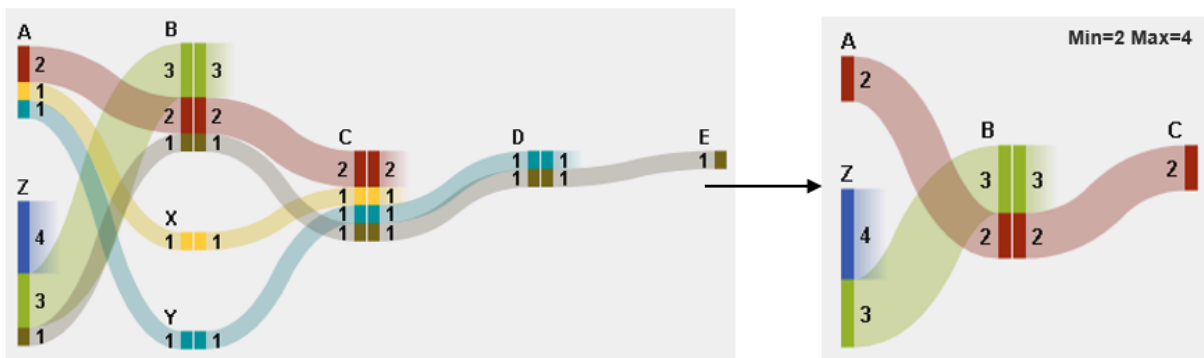
Minimum path length:	7	▲ ▼
Maximum path length:	15	▲ ▼
Minimum frequency:	5	▲ ▼
Maximum frequency:	10	▲ ▼

**Display 18. Minimum and Maximum Path Length and Frequency**

Here are two examples of what happens if you set the minimum and maximum path length and frequency:



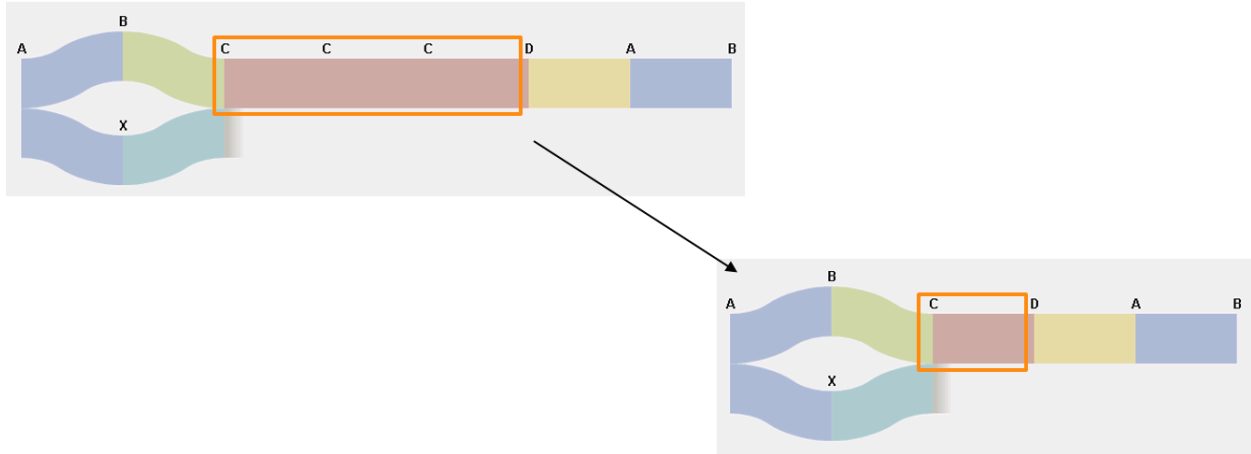
**Display 19. Diagram Reduced to a Minimum Path Length of 1 and a Maximum of 3**



**Display 20. Diagram Reduced to a Minimum Frequency of 2 and a Maximum of 4**

Path ranking in SAS Visual Analytics is equally straightforward. Paths can be ranked by either top or bottom N frequency or weight. In short, this means that you will see only the top or bottom N paths in the Explorer window.

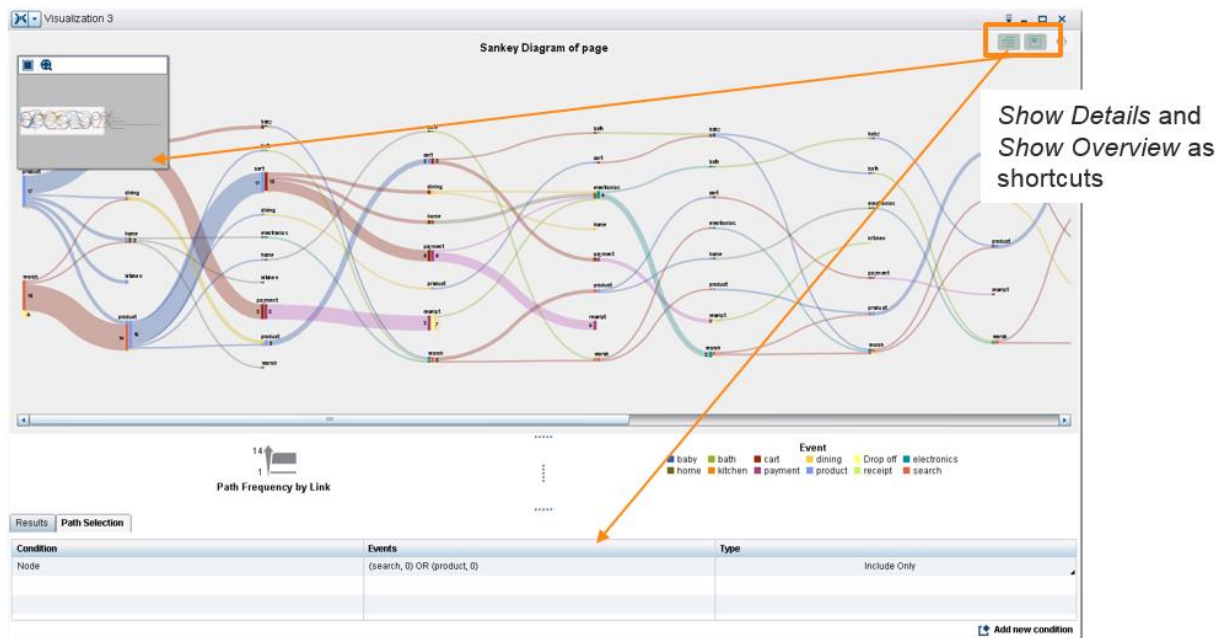
The **Compress** option at the bottom of the **Properties** tab is useful for combining events that appear sequentially into just one node. This saves valuable path space since it shortens the length of the path without altering the sequence. Display 21 explains this concept:



**Display 21. After the Compress Option Has Been Checked, Event C Is Condensed from Three to One**

## THE POP-UP MENU

So far, we have covered many of the properties that can also be found by right-clicking a Sankey diagram, on either a path or a node (event). The options **Show Details** and **Show Overview** are not found on the **Properties** tab. They are available on the pop-up menu as well as shortcuts on the top right of the diagram.



**Display 22. Show Details and Show Overview Options**

The **Show Overview** option is extremely useful because it enables you to zoom in and highlight a certain area of the Sankey diagram. This is a must for large diagrams.

**Show Details** brings two tabs at the bottom of the diagram. The **Results** tab can be used to see events, paths, frequencies, and weights in tabular format, as well as highlight elements on the diagram. The **Path Selection** tab shows a summary of selection conditions currently applied to the diagram, as well as management tools that enable you to add, remove, and change the type of current selection conditions. Conditions are presented next.

Results		Path Selection					
Type	Event	Path	From Node	To Node	Path Frequency	Path Weight	
Link	Drop off	1	(C, 2)	Drop off	2	18	
Link	Drop off	2	(C, 2)	Drop off	1	9	
Link	A	1	(A, 0)	(B, 1)	2	18	
Link	A	2	(A, 0)	(X, 1)	1	9	
Link	A	3	(A, 0)	(Y, 1)	1	15	
Link	B	1	(B, 1)	(C, 2)	2	18	

Results		Path Selection		
Condition	Events			Type
Node	(B, 1)	Remove Selected Conditions Remove All Conditions Create Visualization from All Conditions		Include Only
Starts with	(A)			Exclude

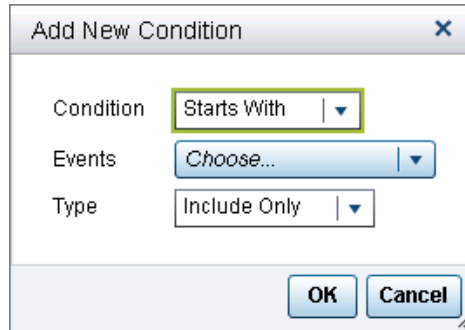
**Display 23. Results and Path Selection Tabs Provided by Show Details Option**

Here are four more options that can help you gain more insights when working with a path analysis diagram:

Add Category	▶
Add Measure	▶
Remove	▶
Rename...	
Comment...	
Duplicate	
Delete...	
Export Image...	
Export Data...	
Show Details	
Chart Type	▶
Show Overview	
✓ Show Labels	
✓ Show Path Widths	
Show Outline	
Vertical Layout	
Add New Condition...	
Include Only	▶
Exclude	▶
Create New Visualization	▶

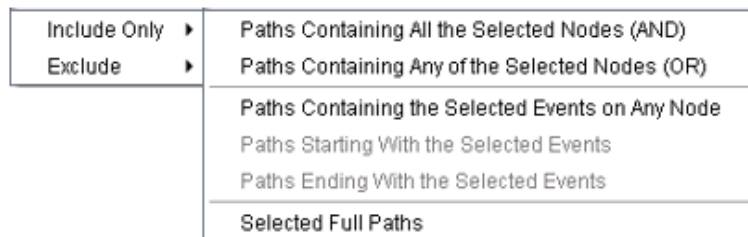
**Display 24. Pop-Up Menu Options Not Found in Properties**

When you select **Add New Condition**, you get the option to include or exclude paths that start, end with, or contain certain events. The events that are part of your diagram will be listed in the drop-down menu.



**Display 25. Add New Condition Menu Options**

The **Include Only** and **Exclude** options, as well as the **Create New Visualization** option, enable you to focus on just parts of the diagram. This means that you can quickly find and filter by certain nodes (events) or paths of interest. Note that in order to use these options, you must right-click on a node in a diagram, and because this menu is context sensitive, certain options can be disabled.



**Display 26. Focus / Narrow Down Options**

## A WEB DATA EXAMPLE






As previously mentioned, Sankey diagrams in SAS Visual Analytics were specifically developed for path analysis. One of the best applications for a path analysis diagram is analyzing web data. Now that you're familiar with the basic data requirements and properties of Sankey diagrams, let's look at a fictitious website data example to see how the data is set up and what insights we can derive from using path analysis.

The table we are going to analyze was kept simple for educational purposes. However, the data is very similar to web data that you might find at a large retailer with multiple pages on their site. These pages will be used to identify the events. We created pages for several departments: home page, baby, bath, dining, electronics, and kitchen. We also created product-specific pages, search results, a shopping cart, a purchase page, and a receipt page.

For each visited page, we defined a unique session ID, page ID, timestamp (the start time of when the user entered the page), and duration (time spent on a given page). We also defined some criteria for following through on a purchase. Every purchase has to go through the shopping cart page. After a payment has been made there, a receipt page is shown as confirmation of the purchase. The purchase amount is shown on the receipt page as well.


Let's look at what a section of the data table for this web data looks like. Below is a partial table view that was created in SAS® Enterprise Guide. SAS Visual Analytics supports a wide variety of data sources and databases. Real-life web data does need to be collected through a different type of solution or product.


As you can see in the table below, session ID 885 shows the sequence of pages visited (first home, then search, product, cart, payment, and receipt) that resulted in a purchase of \$132.92.


	 session_id	 time	 page	 duration	 purchase
5520	884	07JAN14:05:17:19	home	9	.
5521	884	07JAN14:05:17:28	electronics	5	.
5522	884	07JAN14:05:17:33	search	14	.
5523	884	07JAN14:05:17:47	product	44	.
5524	885	03JAN14:10:16:52	home	25	.
5525	885	03JAN14:10:17:17	search	12	.
5526	885	03JAN14:10:17:29	product	20	.
5527	885	03JAN14:10:17:49	cart	9	.
5528	885	03JAN14:10:17:58	payment	48	.
5529	885	03JAN14:10:18:46	receipt	4	\$132.92
5530	886	04JAN14:12:10:25	home	12	.
5531	887	10JAN14:00:30:41	home	7	.
5532	888	15JAN14:23:28:50	home	22	.
5533	888	15JAN14:23:29:12	search	13	.
5534	888	15JAN14:23:29:25	product	57	.
5535	888	15JAN14:23:30:22	electronics	7	.
5536	888	15JAN14:23:30:29	search	9	.
5537	889	06JAN14:07:31:19	home	18	.

**Display 27. Partial Table View**

In SAS Visual Analytics Explorer, you can see that the following columns were assigned to the three required data categories:

Event  
 page

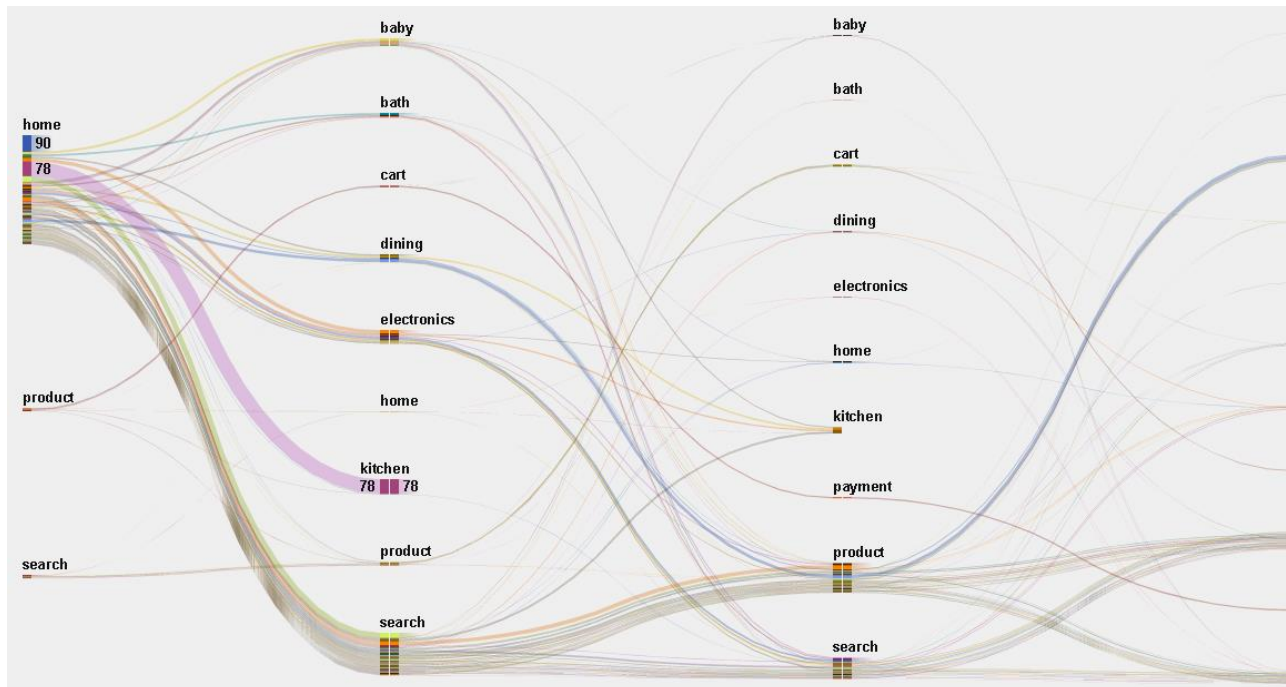
Sequence Order  
 time

Transaction Identifier  
 session\_id

**Display 28. The page, time, and session\_id Are Assigned to Event, Sequence Order, and Transaction Identifier**

The resulting Sankey diagram can be found below. Note that we decided to color the diagram by paths in order to distinguish the different pages from each other.





**Display 29. Initial View of Website Paths**

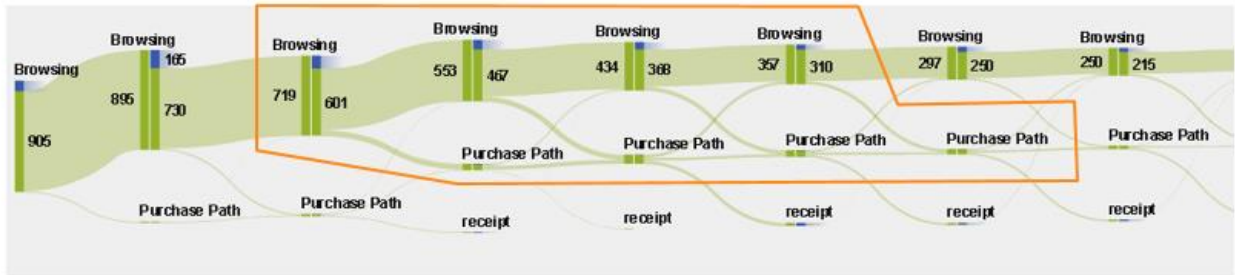
What makes path analysis with Sankey diagrams so amazing is that we can draw a few important insights by just looking at this diagram, without doing any math or crunching numbers ourselves. Here is what we noticed:

- Most sessions start from the home page, but we see other sessions that start from the product and search pages as well. This could mean that visitors are either searching for a specific product by entering the product page or landing on the site via organic search results.
- When you look at how busy those nodes are, it's apparent that search is heavily used by the visitors, because we see many paths running through the search pages.
- Drop off (the point where people abandon the transaction, or the web site in this example) is apparent for the kitchen page. This might be something to look into because it could signal a page error, lack of interest in the products, site usability and navigation issues, or another problem.

Let's take a closer look at how visitors are using our site. We decided to simplify things by categorizing each page into one of two groups: Browsing and Purchase Path. In SAS Visual Analytics, you can use a custom category to group items together. Here is the breakdown of what falls into each category:

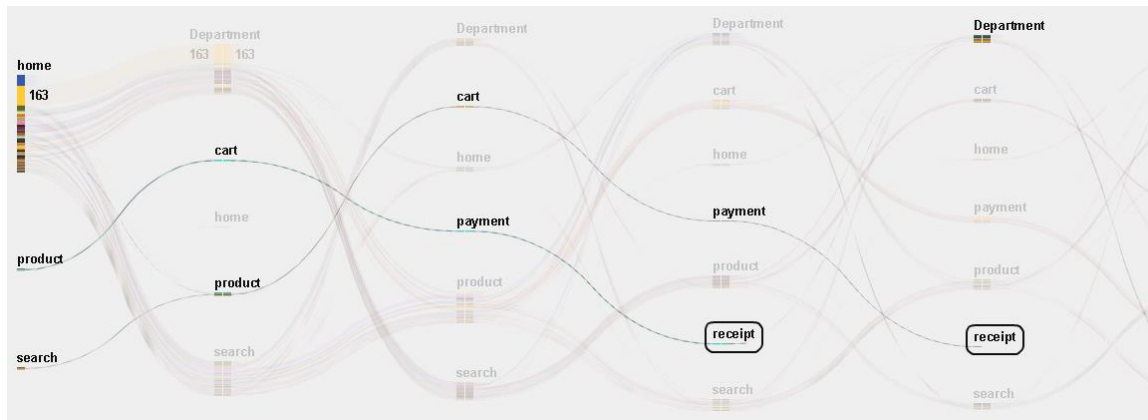
- Browsing – includes home and department pages, as well as all the search and product pages.
- Purchase path – contains cart and payment pages.
- Receipt page – We left this alone since it indicates a purchase, so it is best to view it by itself.





**Display 30. Sankey Diagram with Two Custom Categories**

The display above is the result of creating the different custom categories. It's easy to see that a lot of time is spent browsing, while only some browsing behavior ends in a purchase. This is expected and fairly normal for a retail website. After 2 to 5 browsing events, we see the thickness of the purchase paths increase. This tells us that it takes some browsing for most people to make a purchase. Most people like to do some research before they buy. We also notice that some people continue to browse after a purchase, which is not a bad action, and drop off (indicated by the blue, fading lines) is fairly minimal.

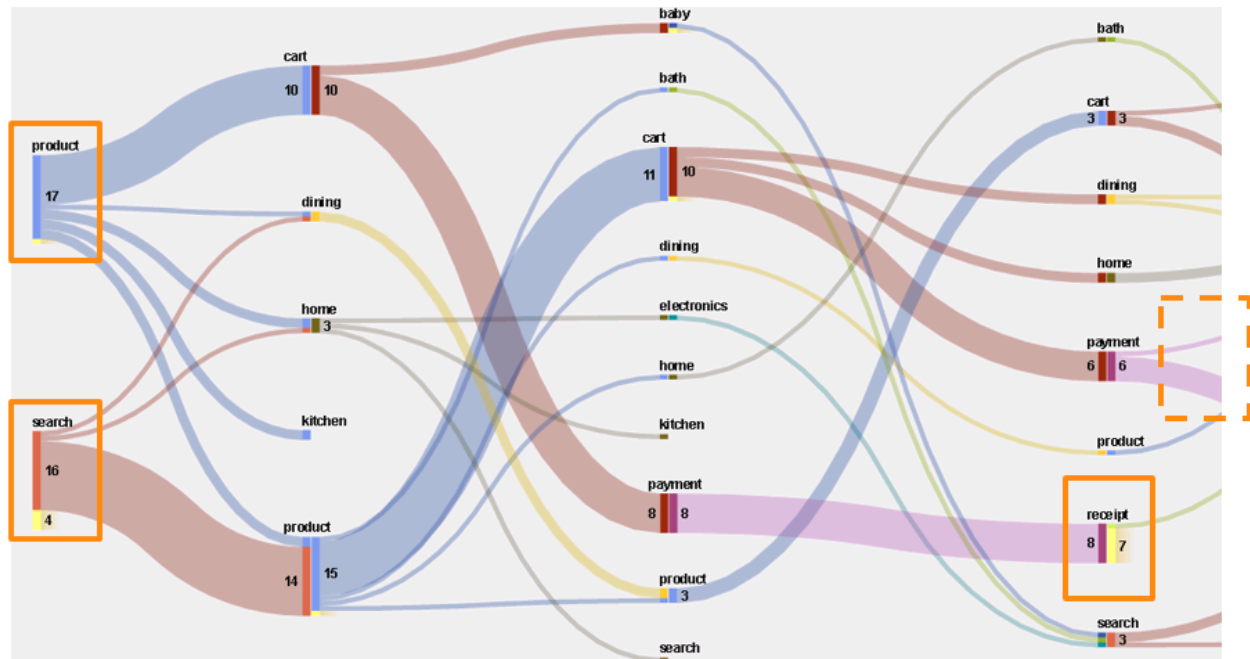


**Display 31. Sankey Diagram with One Custom Category - Department**

Display 31 contains a different view of the diagram. We grouped all department pages into one custom category called Department. We also compressed the department-to-department links to get rid of multiples and then colored these by path. The main advantage in doing this is that we get a high-level view of the paths without the granularity of the department-level data. The two paths that you see highlighted are the first two shortest paths that led to purchases.

What types of insights can we draw from this view of the diagram? Well, it looks like the first two shortest paths came from shortcuts to the product and search pages since they did not go through the home page of the site. This tells us that the visitors already knew what they wanted, or were fairly close to knowing what they were looking for. The paths quickly converge on payment and then receipt, indicating to us that this group of people most likely visited the site with the intent to purchase.

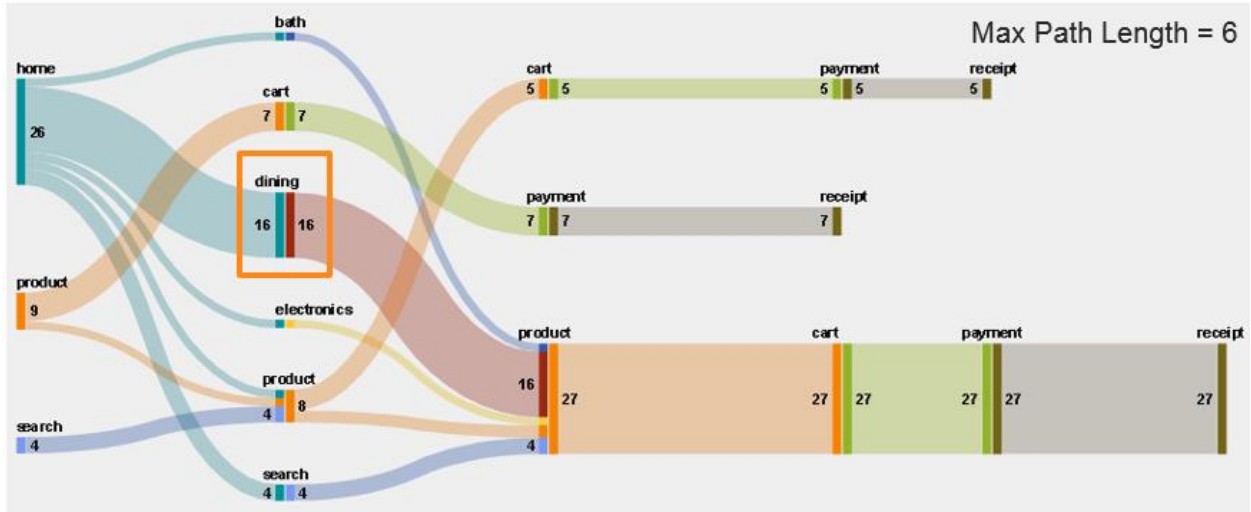
Let's home in on product, search a bit more, and then analyze the original diagram without compression and switching to color by event. We also used the **Include only** menu option to limit the diagram to paths that start with product and search pages.



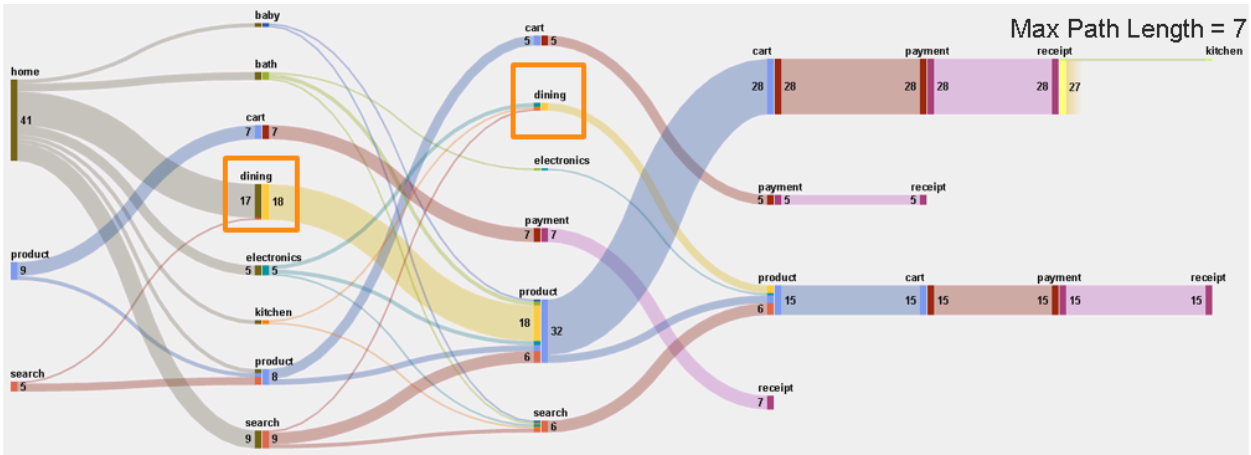
**Display 32. Product and Search Paths Focus**

What's most apparent from this view of the diagram is the high convergence rates, which translate into fairly high purchase rates. Eight out of the 18 visits that started on a specific product page ended up making a purchase. When we start from search, we notice that 5 out of 20 visits are resulting in a purchase. Some actionable activities based on these insights might include ensuring that the product and search pages have easy navigation and usability, and include product suggestions or recommendations, in order to promote cross-sells and up-sells.

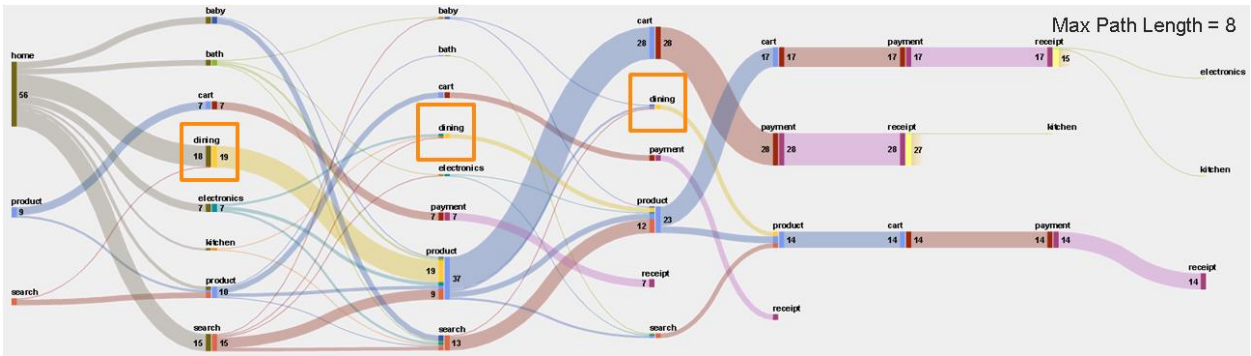
We're now going to analyze all paths that led to a purchase, by including those paths with the receipt node or event. We're also limiting the path length to 6, 7, and 8, in order to find out if there is a difference in convergence rates from the shortest to longest purchase experiences. From analyzing Displays 33, 34, and 35, it becomes apparent that dining is the only department that consistently converges in a sale (receipt event). This prompts us to look into the dining page a bit more (outside of the Sankey diagram). Some factors we might want to analyze include the promotions and any advertisements on this page, the page design, products offered, navigation and menu items, as well as the load speed and links to this page from other pages. Since dining is doing so well from a sales standpoint, we might not want to change anything on this page, but we might want other department level pages to follow a similar format or offer the same types of promotions. Much food for thought.



Display 33. Paths with Receipt Nodes Where Max Path Length = 6

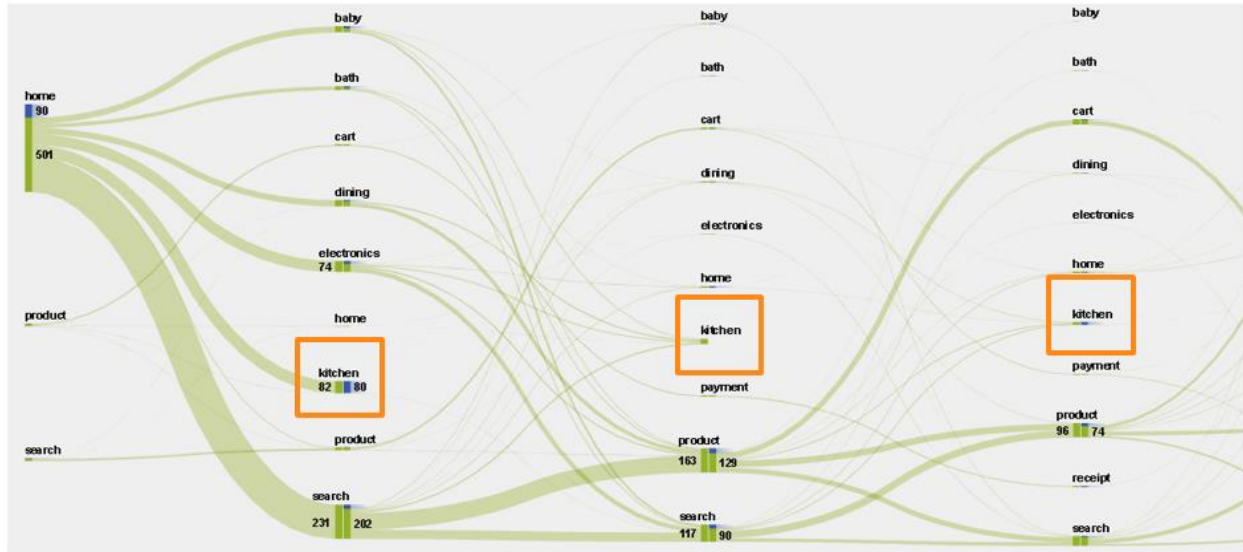


Display 34. Paths with Receipt Nodes Where Max Path Length = 7



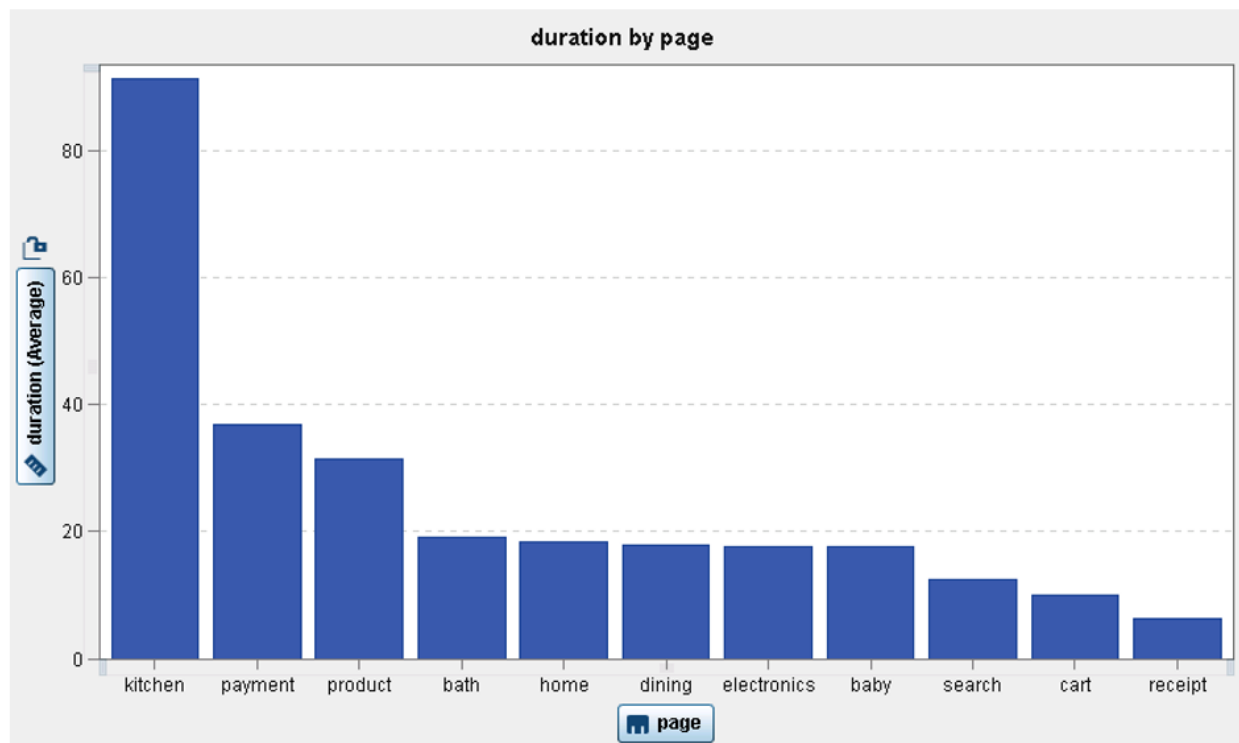
Display 35. Paths with Receipt Nodes Where Max Path Length = 8

We're going back to analyze drop off in our original diagram. All we have to do for this is to turn off the previous filters and switch color to drop off. Here is what we get:



Display 36. Drop Off in the Original Diagram

We don't see much drop off overall, but what we do see is some drop off in the kitchen category. This could be due to several factors, such as a poorly designed page, errors on the page, or poorly placed or designed products. The page should be analyzed from both a product and web perspective. We selected **Create New Visualization** from this diagram to see if there is a possible problem with page load times. We then selected the duration variable to see what the time spent on each department page is like. This resulted in the following bar chart:



Display 37. New Visualization Showing Duration on Each Page

What an insight into duration we have now! The average time spent on the kitchen page is more than double that of any other page. Definitely an indication that visitors either can't quite decide what to

purchase, do more reading on the kitchen page, or find something wrong with the page such as a slow loading time or time-out.

## CONCLUSION

The analysis of different views and the use of the different options found in SAS Visual Analytics demonstrates that Sankey diagrams for path analysis can be extremely beneficial in analyzing single paths of event flow. In an age of web page after web page, this can be especially useful for analyzing web content. You might find many Sankey examples online, but it is crucial to remember that the ones in SAS Visual Analytics are specifically designed for path analysis. This means that you will always need to have events, a sequence, and a transaction identifier to build your diagram.

We would love to hear about what insights you uncovered by building these diagrams using SAS Visual Analytics. Feel free to reach out to us or find us on the [SAS Visual Analytics Community](#) to learn more about SAS Visual Analytics and everything that comes with it.

## REFERENCES

“Sankey Diagram.” In Wikipedia. December 29, 2014. Available at [http://en.wikipedia.org/wiki/Sankey\\_diagram](http://en.wikipedia.org/wiki/Sankey_diagram).

## RECOMMENDED READING

- *SAS Visual Analytics 7.1: User's Guide*

## CONTACT INFORMATION

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

Varsha Chawla  
SAS Institute Inc.  
919-531-1446  
[varsha.chawla@sas.com](mailto:varsha.chawla@sas.com)

Renato Luppi  
SAS Institute Inc.  
919-531-1650  
[Renato.luppi@sas.com](mailto:Renato.luppi@sas.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.