Amazon Network Analysis

Assignment im Rahmen der Vorlesung 'Social Network Analyis' von Philipp Mendoza an der DHBW Stuttgart

Ferdinand Bubeck

Inhaltsverzeichnis

1 Einleitung **2**
- 1.1 Zielsetzung . . . 2
- 1.2 Vorgehensweise . . . 2

2 Hauptteil **3**
- 2.1 Business Understanding . . . 3
- 2.2 Data Understanding . . . 3
- 2.3 Data Preparation . . . 3
- 2.4 Modeling . . . 3
- 2.5 Data Visualization . . . 4
- 2.6 Experimental Data . . . 4

3 Fazit **8**
- 3.1 Evaluation der Ergebnisse . . . 8
- 3.2 kritische Reflexion . . . 8

1 Einleitung

1.1 Zielsetzung

1.2 Vorgehensweise

Als Vorgehensweise wird in diesem Projekt das für das Feld Data Science etablierte Standard-Vorgehen CRISP-DM gewählt (Cross Industry Standard Process for Data Mining). In mehreren Phasen werden so von dem richtigen Verständnis der Daten, dem Data Wrangling und Data Preprocessing bis hin zum Modelfitting und der Evaluation alle entscheidenen Schritte strukturiert durchlaufen, um ein optimales Ergebnis aus den Daten zu generieren. In der Abbildung 1 ist das Vorgehensmodell abgebildet. Da es sich in diesem Projekt um ein PoC handelt, wird die letzte Phase 'Deployment' ausgelassen.

Figure 1: CRISP-DM (Source: https://statistik-dresden.de/archives/1128))

2 Hauptteil

2.1 Business Understanding

- Welche Produkte werden nur in Verbindung mit anderen Produkten gekauft?
- Welche Produkte sind zentral?

2.1.1 Laden der Libraries

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

2.1.2 Importieren der Daten

```
amazon <- read.table("Data/Amazon0302.txt")
```

2.2 Data Understanding

```
head(amazon)
```

```
##   V1 V2
## 1  0  1
## 2  0  2
## 3  0  3
## 4  0  4
## 5  0  5
## 6  1  0
```

```
# Count NAs
which(is.na(amazon))
```

```
## integer(0)
```

2.3 Data Preparation

```
dat <- amazon %>%
  rename(
    from = V1,
    to = V2
  ) %>%
  mutate(
    from = from+1,
    to = to+1
  )
```

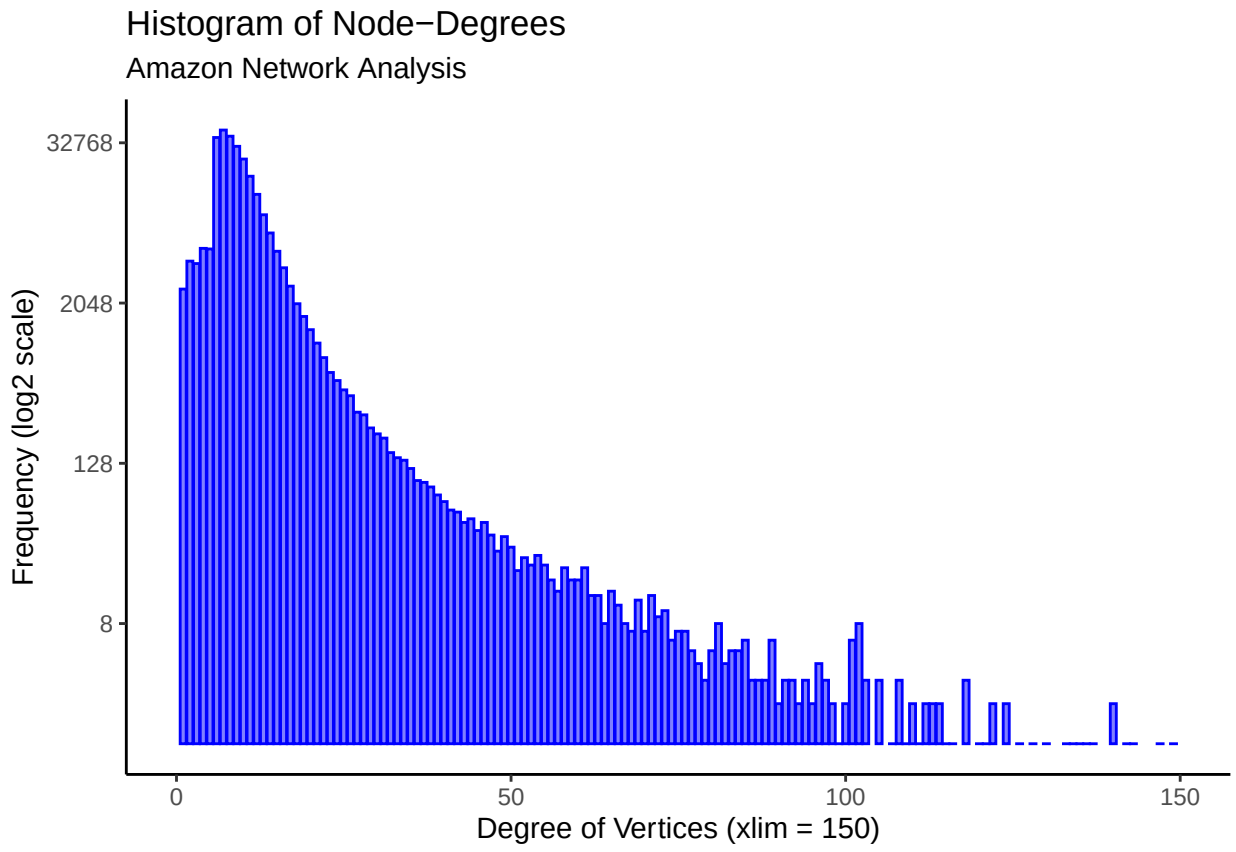
2.4 Modeling

```
net <- as_tbl_graph(dat)


degree <- degree(net)
```

```
# Adjacency Matrix
adjacencyMatrix <- net[]
```

2.5 Data Visualization

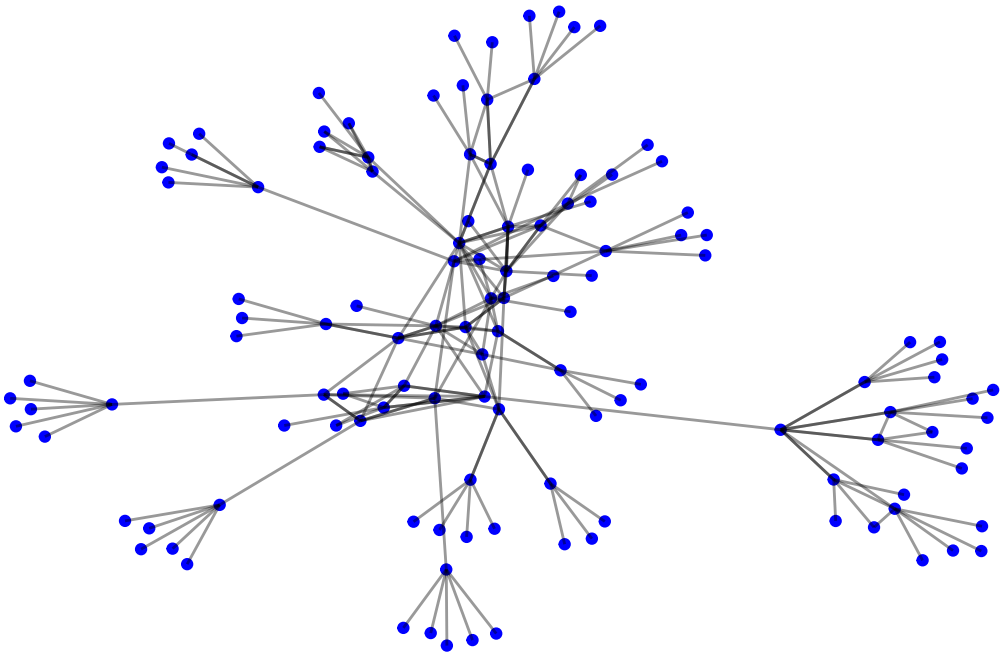
```
degree_df <- as.data.frame(degree)


ggplot(data = degree_df, aes(x=degree))+
  geom_bar(fill = "blue", colour = "blue", alpha=.5)+
  scale_y_continuous(trans='log2')+
  xlim(0,150)+
  labs(title = "Histogram of Node-Degrees", subtitle = "Amazon Network Analysis",
       y = "Frequency (log2 scale)", x = "Degree of Vertices (xlim = 150)")+
  theme_classic()
```



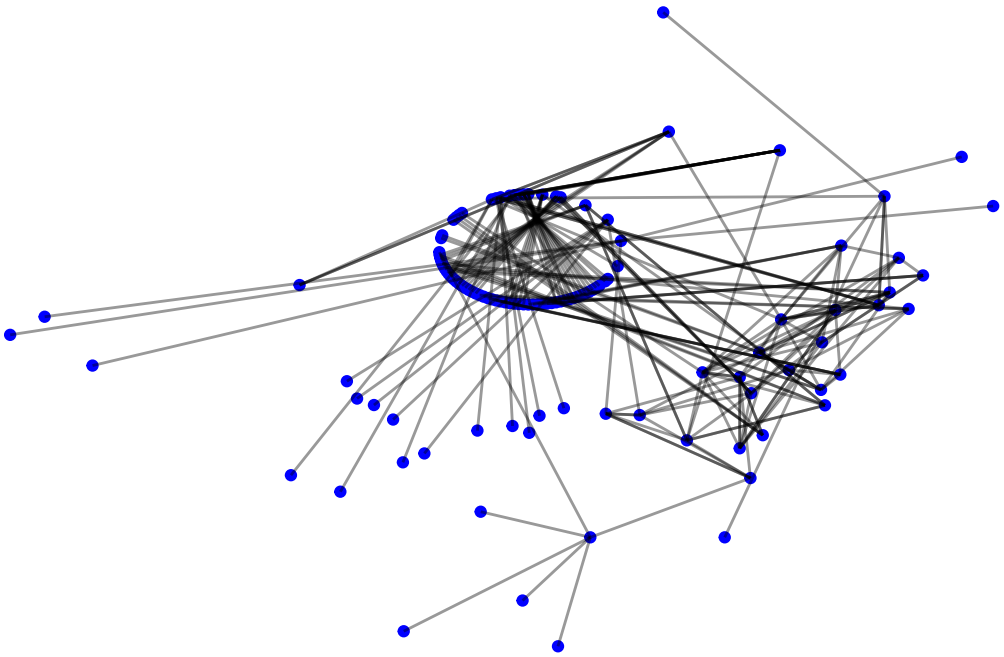
2.6 Experimental Data

```
# Subsetting Data
dat_exp <- dat[1:200,]

net_exp <- as_tbl_graph(dat_exp)

net_exp <- net_exp %>%
  activate(nodes) %>%
```

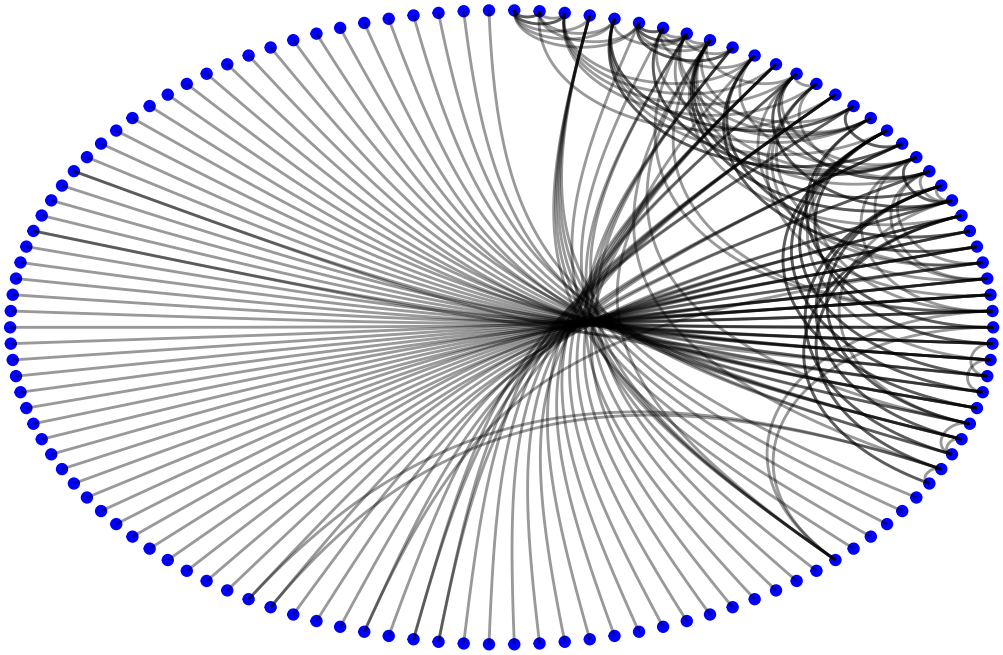
```
  mutate(
    degree = centrality_degree()
  )

# Data Viz for Subset
# network diagramm
ggraph(net_exp, layout = 'fr', maxiter = 100) +
  geom_node_point(colour="blue") +
  geom_edge_link(alpha = .4) +
  theme_graph()
```



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



```
# coord diagramm
ggraph(net_exp, layout = 'linear', circular = TRUE) +
  geom_node_point(colour="blue") +
  geom_edge_arc(alpha = .4) +
  theme_graph()
```



3 Fazit

3.1 Evaluation der Ergebnisse

tbd

3.2 kritische Reflexion

tbd