How to Use R to Create a Migration Sankey Chart

Chiung-Ya Tang, Ph.D.

Institutional Knowledge Management
University of Central Florida
Feb. 24, 2020
Learning Objectives

Audience will be able to:

• learn basic data preparation and codes for creating a migration Sankey chart

• revise codes freely to fit their needs of presenting the data flow

• discuss questions related to creating a Sankey chart with the presenter and other attendees
Sankey Diagrams

• Function:
  A visualization technique used to present the flow of data from one entity (node) to another.

• Type:
  Depending on the structure of data and the purpose (example).

• Data structure:
  • Link Data: Source, Target, and Counts
  • Node Data: Names
Example of Sankey Diagrams

Show the progress between entities

Display the intermediates between entities

Reference: Data to Viz (https://www.data-to-viz.com/graph/chord.html)
Example of Link Data

<table>
<thead>
<tr>
<th></th>
<th>from</th>
<th>to</th>
<th>counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CBA</td>
<td>CBA</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>CBA</td>
<td>CON</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>CBA</td>
<td>CAH</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>CBA</td>
<td>CCIE</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>CBA</td>
<td>COS</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CBA</td>
<td>COM</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>CON</td>
<td>CBA</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>CON</td>
<td>CON</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>CON</td>
<td>CAH</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>CON</td>
<td>CCIE</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>CON</td>
<td>COS</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>CON</td>
<td>COM</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>CAH</td>
<td>CBA</td>
<td>2</td>
</tr>
</tbody>
</table>
What is R?

- Definition
  - a language and environment for statistical computing and graphics

- Environment
  - R
  - R studio: an integrated development environment (IDE) for R
Bring on
the “R”
### R Codes for creating a Sankey Chart

```r
setwd("The Path of your Working Directory")
# If you don't know your working directory, do this:
getwd()

# Install libraries
library(dplyr) # for data management
library(tidyverse) # for data management
library(readxl) # for reading Excel file
library(networkD3) # for making Sankey chart

### Import data file.
# Method 1: Import Excel file using tidyverse
Link <- read_excel("ChangingColleges.xlsx")
Link <- as.data.frame(Link)
# Method 2: Import comma delimited file.
Link1 <- read.csv("ChangingColleges.csv")
rm(Link1) # remove unnecessary data file

# Paste a space in the target value with no separator
Link$target <- paste(Link$target, " ", sep=""")

#checking data information.
head(Link, 10)

### Preparing the NODES file.
# From these flows we need to create a node data frame that store the names of all the source
# and target units involved in the flow
nodes <- data.frame(name = c(as.character(Link$source), as.character(Link$target)) %>% unique())

# Based on the networkD3 package, connection must be provided using numerical id, not using real label.
# So we need to reformat it.
Link$IDsource = match(Link$source, nodes$name) - 1
Link$IDtarget = match(Link$target, nodes$name) - 1

# prepare colourscale
colourscale = d3.scaleOrdinal().range(["red", "orange", "yellow", "green", "blue", "purple"])

# Make the Network
chart <- sankeyNetwork(Links = Link, Nodes = nodes,
Source = "IDsource", Target = "IDtarget", Value = "counts", NodeID = "name",
sinksRight = FALSE, colourScale = colourscale, LinkGroup = "source", NodeGroup = "name",
nodeWidth = 40, fontSize = 20, nodePadding = 20)

# save the widget
library(htmlwidgets)
library(htmltools)
chart <- prependContent(chart, tags$h1('Sankey Chart', noWS = NULL))
saveWidget(chart, file = paste0(getwd(), "/sankeychart.html"))

# Keeping the path that has value greater than 5 in order to make the graph parsimonious
Link <- Link[Link$counts > 5, ]
```
Comparison

- Low or no cost
- Adapting to data updates easily
- Graph saved in html format

- Higher cost
- Need extra work when updating data
- Internet access for interactive effect
Use Sankey Effectively

• A Sankey chart when used effectively, it is a good tool to show the flow of data.

• In addition to R, other software/program such as Excel and Tableau could produce the diagram also.

• Consider reducing the number of paths to be parsimonious.
Other Resources

• Use Tableau
  (e.g., Evergreen Data Academy, https://academy.stephanieevergreen.com/)

• Use Python
  (e.g., https://plot.ly/python/sankey-diagram/)

• Use Excel
  (e.g., third party Add-in product)
Questions?