Assignment 7 Graph Theory

Ridhwan Firdaus bin Mohd Jata 16000250 EDB4623:Big Data Analytics

3/31/2021

## Introduction

The two primary aspects of networks are a multitude of separate entities and the connections between them.The entities are referred to as nodes or vertices of a graph, while the connections are edges or links

## Data used

The data sets that will be used relates to French trains; it contains aggregate daily total trips per connecting stations.

library(readr)

## Warning: package 'readr' was built under R version 3.6.3

url <- "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-02-26/small\_trains.csv"  
small\_trains <- read\_csv(url)

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## year = col\_double(),  
## month = col\_double(),  
## service = col\_character(),  
## departure\_station = col\_character(),  
## arrival\_station = col\_character(),  
## journey\_time\_avg = col\_double(),  
## total\_num\_trips = col\_double(),  
## avg\_delay\_all\_departing = col\_double(),  
## avg\_delay\_all\_arriving = col\_double(),  
## num\_late\_at\_departure = col\_double(),  
## num\_arriving\_late = col\_double(),  
## delay\_cause = col\_character(),  
## delayed\_number = col\_double()  
## )

head(small\_trains)

## # A tibble: 6 x 13  
## year month service departure\_station arrival\_station journey\_time\_avg  
## <dbl> <dbl> <chr> <chr> <chr> <dbl>  
## 1 2017 9 National PARIS EST METZ 85.1  
## 2 2017 9 National REIMS PARIS EST 47.1  
## 3 2017 9 National PARIS EST STRASBOURG 116.   
## 4 2017 9 National PARIS LYON AVIGNON TGV 161.   
## 5 2017 9 National PARIS LYON BELLEGARDE (AIN) 164.   
## 6 2017 9 National PARIS LYON BESANCON FRANCHE COMT~ 129.   
## # ... with 7 more variables: total\_num\_trips <dbl>,  
## # avg\_delay\_all\_departing <dbl>, avg\_delay\_all\_arriving <dbl>,  
## # num\_late\_at\_departure <dbl>, num\_arriving\_late <dbl>, delay\_cause <chr>,  
## # delayed\_number <dbl>

## Data Cleaning and preparation

Data will be changed in order to create a new category called routes. It contains a single entry for each connected station. It also includes the average journey time it takes to go between stations.

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

routes <- small\_trains %>% group\_by(departure\_station, arrival\_station) %>% summarise(journey\_time = mean(journey\_time\_avg)) %>% ungroup()%>% mutate(from = departure\_station,to = arrival\_station) %>% select(from, to, journey\_time)

## `summarise()` has grouped output by 'departure\_station'. You can override using the `.groups` argument.

head(routes)

## # A tibble: 6 x 3  
## from to journey\_time  
## <chr> <chr> <dbl>  
## 1 AIX EN PROVENCE TGV PARIS LYON 186.   
## 2 ANGERS SAINT LAUD PARIS MONTPARNASSE 97.5  
## 3 ANGOULEME PARIS MONTPARNASSE 146.   
## 4 ANNECY PARIS LYON 225.   
## 5 ARRAS PARIS NORD 52.8  
## 6 AVIGNON TGV PARIS LYON 161.

## Creating the adjency matrix

adjmatrix <- subset(routes, select = -c(journey\_time))  
head(adjmatrix)

## # A tibble: 6 x 2  
## from to   
## <chr> <chr>   
## 1 AIX EN PROVENCE TGV PARIS LYON   
## 2 ANGERS SAINT LAUD PARIS MONTPARNASSE  
## 3 ANGOULEME PARIS MONTPARNASSE  
## 4 ANNECY PARIS LYON   
## 5 ARRAS PARIS NORD   
## 6 AVIGNON TGV PARIS LYON

edges\_mat = matrix(nrow = length(adjmatrix$from), ncol = length(adjmatrix$to), data=0, dimnames = list(adjmatrix$from, adjmatrix$to))  
edges\_mat[as.matrix(adjmatrix)] = 1  
head(edges\_mat)

## PARIS LYON PARIS MONTPARNASSE PARIS MONTPARNASSE PARIS LYON  
## AIX EN PROVENCE TGV 1 0 0 0  
## ANGERS SAINT LAUD 0 1 0 0  
## ANGOULEME 0 1 0 0  
## ANNECY 1 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 1 0 0 0  
## PARIS NORD PARIS LYON PARIS LYON PARIS LYON PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 1 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## PARIS MONTPARNASSE PARIS VAUGIRARD TOURCOING  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0  
## PARIS MONTPARNASSE PARIS LYON PARIS LYON PARIS NORD  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS NORD PARIS EST PARIS LYON PARIS LYON PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## PARIS MONTPARNASSE PARIS LYON PARIS MONTPARNASSE PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS MONTPARNASSE LYON PART DIEU MARSEILLE ST CHARLES  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0  
## PARIS NORD LILLE MARNE LA VALLEE MARSEILLE ST CHARLES  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## MONTPELLIER PARIS LYON RENNES PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## MARSEILLE ST CHARLES LYON PART DIEU MARSEILLE ST CHARLES  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0  
## LILLE LYON PART DIEU MADRID MARNE LA VALLEE PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## TOURCOING PARIS EST LYON PART DIEU PARIS LYON PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## PARIS EST PARIS MONTPARNASSE PARIS VAUGIRARD STRASBOURG  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS LYON PARIS LYON FRANCFORT METZ NANCY REIMS STRASBOURG  
## AIX EN PROVENCE TGV 0 0 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0 0 0  
## ANNECY 0 0 0 0 0 0 0  
## ARRAS 0 0 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0 0 0  
## STUTTGART AIX EN PROVENCE TGV ANNECY AVIGNON TGV BARCELONA  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## BELLEGARDE (AIN) BESANCON FRANCHE COMTE TGV  
## AIX EN PROVENCE TGV 0 0  
## ANGERS SAINT LAUD 0 0  
## ANGOULEME 0 0  
## ANNECY 0 0  
## ARRAS 0 0  
## AVIGNON TGV 0 0  
## CHAMBERY CHALLES LES EAUX DIJON VILLE GENEVE GRENOBLE  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## ITALIE LAUSANNE LE CREUSOT MONTCEAU MONTCHANIN  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0  
## LYON PART DIEU MACON LOCHE MARSEILLE ST CHARLES MONTPELLIER  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## MULHOUSE VILLE NICE VILLE NIMES PERPIGNAN  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## SAINT ETIENNE CHATEAUCREUX TOULON VALENCE ALIXAN TGV ZURICH  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## ANGERS SAINT LAUD ANGOULEME BORDEAUX ST JEAN BREST  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## LA ROCHELLE VILLE LAVAL LE MANS NANTES POITIERS QUIMPER  
## AIX EN PROVENCE TGV 0 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0 0  
## ANNECY 0 0 0 0 0 0  
## ARRAS 0 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0 0  
## RENNES ST MALO ST PIERRE DES CORPS TOULOUSE MATABIAU TOURS  
## AIX EN PROVENCE TGV 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0  
## ANNECY 0 0 0 0 0  
## ARRAS 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0  
## VANNES ARRAS DOUAI DUNKERQUE LILLE BORDEAUX ST JEAN NANTES  
## AIX EN PROVENCE TGV 0 0 0 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0 0 0 0  
## ANGOULEME 0 0 0 0 0 0 0  
## ANNECY 0 0 0 0 0 0 0  
## ARRAS 0 0 0 0 0 0 0  
## AVIGNON TGV 0 0 0 0 0 0 0  
## RENNES PARIS LYON PARIS MONTPARNASSE PARIS MONTPARNASSE  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS EST LYON PART DIEU PARIS MONTPARNASSE PARIS VAUGIRARD  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS LYON PARIS MONTPARNASSE PARIS MONTPARNASSE NANTES  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## PARIS EST PARIS EST PARIS LYON PARIS MONTPARNASSE  
## AIX EN PROVENCE TGV 0 0 0 0  
## ANGERS SAINT LAUD 0 0 0 0  
## ANGOULEME 0 0 0 0  
## ANNECY 0 0 0 0  
## ARRAS 0 0 0 0  
## AVIGNON TGV 0 0 0 0  
## BORDEAUX ST JEAN MARSEILLE ST CHARLES PARIS MONTPARNASSE  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0  
## PARIS LYON PARIS MONTPARNASSE PARIS LYON  
## AIX EN PROVENCE TGV 0 0 0  
## ANGERS SAINT LAUD 0 0 0  
## ANGOULEME 0 0 0  
## ANNECY 0 0 0  
## ARRAS 0 0 0  
## AVIGNON TGV 0 0 0

## Transform data into graph table

In order to prepare routes for this transformation, it has to contain two variables specifically named from and to.The member of the group is called a node (or vertex) in the graph, and a relationship between nodes is called an edge.

library(tidygraph)

## Warning: package 'tidygraph' was built under R version 3.6.3

##   
## Attaching package: 'tidygraph'

## The following object is masked from 'package:stats':  
##   
## filter

graph\_routes <- as\_tbl\_graph(routes)  
  
graph\_routes

## # A tbl\_graph: 59 nodes and 130 edges  
## #  
## # A directed simple graph with 1 component  
## #  
## # Node Data: 59 x 1 (active)  
## name   
## <chr>   
## 1 AIX EN PROVENCE TGV  
## 2 ANGERS SAINT LAUD   
## 3 ANGOULEME   
## 4 ANNECY   
## 5 ARRAS   
## 6 AVIGNON TGV   
## # ... with 53 more rows  
## #  
## # Edge Data: 130 x 3  
## from to journey\_time  
## <int> <int> <dbl>  
## 1 1 39 186.   
## 2 2 40 97.5  
## 3 3 40 146.   
## # ... with 127 more rows

library(stringr)  
  
graph\_routes <- graph\_routes %>% activate(nodes) %>% mutate( title = str\_to\_title(name), label = str\_replace\_all(title, " ", "\n") )  
stations <- graph\_routes %>% activate(nodes) %>% pull(title)

## Visualizing

In this section we will visualize all the nodes and edges from the dataset

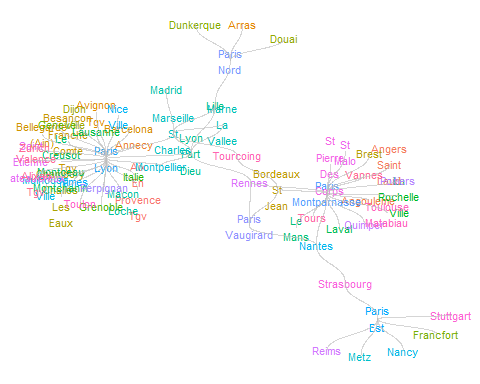
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

library(ggraph)

## Warning: package 'ggraph' was built under R version 3.6.3

thm <- theme\_minimal() +  
 theme(  
 legend.position = "none",  
 axis.title = element\_blank(),  
 axis.text = element\_blank(),  
 panel.grid = element\_blank(),  
 panel.grid.major = element\_blank(),  
 )   
  
theme\_set(thm)  
graph\_routes %>% ggraph(layout = "kk") + geom\_node\_text(aes(label = label, color = name), size = 3) +geom\_edge\_diagonal(color = "gray", alpha = 0.4)

 ## Finding the shortest distance In this section we will identify the shortest route from station Arras to station Nancy

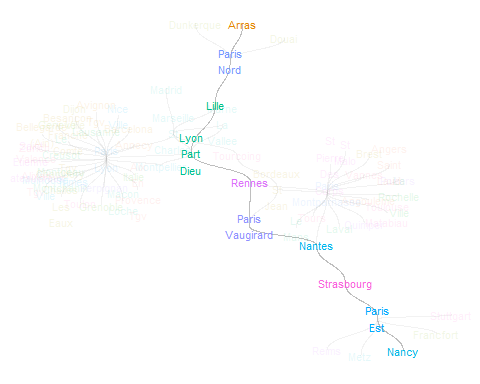
from <- which(stations == "Arras")  
to <- which(stations == "Nancy")  
  
shortest <- graph\_routes %>% morph(to\_shortest\_path, from, to, weights = journey\_time)  
shortest %>% mutate(selected\_node = TRUE) %>% unmorph()

## # A tbl\_graph: 59 nodes and 130 edges  
## #  
## # A directed simple graph with 1 component  
## #  
## # Node Data: 59 x 4 (active)  
## name title label selected\_node  
## <chr> <chr> <chr> <lgl>   
## 1 AIX EN PROVENCE TGV Aix En Provence Tgv "Aix\nEn\nProvence\nTgv" NA   
## 2 ANGERS SAINT LAUD Angers Saint Laud "Angers\nSaint\nLaud" NA   
## 3 ANGOULEME Angouleme "Angouleme" NA   
## 4 ANNECY Annecy "Annecy" NA   
## 5 ARRAS Arras "Arras" TRUE   
## 6 AVIGNON TGV Avignon Tgv "Avignon\nTgv" NA   
## # ... with 53 more rows  
## #  
## # Edge Data: 130 x 3  
## from to journey\_time  
## <int> <int> <dbl>  
## 1 1 39 186.   
## 2 2 40 97.5  
## 3 3 40 146.   
## # ... with 127 more rows

shortest <- shortest %>% mutate(selected\_node = TRUE) %>% activate(edges) %>% mutate(selected\_edge = TRUE) %>% unmorph()   
shortest <- shortest %>% activate(nodes) %>% mutate(selected\_node = ifelse(is.na(selected\_node), 1, 2)) %>% activate(edges) %>% mutate(selected\_edge = ifelse(is.na(selected\_edge), 1, 2)) %>% arrange(selected\_edge)

## Visualizing the shortest distance

shortest %>%  
 ggraph(layout = "kk") +  
 geom\_edge\_diagonal(aes(alpha = selected\_edge), color = "gray") +  
 geom\_node\_text(aes(label = label, color =name, alpha = selected\_node ), size = 3)

 ## Summarizing the graph table

shortest %>%  
 activate(edges) %>%  
 filter(selected\_edge == 2) %>%  
 as\_tibble() %>%  
 summarise(  
 total\_stops = n() - 1,  
 total\_time = round(sum(journey\_time) / 60)  
 )

## # A tibble: 1 x 2  
## total\_stops total\_time  
## <dbl> <dbl>  
## 1 8 23

As we can see there will be a total of 8 stops between Arras and Nancy and the total time is 23 minutes