Transport-Reachability-Network Analysis

Assignment im Rahmen der Vorlesung 'Social Network Analyis'

Johannes Bubeck

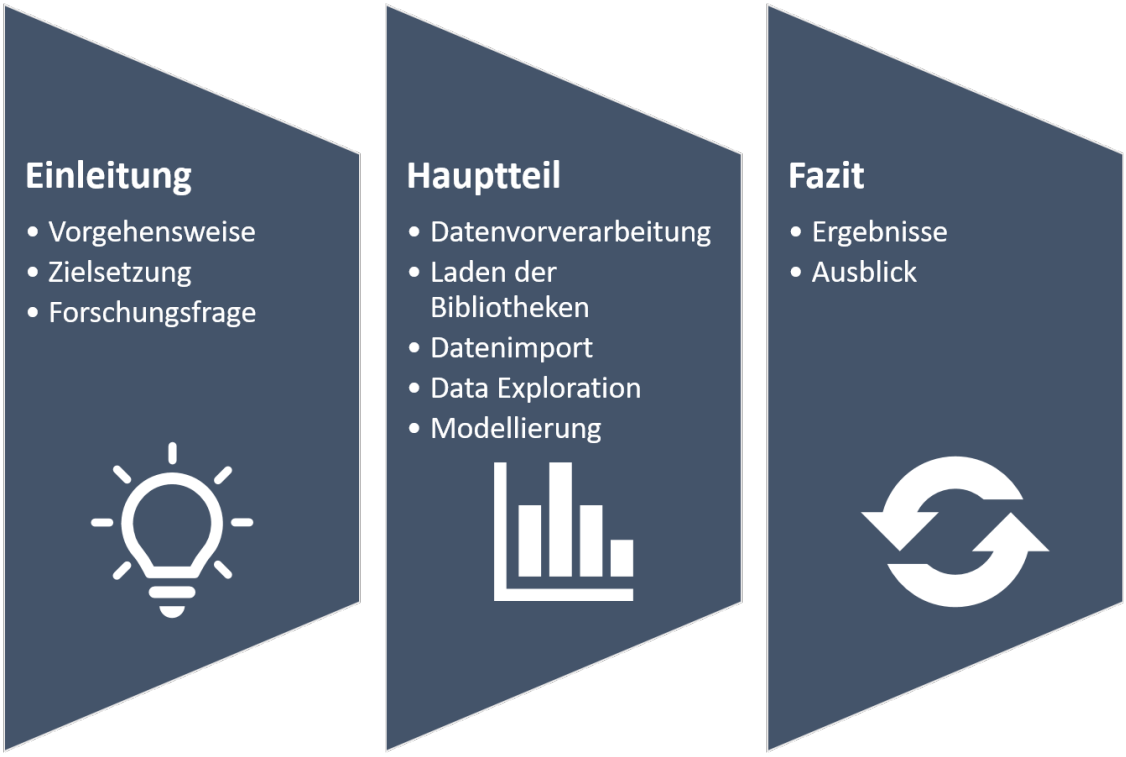
Inhaltsverzeichnis

1 Einleitung

1.1 Vorgehensweise und Zielsetzung

Die Zielsetzung dieses Projekts ist die Durchführung einer "Social Network Analysis" im Rahmen der Leistungsbeurteilung des gleichnamigen DHBW-Kurses. Die Datenbasis der Analyse und des Assignments stellt der Datensatz "Airline travel reachability network" dar (https://snap.stanford.edu/data/reachability.html).

Folgende Abbildung zeigt die Vorgehensweise dieses Projekts. In der Einleitung wird das Vorgehen des "Social Network Analysis Projekts" beschrieben. Weiterhin wird die Zielsetzung und die des Assignment zugrundeliegende Forschungsfrage definiert. Der Hauptteil des Assignments bildet eine "klassische" Vorgehensweise im Data Analytics Bereich ab. In einem ersten Schritt wird zunächst der theoretische Hintergrund des Projekts dargelegt. Dazu werden unter anderem Punkte wie die verfügbaren Daten, die Netzwerkdefinition, Flow Struktur und das Zentralitätsmaß näher beleichtet. Nach der Datenvorverarbeitung und dem Laden der notwendigen Bibliotheken, die für die Analyse notwendig sind, erfolgt der Datenimport. Weiterhin dient eine Explorative Analyse der Daten näheren Einblicken in die verfügbaren Daten und deren Qualität. Schließlich wird im letzten Schritt des Hauptteils die Modellierung und Visualisierung durchgeführt. Das Fazit beschäftigt sich zum Einen mit der Bündelung und Präsentation der Daten und zum Anderen einem Ausblick.



1.2 Forschungsfrage

Der Datensatz "Airline travel reachability network" bildet Flugverbindungen für Flughäfen in den Vereinigten Staaten sowie Kanada ab. Zusätzlich sind dem Datensatz Informationen zu den Einwohnern der jeweiligen Stadt sowie geometrischen Daten in Form von Längen- und Breitengraden. Mit diesen Informationen bietet sich folgende Forschungsfrage an:

Werden Flughäfen in Städten mit mehr Einwohnern häufiger angeflogen, als Städte mit weniger Population?

2 Hauptteil

2.1 Theoretischer Hintergrund

Das Kapitel theoretischer Hintergrund dient dem Verständnis der nachfolgenden Netzwerkanalyse. Es wird unter anderem auf den Datensatz an sich eingegangen, der Netzwerktyp analysiert und Flow Struktur und Zentralitätsmaße erörtert.

2.1.1 Datensatz

Bei dem Datensatz "Airline travel reachability network" handelt es sich um ein Netzwerk für die Erreichbarkeit von Städten in den Vereinigten Staaten und Kanada. Die Knoten im Datensatz bilden die einzelnen Flughäfen der Städte. Die Kanten repräsentieren die Flugverbindung. Diese sind so gewichtet, dass es eine Kante von Stadt i zu Stadt j gibt, wenn die geschätzte Flugreisezeit unter einem Schwellenwert liegt. Die Reisezeit schließt geschätzte Verspätungen bei Zwischenlandungen mit ein. Zusätzlich beinhalet der Datensatz, wie bereits erwähnt, Informationen zu den Einwohnern der jeweiligen Stadt sowie geometrischen Daten in Form von Längen- und Breitengraden.

2.1.2 Netzwerkdefinition

Die Definition des Netzwerktyps beschreibt zum Einen WIE die Daten erhoben worden sind und zum Anderen aber auch WELCHE und WIEVIELE Datenpunkte zur Verfügung stehen. Bei dem zugrundeliegenden Datensatz ist der Ursprung der Erhebung nicht publiziert. Daher lässt sich die Frage ob es sich bei der Erhebung um "Custom-made" Daten oder "Ready-made" Daten handelt nur in der Theorie beantworten. Sollte es sich um Custom-made Daten handeln wurden diese für einen spezifischen Zweck hin gesammelt. Dies kann durch Umfragen, Schneeball Analysen oder mithilfe einer Gesamterhebung stattfinden.

Bei Ready-made Daten ist der Datensatz als Nebenprodukt entstanden. Bei diesem Netzwerk ist es denkbar, dass mithilfe von Web-Crawlern bestimmter APIs die Daten abgezogen worden sind, beispielsweise von Buchungsseiten für Flugverbindungen oder "Arrival-Departure" Seiten.

Weiterhin kann ein Netzwerk über einen Grundtyp definiert werden. Hierbei gibt es einerseits das Ego-Netzwerk, bei dem Beziehnungen eines Akeurs (Ego) zu anderen Akteuren (Alteri) der direkten Netzwerkumgebung, sowie den Beziehungen zwischen den Akteuren analysiert werden. Als zweite Möglichkeit gibt es Schneeball-Netzwerke oder auch referentielles Netzwerk genannt. Hierbei wird mit einem Sample durch eine Erhebung gestartet und dann jeder Akteur zu den Attributdaten befragt. Die Erhebung endet wenn das Netzwerk gesättigt ist. Die Gesamterhebung analysiert alle Elemente eines Netzwerks inklusive deren Attribute und Beziehungen untereinander.

Bei dem "Airline travel reachability network" handelt es sich wahrscheinlich um eine Gesamterhebung aller Flugverbindungen.

2.1.3 Zentralitätsmaß

Um ein Netzwerk beziehungsweise einen Graphen analysieren zu können, ist die Wichtigkeit der Knoten und deren Kanten von hoher Bedeutung. Für die Analyse und die Beantwortung der Forschungsfrage ist die Anwendung von sogenannten Zentralitätsmaßen essentiell.

Die Forschungsfrage "Werden Flughäfen in Städten mit mehr Einwohnern häufiger angeflogen, als Städte mit weniger Population?" wird in zwei Schritten beantwortet. Mithilfe des Zentralitätsmaßes "degree centrality" kann der Knoten mit der höchsten Zentralität im Netzwerk gefunden werden. Dies entspricht der Stadt beziehungsweise Flughafen mit den meisten Kanten, also den meisten Flugverbindungen. Anschließend kann geprüft werden, ob dieser Flughafen auch eine hohe Population hat.

2.2 Datenvorverarbeitung

2.2.1 Laden der Bibliotheken

Als ersten Schritt der Datenvorverarbeitung werden die benötigten Bibliotheken für das Data-Wrangling, DIe Modellierung sowie die Visualisierungen geladen.

```
library("tidyverse")
library("igraph")
library("tidygraph")
library("ggraph")
library("ggplot2")
library("tinytex")
```

2.2.2 Datenimport

Die Datenbasis für das Projekt besteht aus zwei CSV-Dateien, welche mithilfe eines Imports aus dem Ordner "Data" in das Environment geladen werden können. Im Datensatz "Dat" befinden sich neben der node_id alle relevanten Zusatzinformationen wie die Geo-Daten, die Bevölkerung sowie die vollständigen Namen der Städte des Flughafens. "Dat2" beinhaltet zum eine ID-Spalte sowie alle Flugverbindungen als Kantenliste.

```
dat <- read_csv('Data/reachability-meta.csv')
dat2 <- read.table('Data/reachability.txt')
```

Nach erfolgreichem Datenimport und Betrachtung der Daten sind zwei Schritte nötig: Zum Einen müssen die Spalten des Datensatzes "dat2" umbenannt werden und zum Anderen negative Werte in positive umgewandelt werden.

```
dat2 <- dat2 %>%
  rename(
    from = V1,
    weight = V2,
    to = V3
  )
```

```
dat2$to <- dat2$to*(-1)
```

2.3 Daten Exploration

Die Erkundung der zur Verfügung stehenden Datenbasis findet im Kapitel "Daten Exploration" statt. Als ersten Schritt wird jeweils mit dem Befehl "head()" der Kopf des Datensatzes ausgegeben und betrachtet.

Weiterhin kann über "summary()" ein Blick auf die Datentypen und erste Statistiken des Datensatzes geworfen werden. Damit ist es möglich ein erstes Gefühl für die Datenbasis zu bekommen und eventuelle fehlende Daten oder falsche Datentypen in einem "Data-Cleaning" zu beheben. In diesem Fall sind neben der Spaltenumbenennung und der Konvertierung der negativen Werte keine weiteren Manipulationen nötig.

```
head(dat)
```

```
## # A tibble: 6 x 5
##   node_id name             metro_pop latitude longitude
##     <dbl> <chr>                <dbl>    <dbl>     <dbl>
## 1       0 Abbotsford, BC      133497     49.1    -122.
## 2       1 Aberdeen, SD         40878     45.5     -98.5
## 3       2 Abilene, TX         166416     32.4     -99.7
## 4       3 Akron/Canton, OH    701456     40.8     -81.4
## 5       4 Alamosa, CO           9433     37.5    -106.
## 6       5 Albany, GA          157688     31.6     -84.2
```

```
summary(dat)
```

```
##     node_id          name             metro_pop           latitude
##  Min.   :  0.0   Length:456         Min.   :     538   Min.   :19.64
##  1st Qu.:113.8   Class :character   1st Qu.:   49513   1st Qu.:35.22
##  Median :227.5   Mode  :character   Median :  158892   Median :40.79
##  Mean   :227.5                      Mean   :  713095   Mean   :40.69
##  3rd Qu.:341.2                      3rd Qu.:  496238   3rd Qu.:45.47
##  Max.   :455.0                      Max.   :19020000   Max.   :71.29
##    longitude
##  Min.   :-165.40
##  1st Qu.:-110.70
##  Median : -92.45
##  Mean   : -96.72
##  3rd Qu.: -81.24
##  Max.   : -55.61
```

```
head(dat2)
```

```
##   from weight   to
## 1   27      0  757
## 2   57      0   84
## 3   70      0 1290
## 4   74      0  465
## 5   86      0  700
## 6   94      0  526
```

```
summary(dat2)
```

```
##       from           weight            to
##  Min.   :  0.0   Min.   :  0.0   Min.   :  10.0
##  1st Qu.:109.0   1st Qu.:107.0   1st Qu.: 217.0
##  Median :237.0   Median :235.0   Median : 282.0
##  Mean   :227.3   Mean   :226.4   Mean   : 348.1
##  3rd Qu.:342.0   3rd Qu.:340.0   3rd Qu.: 397.0
##  Max.   :455.0   Max.   :455.0   Max.   :2855.0
```

2.4 Modellierung

Ziel des Kapitels Modellierung ist zum Einen ein Netzwerkobjekt als "tbl_graph" zu modellieren und zum Anderen weitere Attribute und Werte zu berechnen, die für die Visualisierungen und die Beantwortung der Forschungsfrage im "Fazit" notwendig sind.

Durch die Funktion "as_tbl_graph" kann die Kantenliste ganz einfach als tbl_object umgewandelt und als "net" gespeichert werden.

```
net <- as_tbl_graph(dat2)
```

Der nächste Schritt in der Netzwerkanalyse ist, wie im Kapitel "Zentralitätsmaße" erwähnt, die Berechnung des Grad der Zentralität. Dafür kann mithilfe der Funktion "degree" ganz einfach der Grad jedes Knoten berechnet werden. Um die berechneten Werte später verwenden zu können, wir das Array zusätzlich noch in einen Data Frame überführt. Außerdem werden die Zeilen mit den 10 höchsten Werten zur Betrachtung ausgegeben.

```
degree <- degree(net)

degree_df <- as.data.frame(degree)

degree_df %>%
  top_n(10)
```

```
##     degree
## 230    873
## 246    662
## 269    648
## 280    704
## 294    618
## 290    676
## 200    618
## 245    618
## 305    616
## 240    667
```

Zur Beantwortung der Forschungsfrage sind die Zusatzinformationen aus dem Datensatz "dat" unerlässlich. Aus diesem Grund werden die beiden Datensätze im nächsten Schritt "gejoined", sodass diese in einem Datensatz vorliegen.

```
degree_df <- rowid_to_column(degree_df, "node_id")

degree_pop_df <- merge(x=degree_df, y=dat, by="node_id")
```

Interssant für die Interpretation der Daten sind die jeweils 10 höchsten Werte von Population und dem Grad der Zentralität. Aus diesem Grund werden diese Im nächsten Schritt berechnet und ausgegeben.

```
calc_top_degree <- top_n(x=degree_pop_df, n=10, wt=degree)
calc_top_degree
```

```
##    node_id degree               name metro_pop latitude  longitude
## 1       13    873       Amarillo, TX    253823 35.20726 -101.83389
```

```
## 2       16    662       Asheville, NC    429017 35.59846  -82.55314
## 3       18    648          Athens, GA    193317 33.95813  -83.37326
## 4       20    704   Atlantic City, NJ    274338 39.36281  -74.42652
## 5       21    618         Augusta, GA    561858 33.47909  -81.97531
## 6       90    676  Corpus Christi, TX    431381 27.79635  -97.40356
## 7      153    618    Grand Rapids, MI    779604 42.96641  -85.67118
## 8      158    618      Greenbrier, WV     35644 37.97558  -80.42690
## 9      163    616 Gulfport/Biloxi, MS    253511 30.41334  -89.07202
## 10     229    667          Laredo, TX    256496 27.53092  -99.50201
```

```
calc_top_pop <- top_n(x=degree_pop_df, n=10, wt=metro_pop)
calc_top_pop
```

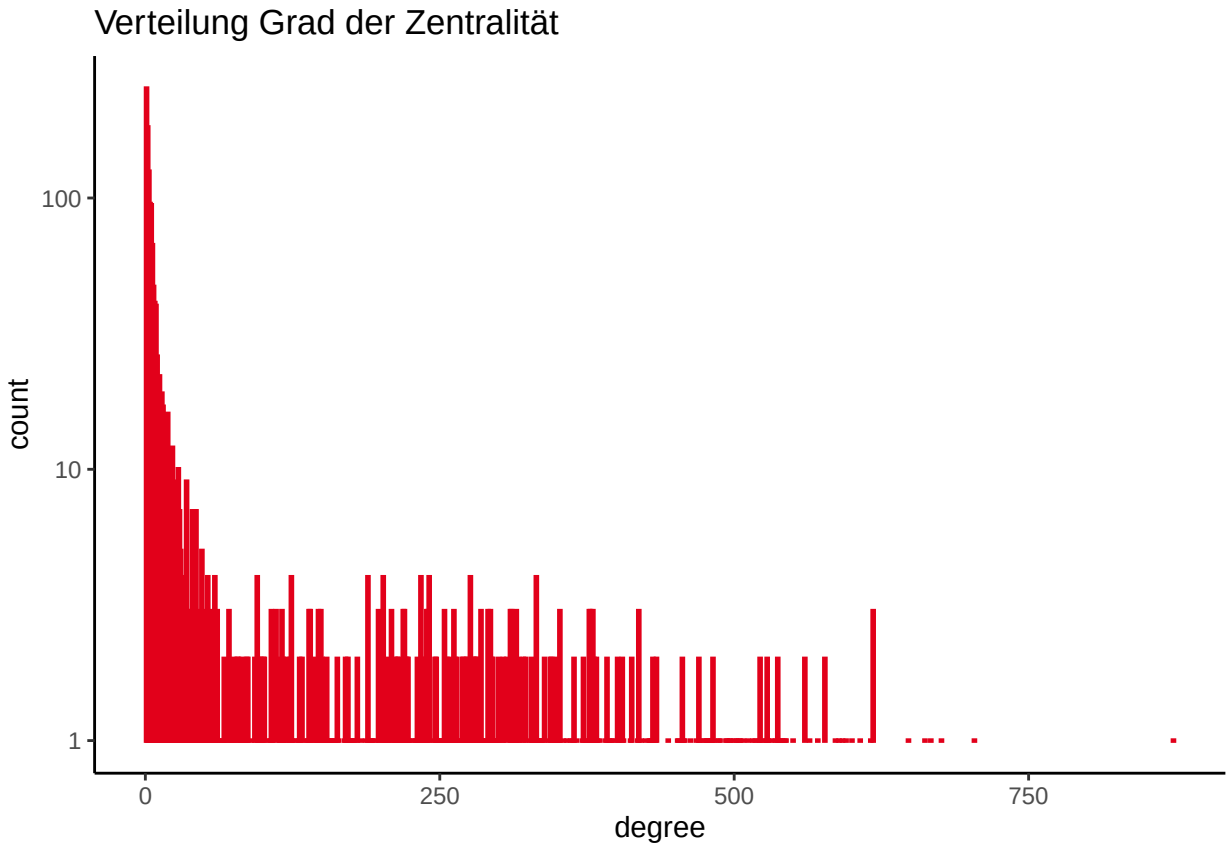
```
##    node_id degree                     name metro_pop latitude  longitude
## 1       53    418              Burbank, CA  12940000 34.18147 -118.30812
## 2       74    392              Chicago, IL   9505000 41.88415  -87.63241
## 3       94    543    Dallas/Fort Worth, TX   6527000 32.92222  -97.04090
## 4      178    380              Houston, TX   6087000 29.76045  -95.36978
## 5      243    342           Long Beach, CA  12940000 33.76642 -118.19239
## 6      244    254 Long Island MacArthur, NY   7568000 40.78913  -73.09839
## 7      246    275          Los Angeles, CA  12940000 34.05329 -118.24501
## 8      294     42             New York, NY  19020000 40.71455  -74.00712
## 9      323    295         Philadelphia, PA   5992000 39.95227  -75.16237
## 10     416    187              Toronto, ON   6324000 43.64856  -79.38533
```

2.5 Visualisierung

Das Kapitel Visualisierung dient der erweiterten Daten Exploration. Mithilfe der nachfolgenden Plots können die Daten noch besser verstanden werden und weitere Erkenntnisse daraus geschlossen werden.

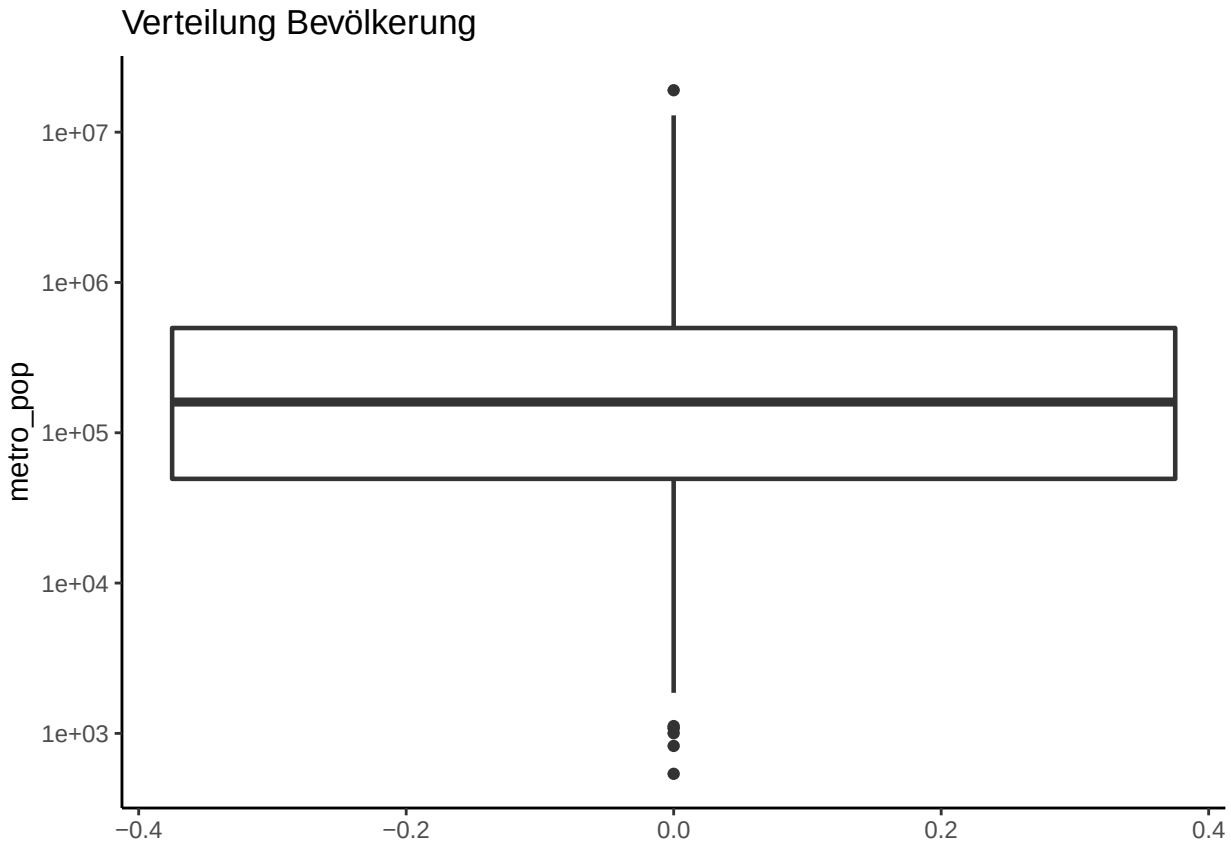
Als ersten Plot wird die Verteilung des Grad der Zentralität für jeden Knoten ausgegeben. Dadurch kann ein Gesamtüberblick erreicht werden. Auffällig hierbei ist, dass es relativ viele Knoten gibt, bei denen der Grad 0 oder sehr klein ist. Interessant ist jedoch auch, dass es Knoten mit sehr großen Werten jenseits von 500 gibt. Diese könnten interessant für die Beantwortung der Forschungsfrage sein.

```r
ggplot(data = degree_df, aes(x = degree)) +
  geom_bar(size = .8, colour="#e2001a") +
  theme_classic() +
  scale_y_log10() +
  ggtitle("Verteilung Grad der Zentralität")
```



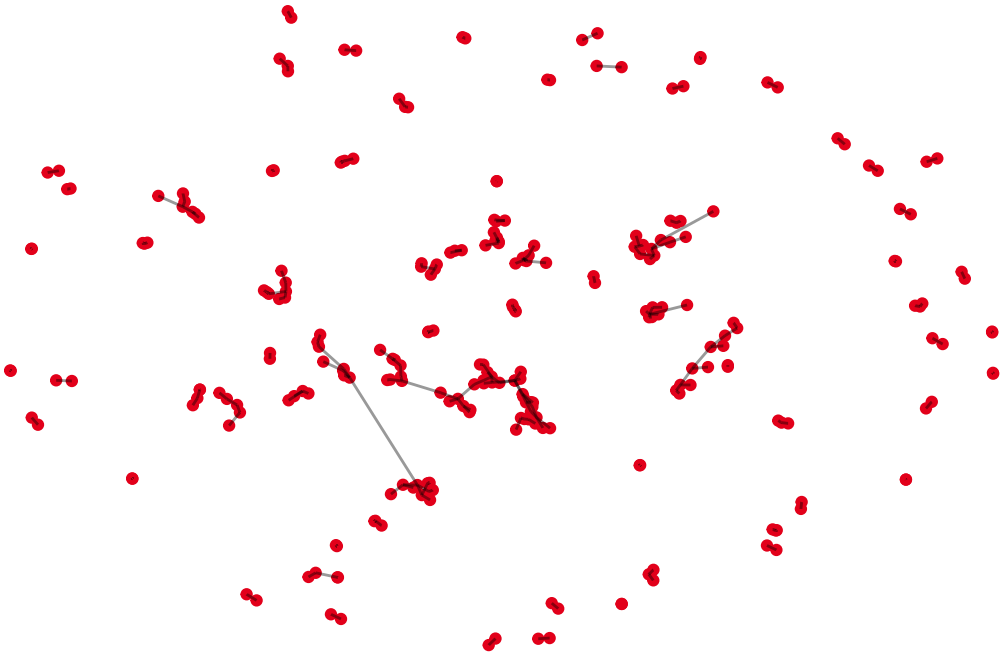
Weiterhin ist die Verteilung der Populationszahlen ein wichtiger Schritt im Verständnis des Datensatzes. Dafür kommt als nächste Visualisierung ein Boxplot zum Einsatz. Dieser zeigt den Durchschnitt der Bevölkerungszahlen sowie Ausreißer nach oben und unten.

```r
ggplot(data = degree_pop_df, aes(y=metro_pop)) +
  geom_boxplot(size = .8) +
  theme_classic() +
  scale_y_log10() +
  ggtitle("Verteilung Bevölkerung")
```

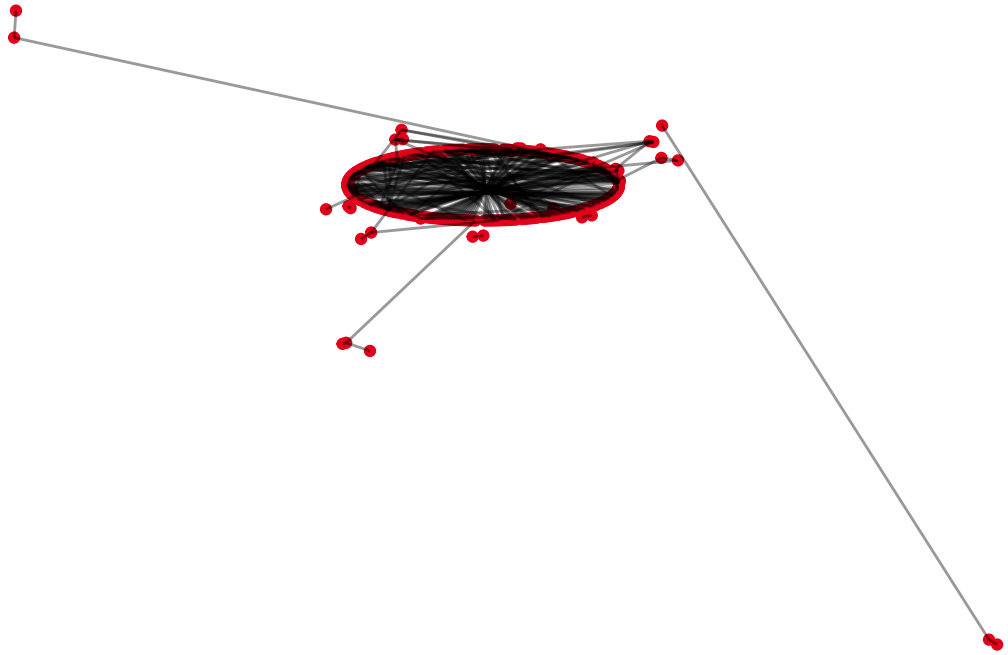


Für die Visualisierung des Netzwerks wird aus Übersichtsgründen als ersten Schritt ein Sample von 200 Datenpunkten extrahiert. Anschließend folgen drei verschiedene Plots des Netzwerks.
Aufrgrund der Tatsache, dass es sich um gewichtete Kanten handelt, ist auffällig, dass die Städte bis auf einige Ausreißer von der Entfernung relativ ähnlich sind. Der dritte Plot zeigt deutlich, dass es verschieden lange Flugverbindungen zwischen den Flughäfen an sich gibt.

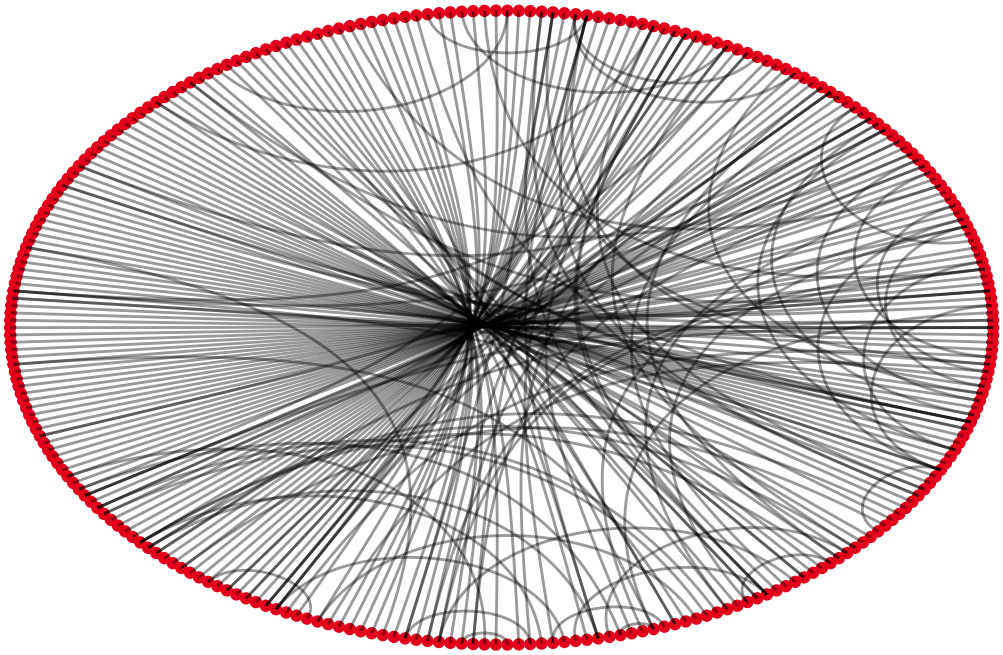
```
dat_sample <- sample_n(dat2, 200)

net_sample <- as_tbl_graph(dat_sample)

ggraph(net_sample, layout = 'fr', maxiter = 100) +
  geom_node_point(colour="#e2001a") +
  geom_edge_link(alpha = .4) +
  theme_graph()
```



```
ggraph(net_sample, layout = 'kk', maxiter = 100) +
  geom_node_point(colour="#e2001a") +
  geom_edge_link(alpha = .4) +
  theme_graph()
```



```
ggraph(net_sample, layout = 'linear', circular = TRUE) +
  geom_node_point(colour="#e2001a") +
  geom_edge_arc(alpha = .4) +
  theme_graph()
```



3 Fazit

3.1 Ergebnisse

Im Folgenden werden nun die Ergebnisse aus den vorangegangenen Kapiteln zusammengefasst und gebündelt. Weitherhin wird eine Antwort auf die Forschungsfrage, ob Flughäfen in Städten mit mehr Einwohnern häufiger angeflogen werden, als Städte mit weniger Population, gegeben.

Das vorangegangene Kapitel zeigt die 10 Städte, bei denen der Grad der Zentralität am höchsten ist. Dies sind also die Flughäfen, die am meisten Kanten also Flugverbindungen haben. Da es sich um gerichtete Kanten handelt, entspricht der Grad den ein- und ausgehenden Kanten. Mit diesen "zentralen" Flughäfen des Datensatzes kann nun also eine Interpretation bezüglich der Bevölkerung in diesen Städten vorgenommen werden.
Dafür werden im Folgenden das globale Maximum und Minimum aller Populationszahlen ermittelt:

```
max_population <- max(degree_pop_df$metro_pop)
min_population <- min(degree_pop_df$metro_pop)
mean_population <- mean(degree_pop_df$metro_pop)

mean_population_df <- data.frame(mean_population)


calc_top_degree$mean_population <- mean_population
```

Das Maximum der Populationszahlen liegt bei:

```
max_population
```

```
## [1] 19020000
```

Und das Minimum liegt bei:

```
min_population
```

```
## [1] 538
```

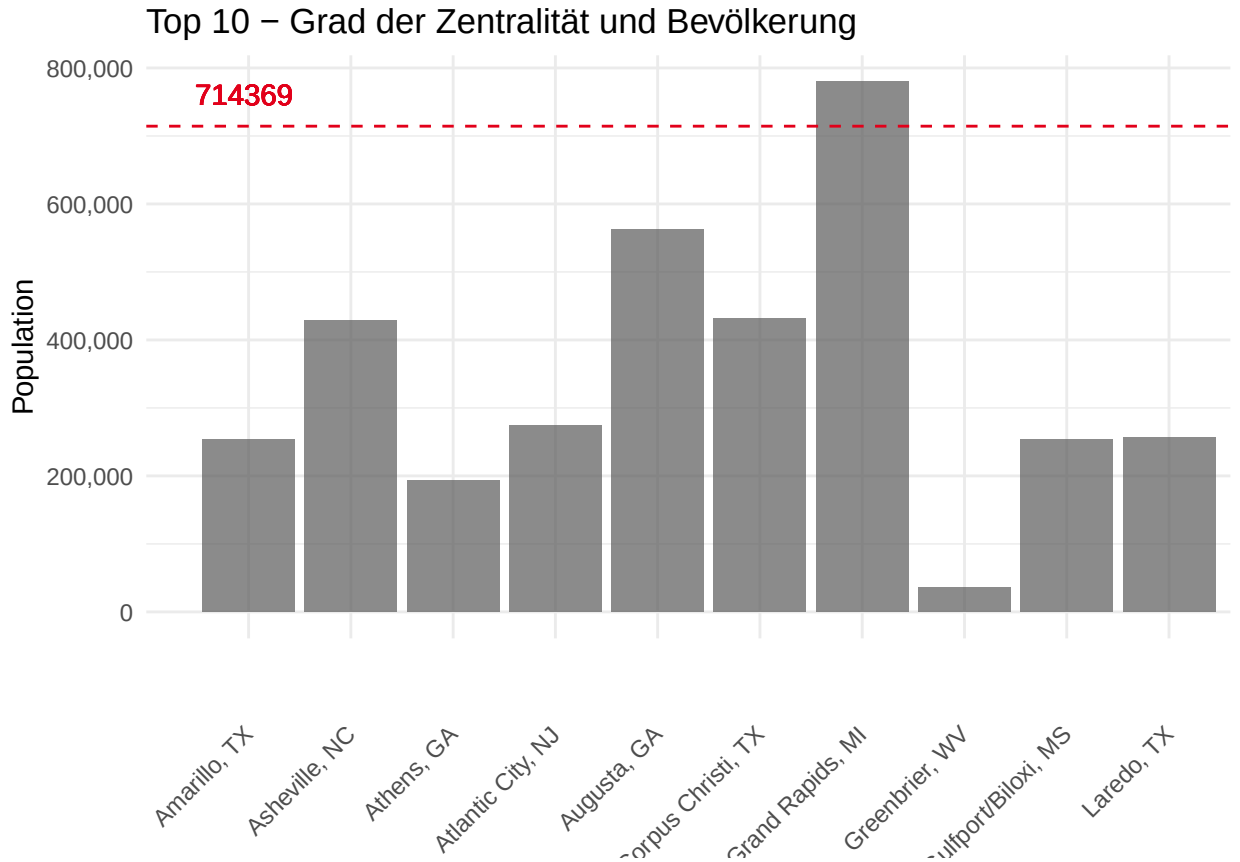
Um zu beurteilen, ob eine Stadt besonders viel oder wenig Bevölkerung hat, wird hier als Metrik das Arithmetische Mittel verwendet. Dieses liegt bei:

```
mean_population
```

```
## [1] 714369.2
```

Um nun die Verbindung aus den Zentralitätsdaten und den Bevölkerungszahlen herzustellen, wird dafür nachfolgend eine Visualisierung verwendet.
Diese zeigt die Top 10 Städte mit dem höchsten Grad der Zentralität im Datensatz sowie das arithmetische Mittel der Population.

```r
ggplot(calc_top_degree, aes(x=name, y=metro_pop))+
  geom_col(size=5, alpha=.7) +
  scale_y_continuous(labels = scales::comma) +
  geom_hline(yintercept=mean_population, colour = "#e2001a", linetype="dashed") +
  geom_text(aes(0,round(mean_population, digits = 0), label = round(mean_population, digits = 0), vjust
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1)) +
  xlab("") +
  ylab("Population") +
  ggtitle("Top 10 - Grad der Zentralität und Bevölkerung")
```



Der Plot zeigt, dass die anfangs aufgestellte Forschungsfrage nicht zutrifft. Die top 10 der "zentralsten" Flughäfen dieses Datensatzes liegen nicht wie vermutet weit über der durchschnittlichen Bevölkerung, sie liegen sogar leicht darunter. Dennoch sind diese Städte keine "kleinen Städte". Mit über 200 000 Einwohnern kann man durchaus von einer Großstadt sprechen. Dies macht auch Sinn, da in größeren Städten nactürlich mehr Personen befördert werden müssen und sich dort zum Teil natürlich auch Umsteigeflughäfen befinden.

3.2 Reflexion

Kritisch reflektiert lässt sich anmerken, dass durch das erzeugen eines Samples von 200 Datenpunkten, die Netzwerkvisualisierungen nicht vollkommen der Realität entsprechen. Dennoch repräsentieren diese einen Teil der Daten und können dennoch repräsentativ betrachtet werden.
Trotz der widerlegten Forschungsfrage ist das Projekt von Datenvorverarbeitung über Daten Exploration, Modellierung und Visualisierungen ohne größere Probleme und Komplikationen verlaufen.