Paper 135-2017

Sankey Diagram- A Compelling, Convenient, and Informational Path Analysis with SAS® Visual Analytics

Abhilasha Tiwari, Accenture, Netherlands

ABSTRACT

SAS® Visual Analytics provides a complete platform for analytics visualization and exploration of the data. There are several interactive visualizations such as charts, histograms, heat maps, decision tree, and Sankey diagram. A Sankey diagram helps in performing the path analytics and offers a better understanding of complex data. It is a graphic illustration of flows from one set of values to another as a series of paths, where width of each flow represents the quantity. It is a better and more efficient way to illustrate which flows represent advantages and what flows are responsible for the disadvantages or losses. Sankey diagrams are named after Matthew Henry Phineas Riall Sankey, who first used this in a publication on energy efficiency of a steam engine in 1898. This paper begins with information regarding the essentials or parts of Sankey: nodes, links, drop-off links, and path. Later, the paper explains the method for creating a meaningful visualization (with the help of examples) with a Sankey diagram by looking into the data roles and properties, describing ways to manage the path selection, exploring the transaction identifier values for a path selection, and using the spotlight tool to view multiple data tips in SAS Visual Analytics.

INTRODUCTION

Sankey diagram provides the illustration of different kinds of flows like: energy, material or money. In other words, it provides the summary of all the path involved in a process. Sankey diagram are best used to show the mapping between different domains or for identifying the different paths involved in a process.

Sankey diagrams are named after an Irishman- Matthew Henry Phineas Riall Sankey. He first used it to show the energy efficiency of a steam engine in 1898 in a publication.

In Sankey diagram, thickness of the line or arrow represent the quantity, so thicker the arrow, larger is the quantity and vice versa. Direction of the arrow points towards the flow direction. This is illustrated and explained in the figure below.

Figure 1. Simple illustration of Sankey diagram depicting the relation between the width and direction of the arrow with number of satisfied and dissatisfied customers from number of call received.

Additionally, different colors can also be used to indicate different categories or to show the different transition state or path involved in a process.

ESSENTIALS OF SANKEY DIAGRAM

MAIN COMPONENTS

Sankey diagram mainly consist of following:

Nodes – It represents the events of each path. The node refers to the width of link that enters and exit the path in the diagram.

Links – It represents the path in diagram. The link refers to the distance between two nodes and thickness of each link represents the frequency of the path or value of the measure.

Drop off links – It represents the end of the path. The drop off links refer to the link that end at current node. In other terms, it occurs only when at least one of the path continues beyond the node in the diagram.

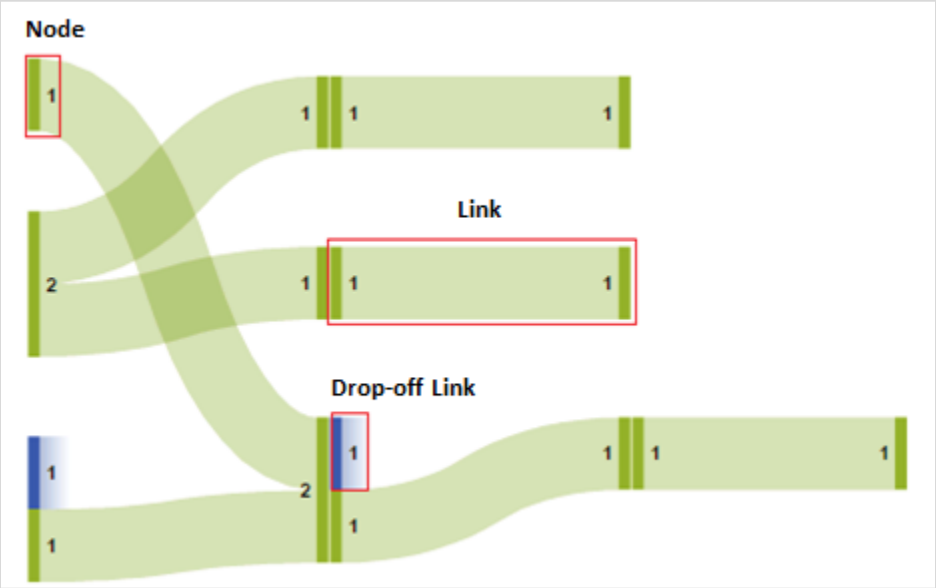


Figure 2. Essentials of Sankey Diagram

DATA REQUIREMENTS IN SAS VISUAL ANALYTICS

Event: It requires the variable to be a category. These values are represented as nodes in Sankey diagram.

Sequence Order: It requires the variable to be a date-time or a measure. These values define the order of the events within each transaction.

Transaction Identifier: It requires the variable that is either numeric or a category. These values identifies the sequence of each event.

Let's consider a very simple data set to have better understanding of main components and data requirements for Sankey diagram.

Customer	Call Subject	Date
1	Bill	01 January 2016
1	Recharge	02 January 2016
2	Free trial	01 January 2016
2	Free trial	02 January 2016
2	Product Purchase	03 January 2016
3	Product Purchase	01 January 2016
3	Recharge	02 January 2016
3	Online payment	03 January 2016
3	Bill	04 January 2016
4	Free trial	01 January 2016
4	Product Purchase	02 January 2016
4	Bill	03 January 2016
5	Product Purchase	01 January 2016

Figure 3. Sample dataset of a customer journey.

The dataset contains the information for five Customers regarding the type and order of call made by them.

For this paper, "Call Subject" data is used as "Event" for identifying the nodes. "Date" data in "Sequence Order" to know the order and "Customer" data in "Transaction Identifier" to identify the sequence of each event.

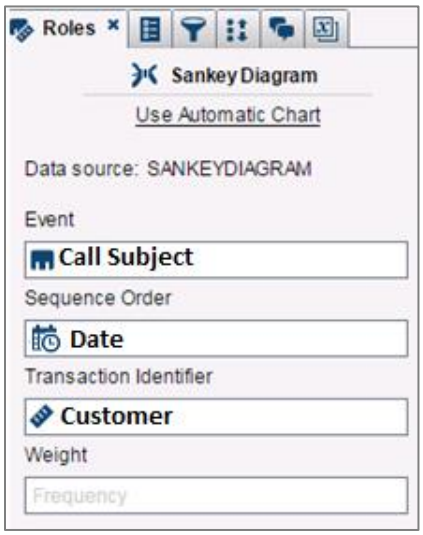


Figure 4. Data requirements filled in Role tab.

With the above mentioned roles, below is the Sankey diagram, depicting the paths followed by each of the customer.

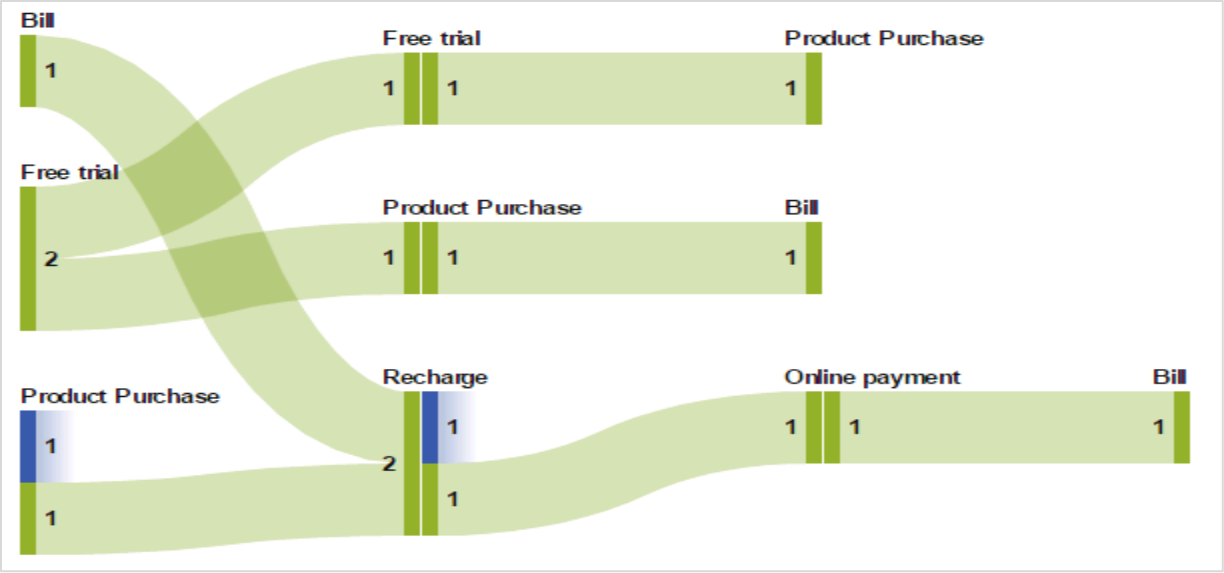


Figure 5. Sankey diagram showing the path of each Customer.

Each node represents the event, and number associated with the node represent the quantity of the event.

Let's look at each node briefly to understand the information associated with it.

Node 1 – On 1st January 2016, only one customer enquired about "Bill" information and next day same customer enquired about 'Recharge'.

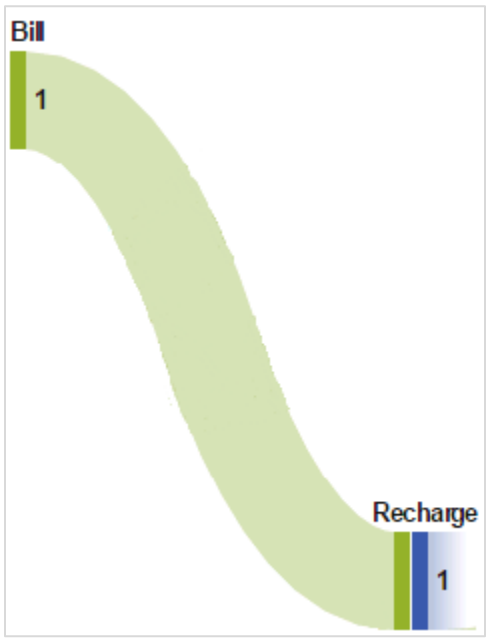


Figure 6. Path analysis for Node 1

Node 2 – Two customer's from five inquired for "Free Trial" on 2nd January 2016. Later both of them followed two different paths.

One of the customer inquired next day again for 'Free trail' and then for 'Product Purchase'.

While the other customer inquired on 'Product Purchase' and 'Bill'.

Figure 7. Path analysis for Node 2

On the basis of Event - Call subject, path analysis of Node 2 depicts that most of the calls were made for free trial. Also, its most likely that if customer goes for the free trial, will also go for the product purchase.

Node 3 – Two of the customer inquired firstly on Product purchase and took different paths.

It's clear from the diagram below, that one of the customer did not go ahead with any other inquiry after product purchase, but the other one did ask for recharge, online payment and bill over the period of time.

Figure 8. Path analysis for Node 3

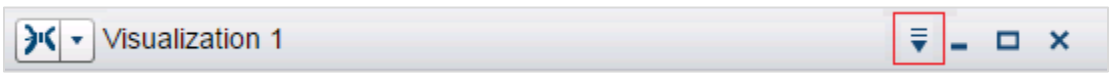
EFEECTIVE AND EFFICIENT WAYS TO MANAGE PATH SELECTION IN SANKEY DIAGRAM

Path selection is an efficient way to manage path analysis by selecting or deselecting the nodes or events of interest. It is of great help when there is noise in the data or need to focus on some particular events.

Path selection, or filtering can be managed by creating a new condition either using 'Options' functionality or from node selection. Both methods for path filtering are briefly discussed below:

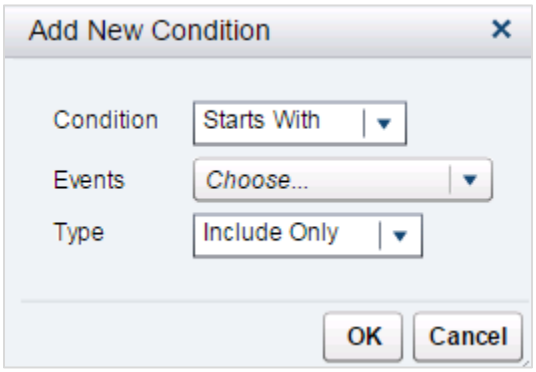
CREATE A NEW CONDITION FROM OPTIONS

- To create a new condition, click "Options" icon in the visualization toolbar.



- From the drop-down list available, select "Add New Condition"

- Add New Condition window appears with three options - Condition, Events and Type.



- Select the conditions of interest or as required and click "Ok".

The Type "Include Only" allows the path to be selected on the basis of the selected event. Whereas the Type "Exclude" allows the path to be removed from path analysis on the basis of the selected event.

In order to demonstrate the functionality of path selection via adding new conditions, below mentioned conditions are used in this paper.

Condition: Start With

Event : Free trial

Type : Include only

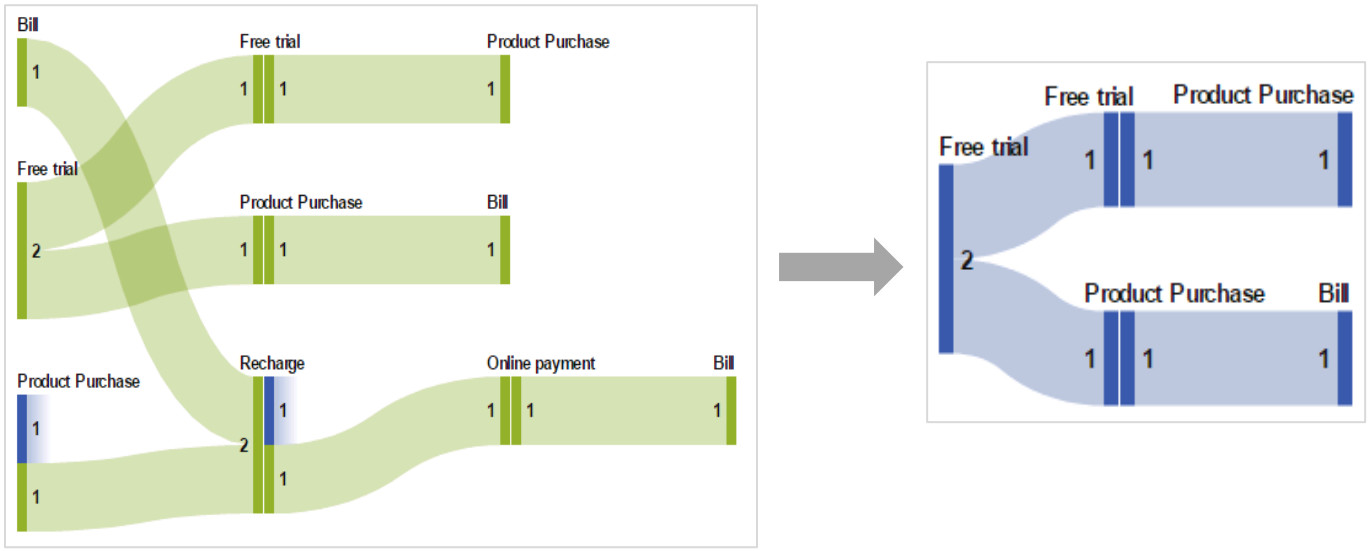


Figure 11. Result of path selection for Event - Free Trial

The Sankey diagram obtained through path selection with above conditions, has only event that start with "Free Trail". In this way it's very convenient to narrow down the path analysis on some particular events of interest. Also, it gives the insight on number of events starting with "Free Trial" and thus this analysis can help in making strategic decisions for business.

CREATE A NEW CONDITION FROM THE SELECTED NODE

- To create a new condition from the existing nodes in path analysis, select one or more nodes in

the diagram. Multiple nodes can be selected with Ctrl key.

Figure 12. Selection of Node from Sankey diagram

- Right click on the selected node and navigate through the options to either “Include Only” or “Exclude”.

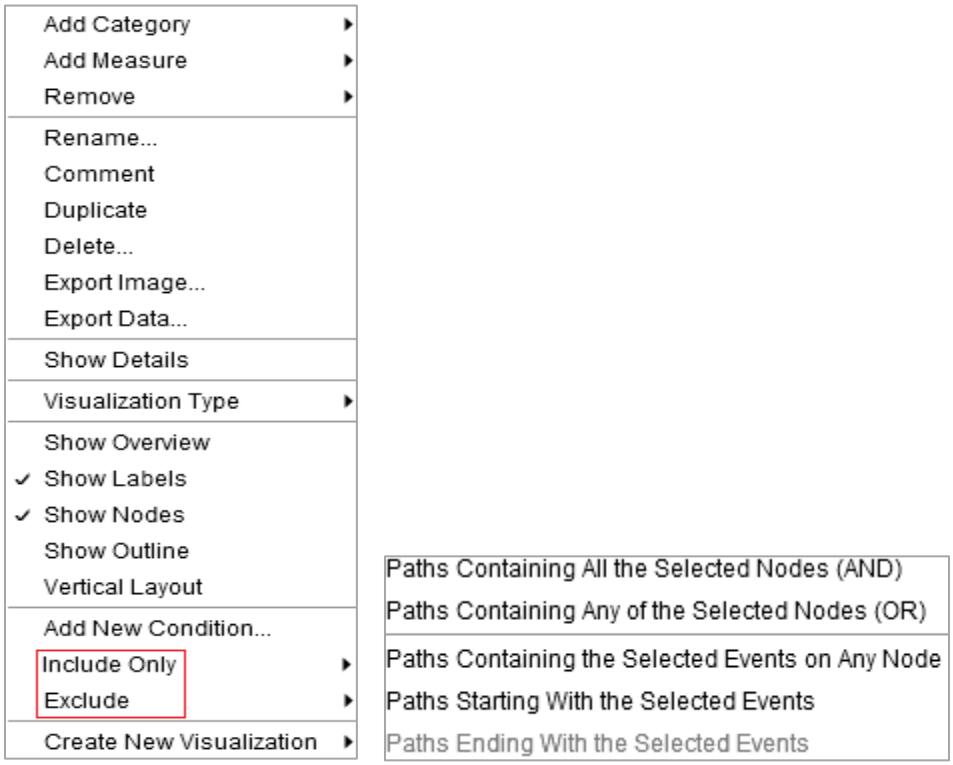


Figure 13. Adding condition from selected node

Both of these options have following sub conditions:

- **Paths Containing All the Selected Nodes**

It includes or excludes path on the basis of the selection made on node for path analysis.

- **Paths Containing Any of the Selected Nodes**

It includes or excludes path on the basis of any of the selected nodes for path analysis.

- **Paths Containing the Selected Events on Any Node**

It includes or excludes path on the basis of the selected events, on any node for path analysis.

- **Paths Starting With the Selected Events**

It includes or excludes paths that start with the selected events for path analysis.

- **Paths Ending With the Selected Events**

It includes or excludes paths that end with the selected events for path analysis.

These methods to filter the paths in Sankey diagram by selecting conditions, gives more flexibility and ability to dive deep into the data and analyze from different perspectives.

EDIT A CONDITION FOR A PATH SELECTION

- To edit a condition for a path selection, click “Options” icon in the visualization tool bar as below:

- Open the details table by selecting “Show Details”.
- In the details table, select the tab “Path Selection” and in the “Type”, select the condition of choice.

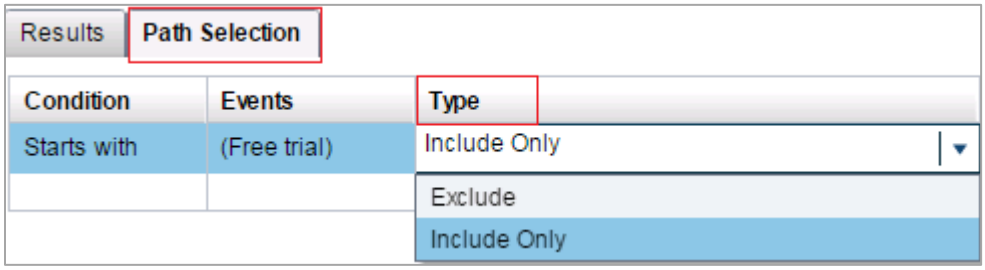


Figure 14. Edit condition from details table

REMOVE CONDITIONS FROM A PATH SELECTION

- To remove a condition from selected path, click “Options” icon in the visualization tool bar as below:

- Open the details table by selecting Show Details.
- In the details table, select the tab “Path Selection”
- Select the existing condition and right click for options.

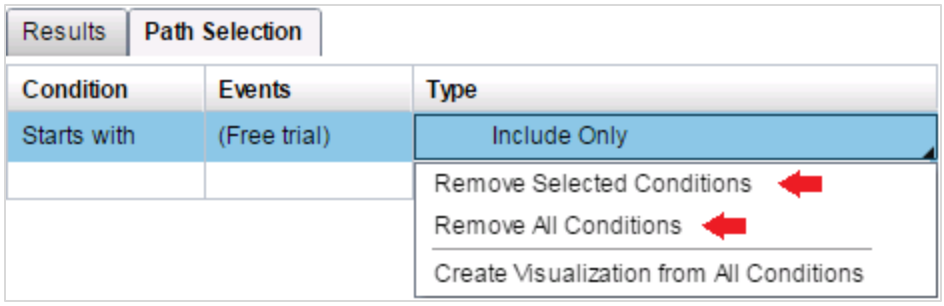


Figure 15. Removing conditions from path selection

Select the option for removing the selected conditions or removing all the conditions in case multiple selections are made for path filtering.

WAYS TO EXPLORE TRANSACTION IDENTIFIER IN SANKEY DIAGRAM

One of the interesting functionality of Sankey is to be able to explore the transaction identifier used in path analysis. This exploration is available only when one or more path has been managed or selected by “Add New Condition”, as described above.

PRE - SELECTED PATH FOR TRANSACTION IDNETIFIER EXPLORATION:

If node is pre-selected for filtering, the results table is available in visualization window.

- Select a condition under the tab “Path Selection” in the details table.
- Right click on the condition and select “Create Visualization from all conditions”. This will select all the conditions used for path selection for filtering data into a new visualization.

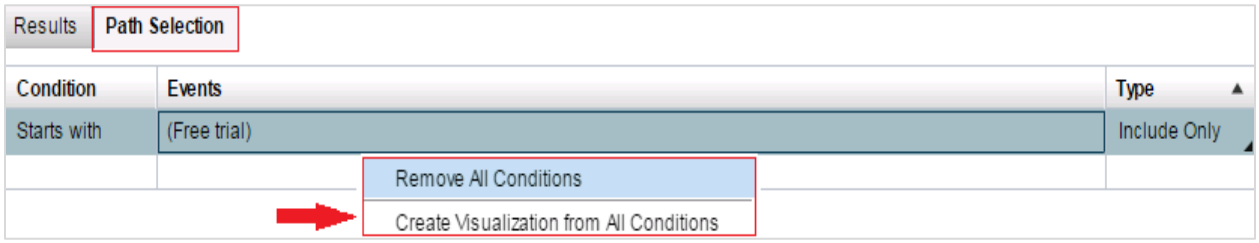


Figure 16. Creating Visualization from selected path

A new visualization window opens with the bar chart. The frequency on Y-axis, indicates the number of times the Call subjects have been inquired by the customer. In example below, Event – “Free Trial” has been used to explore the transaction identifier – “Customer” data values.

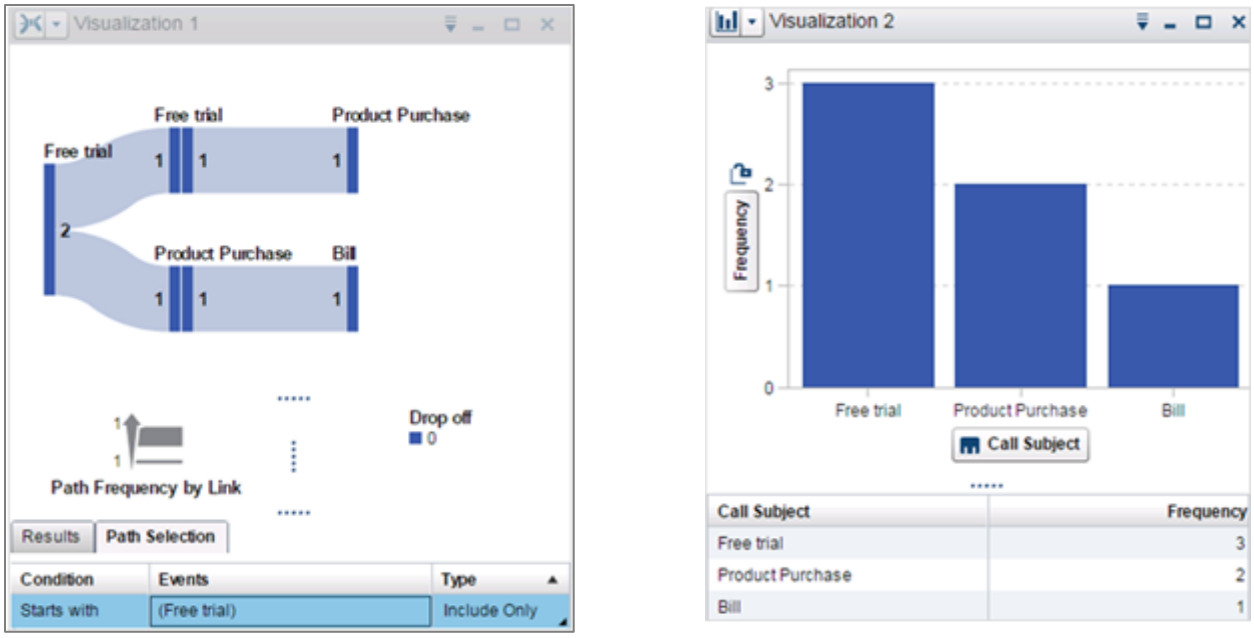


Figure 17. Exploring the transaction identifier

PATH SELECTION AND TRANSACTION IDENTIFIER EXPLORATION

It is also possible to apply filters to the path and analyse transaction identifier at the time.

- Select the node of interest from Sankey diagram, right click for options and then select “Create New Visualization”.

Before opening new visualization window, it will prompt to make choice of path / node using conditions.

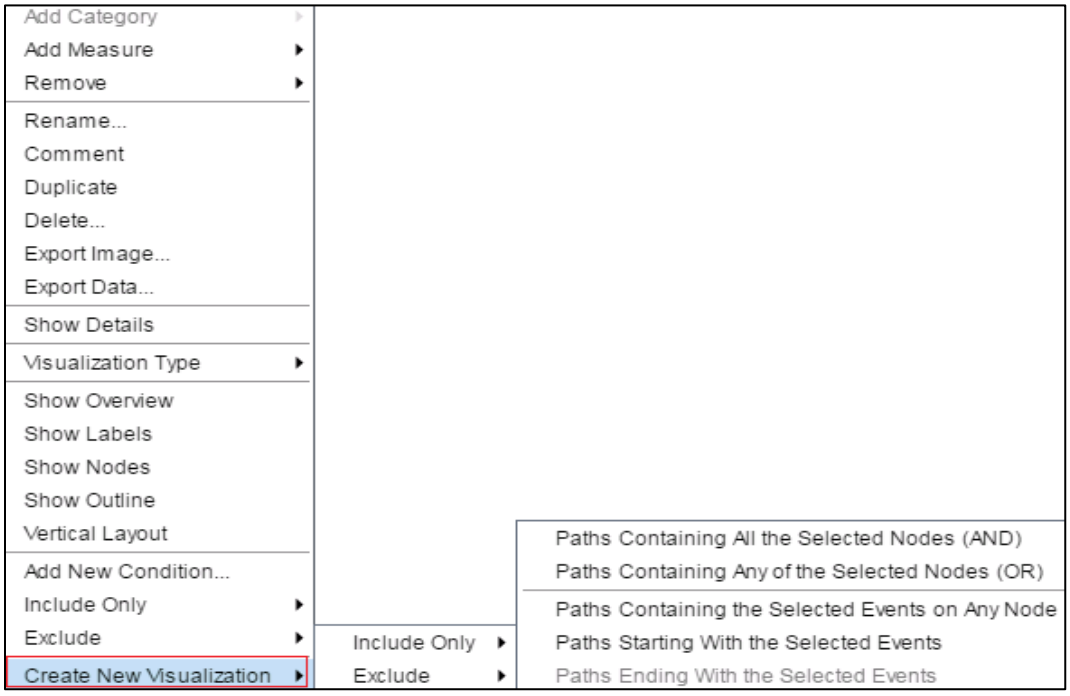


Figure 18. Exploring Transaction Identifier through selecting path

Select any one of the conditions to filter path and explore transaction identifier.

In example below, following condition are used :

Create New Visualization > Exclude > Path containing all selected nodes

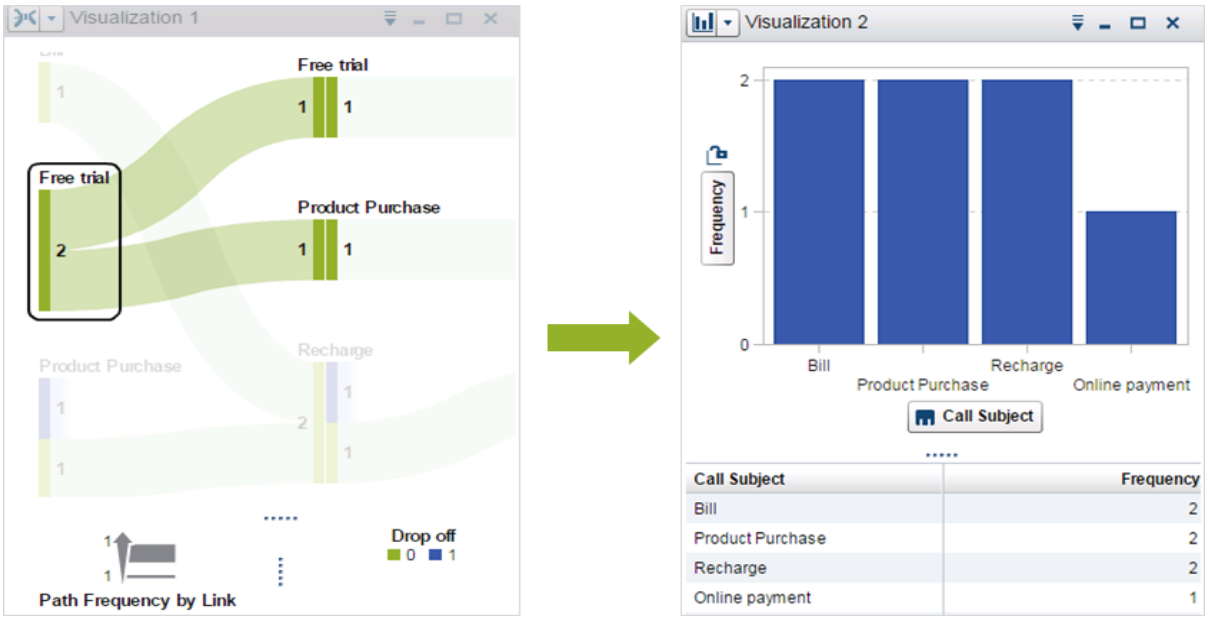


Figure 19. To demonstrate the exploration through path selection

The new visualization window contains the result that shows only those paths that do not start with event “Free Trial” and it becomes handy to analyze the transaction identifier of the available paths.

WAYS TO PERFORM PATH ANALYTICS IN SANKEY DIAGRAM

Sankey diagram for a data can be visualized and analyzed using link color, link width, path ranking, minimum & maximum frequency and length. They are very useful in limiting or narrowing down the path for analysis. There are available under “Properties” tab in SAS® Visual Analytics and have been discussed briefly below:

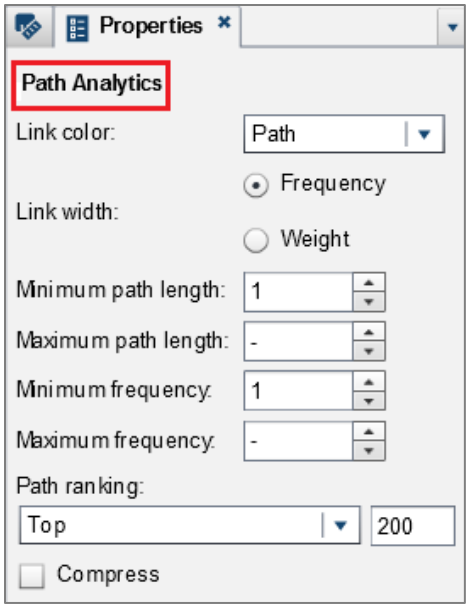


Figure 19. Path analytics under properties tab in SAS® Visual Analytics

LINK COLOR

It allows the user to give different colours to the Sankey diagram on the basis of an event, path or drop-off

1) **By Path**: It displays each path distinctly and gives colour to each group. This is a very good way to analyse data and track the path followed. For example, in the fig below, Sankey diagram clearly shows the path followed by each customer

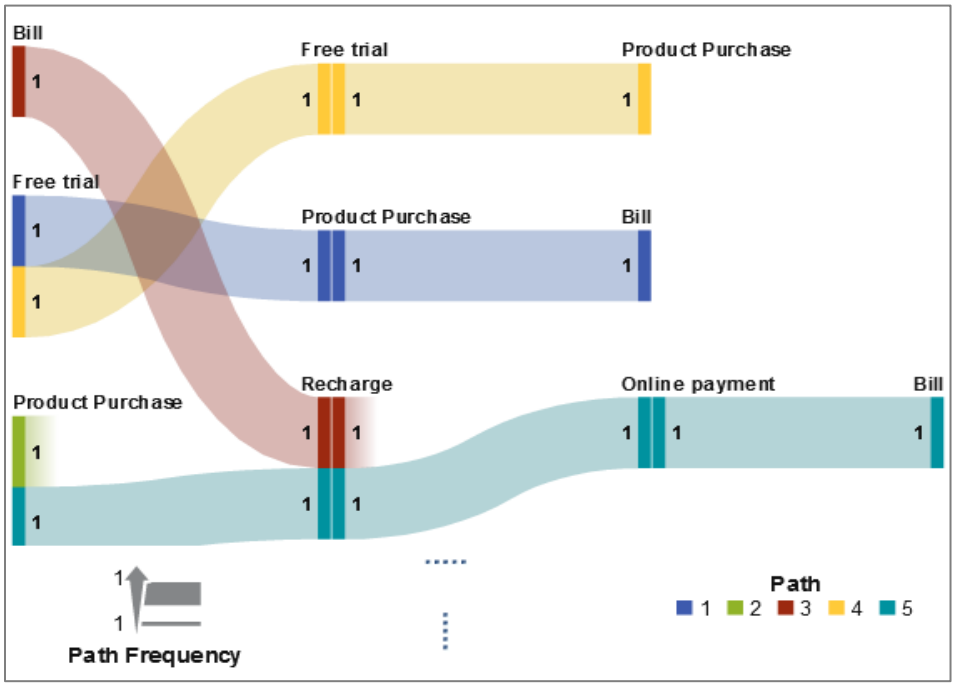


Figure 20. Link color by path

2) **By Drop-off:** It displays the path that are group on the basis of the drop off links. It provides a helpful insight and helps in identifying the cause for the drop off- events, path and width

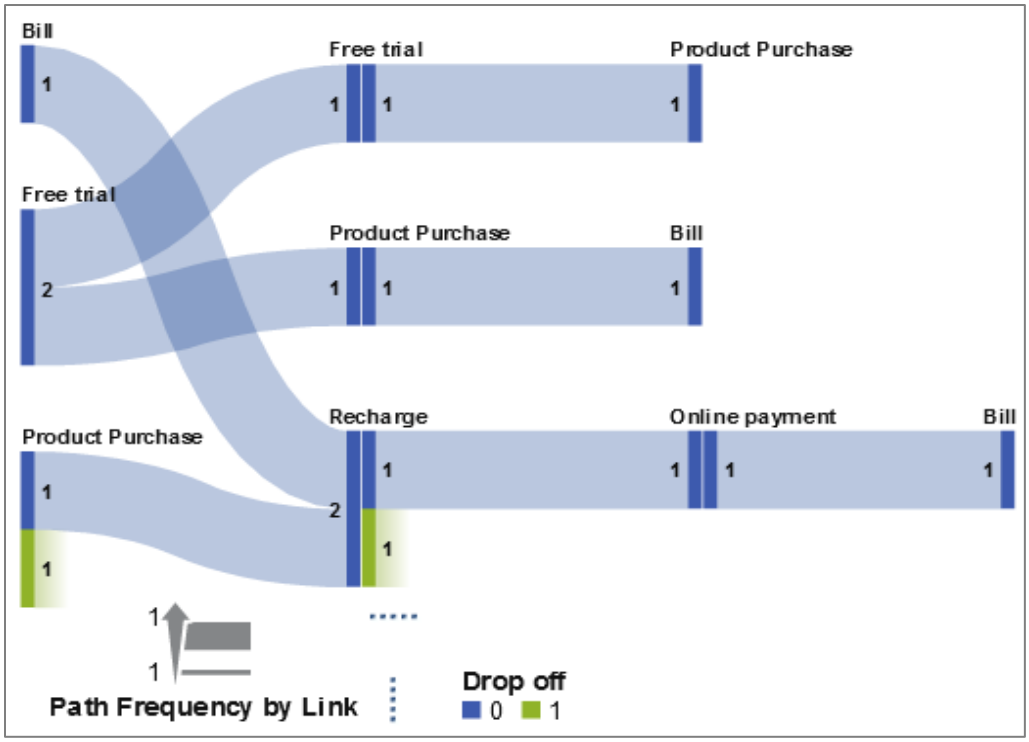


Figure 21. Link color by drop-off

3) **By Event:** It displays the path that links the group on the basis of event and provide same color to the links that originate from the same event.

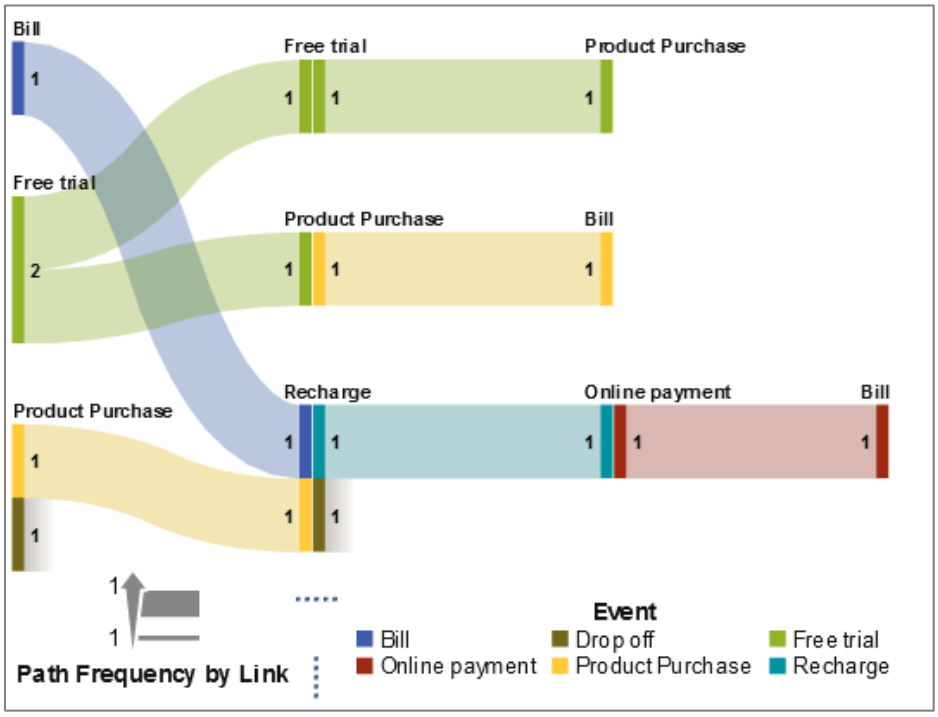


Figure 22. Link color by Event

LINK WIDTH
This option is available if the Sankey diagram has a Weight measure. You can chose to show the width of a link either as a frequency or as a aggregated measure in a path. By default, Sankey diagram in SAS® Visual Analytics always shows the link width in terms of frequency. Also, if the Weight data has negative, zero or missing values for a path, then frequency is used to define the link width.

MINIMUM & MAXIMUM PATH LENGTH
It allows you to specify the desired minimum and maximum number of nodes in a path. For instance- if minimum = 2 and maximum = 4 is selected, then the diagram displays only those paths that have at least 2 nodes and not more than 4 nodes.

MINIMUM & MAXIMUM FREQUENCY
It allows you to specify the desired minimum and maximum range of frequency for a path. For instance- if minimum = 2 and maximum = 4 is selected, then the diagram displays only those paths that have the frequency of 2, 3 or 4.

PATH RANKING
It allows to select the path in the diagram on the basis of ranks. If Weight is used as measure, then ranking is based on the aggregated value of the weight measure for each path, else it is based on the frequency of each path.
Path ranking can be selected to choose paths either from the higher values or from lower values.

By default in SAS® Visual Analytics, rank is selected for “Top 200” paths.
Let’s consider an example with conditions as “Bottom” and “2” to see how the path ranking affects the path in Sankey diagram.

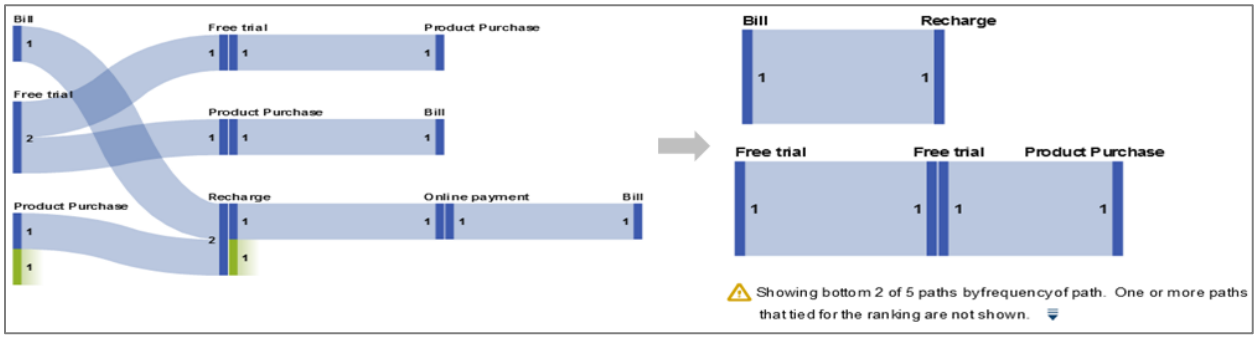


Figure 23. Path Ranking in Sankey diagram

Path ranking with conditions above results in only two paths, and thus helps to analyse the paths that are less travelled and gives the direction to improve the service. When path ranking is applied, it also gives an indication in the visualization, stating that paths shown are on the top or bottom approach and also the number of paths displayed from the count.

COMPRESS

This option is useful, when the data has lot of repetitive events. Upon selection, it combines all the repetitive and consecutive events into a single event for each path.

HANDY USAGE TIPS

SAS® Visual Analytics comes with handy and efficient tools to visualize Sankey diagram.

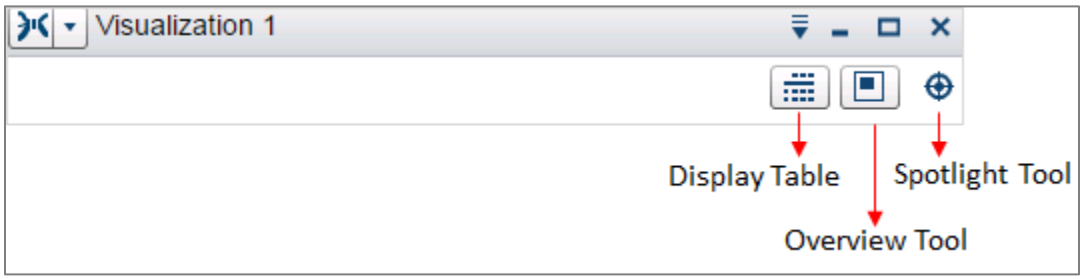


Figure 24. Hidden Features (Visible upon hovering)

SPOTLIGHT TOOL

- In order to use the spotlight tool, hover over the visualization area, to enable or disable the spotlight tool.
- Move the cursor to the selected node with white circle showing the details of the node.

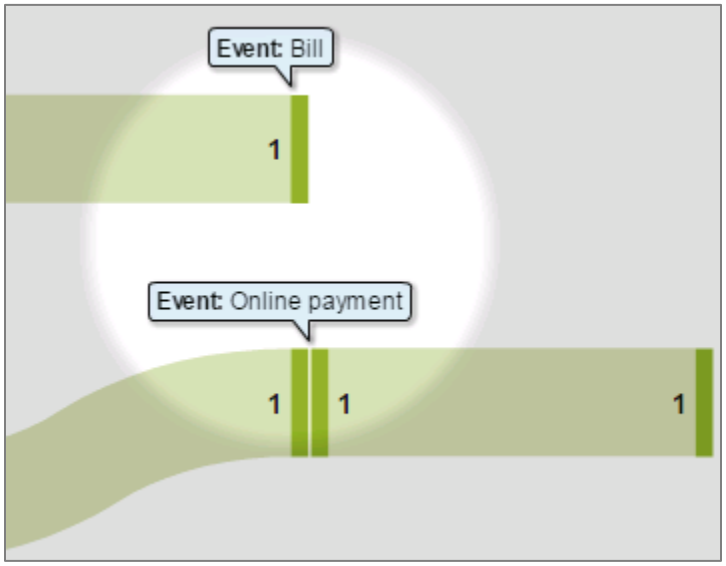


Figure 25. Spotlight tool with details of the selected node

OVERVIEW TOOL

It is helpful, when Sankey diagram is too big and need to focus only on some points/nodes.

- Hover over the visualization area, to enable or disable the overview.
- A new window opens, which allows to select the area of interest to be zoomed in and out.

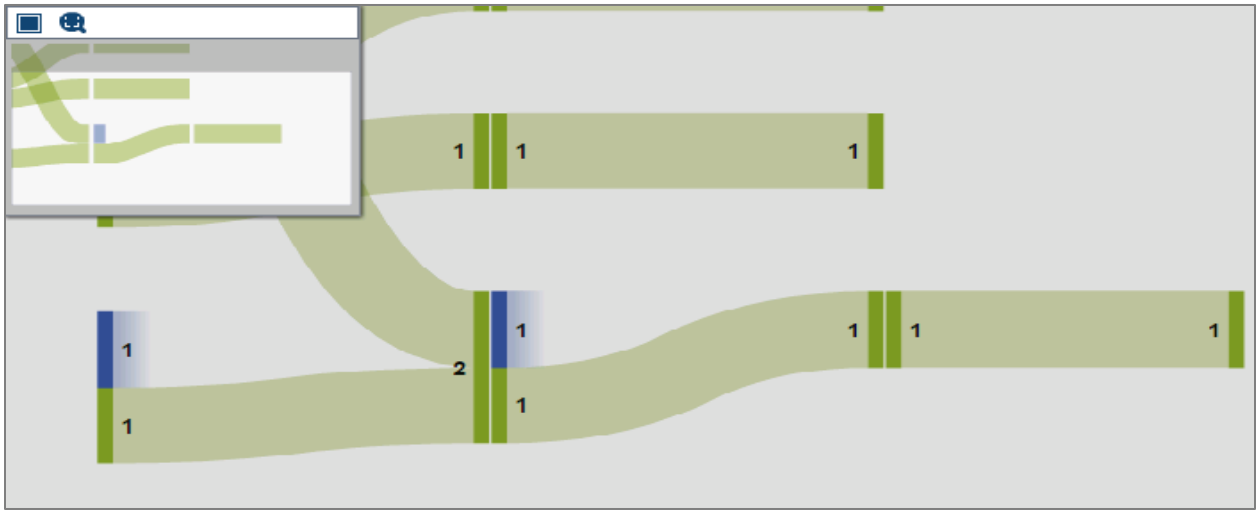


Figure 26. Overview window to focus or change the area of visualization.

MANAGE VISUALIZATIONS

Sometimes, the data analysis requires data to be visualized under different conditions, and this often led to too many visualizations windows on main screen. Also, it takes a lot of space in the workspace.

These visualization can be managed easily by a single click on “Manage Visualization” icon or by selecting any of the minimized Visualization window.

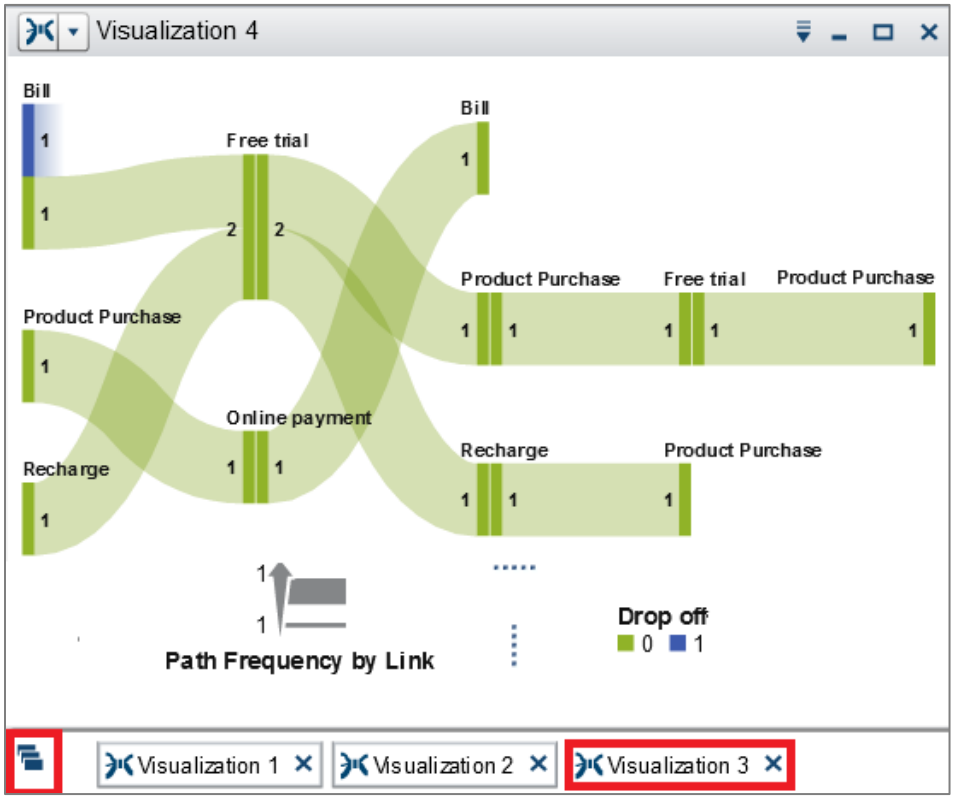


Figure 27. Manage Visualizations in SAS® Visual Analytics

Clicking on the icon, opens a new window with all the visualizations that were made for the analysis. Choice can be made to select the visualization of interest to be displayed on the workspace, while the rest can still be minimized and saved for further analysis.

Alternatively, selection can also be made by single click on the minimized visualizations, to appear on the workspace.

CONCLUSION

Path Analysis is very important in order to understand customer journey and their behavior pattern for gaining actionable insights into data. Sankey diagram in SAS® Visual Analytics not only provides a complete and robust means for path analysis, but also helps in providing insights on different path taken by the user.

Different ways to perform paths filtering and ability to explore the transaction identifier, makes it easy for the organization to understand their customer and predict progression of events that would help in steering the organization and in making decisions.

REFRENCES

Unternehmensberatung und Forschungsgesellschaft fur Umweltfragen. "STENUM GmbH". Accessed January 10, 2016. http://www.stenum.at/en/?id=software/sankey/sankey-glossar .

"Google developers". Accessed January 10, 2016. https://developers.google.com/chart/interactive/docs/gallery/sankey

"The Data Visualisation Catalogue". Accessed January 10, 2016. http://www.datavizcatalogue.com/methods/sankey_diagram.html

SAS Institute Inc. "SAS(R) Visual Analytics 7.1: User's Guide". Accessed January 10, 2016. http://support.sas.com/documentation/cdl/en/vaug/67500/HTML/default/viewer.htm#n08zz749uilz3cn1u3hm8b8dmos2.htm

Varsha Chawla and Renato Luppi. 2015. Sankey Diagrams in SAS® Visual Analytics. SAS Institute Inc., Cary, NC http://support.sas.com/resources/papers/proceedings15/SAS1808-2015.pdf

Falko Schulz, Brisbane, Australia and Olaf Kratzsch, Cary, NC. 2015.Taking the Path More Travelled – SAS Visual Analytics® and Path Analysis. SAS Institute Inc. http://support.sas.com/resources/papers/proceedings15/SAS1444-2015.pdf

CONTACT INFORMATION

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

Abhilasha Tiwari
Accenture Netherlands
+31 (0) 61 0120389
tiwari.abhilasha1@gmail.com