Social network analysis HW1: export-network

Yijia Lin, Diego Paroli

2025-04-23

Libraries

```r
rm(list = ls())
library(tidyverse)
library(httr2)
library(igraph)
library(visNetwork)
library(tidygraph)
library(ggraph)
```

Get the data

Data has been obtained following the procedure in the chunk below. The code is commented to avoid repeating the download every time the notebok is run.

```r
# zip_data <- request("https://networks.skewed.de/net/product_space/files/SITC.csv.zip") |>
#   req_perform()
#
# writeBin(resp_body_raw(zip_data), "exports_SITC.csv.zip")
#
# unzip("exports_SITC.csv.zip", exdir = "network-data")
#
# file.remove("exports_SITC.csv.zip", )
```

```r
nodes <- read_csv("network-data/nodes.csv")
links <- read_csv("network-data/edges.csv")
```

Description of the dataset

Our network represents economic products. Two products are connected if two or more countries export both products in significant quantities (above world average). The meaning of a link is that two products are connected if the same countries “specialize” in making them, hence, basically, products are connected if they require the same capabilities to be made. Edges weights represent a similarity score (called “proximity”). Data is based on UN Comtrade worldwide trade patterns using the SITC (Standard International Trade Classification) for classifying product categories.

Source: The Product Space.

Properties:

Weighted, Undirected

Nodes and links

```r
head(nodes)
```

```
## # A tibble: 6 x 9
##   `# index`   pid community  size pos                  leamer name  color `_pos`
##       <dbl> <dbl>     <dbl> <dbl> <chr>                 <dbl> <chr> <chr> <chr>
## 1         0  6932         0  48.8 array([4551.8996582~      8 WIRE~ "#9c~ array~
## 2         1  7362         0  65.2 array([ 216.8350982~      9 META~ "#40~ array~
## 3         2  7911         0  54.0 array([ 538.9149017~      9 RAIL~ "#40~ array~
## 4         3  8946         0  57.7 array([ 696.3942565~      7 NON-~ "#40~ array~
## 5         4  7264         0  73.3 array([  57.2840652~      9 PRIN~ "#40~ array~
## 6         5  2783         0  58.3 array([4662.2502441~      2 COMM~ "#ff~ array~
```

```r
head(links)
```

```
## # A tibble: 6 x 4
##   `# source` target width color
##        <dbl>  <dbl> <dbl> <chr>
## 1          1    328  5.58 "#727272\n"
## 2          4    475  6.36 "#7b7b7b\n"
## 3          6     69  5.71 "#737373\n"
## 4          8     18  5.12 "#6c6c6c\n"
## 5          8      9  3.72 "#545454\n"
## 6         10    480  8.92 "#949494\n"
```

Most of the columns in the `nodes` dataset will not be useful for us. The column `# index` is the one indicating the index of the nodes and the column `name` is the one indicating their corresponding name. In the `links` dataframe, the column `# source` and `target` indicate the 2 nodes forming a link, while `width` is the weight of that link.

Graph:

```r
graph <- graph_from_data_frame(links, directed = FALSE, vertices = nodes)
graph
```

```
## IGRAPH 9855ac1 UN-- 774 1779 --
## + attr: name (v/c), pid (v/n), community (v/n), size (v/n), pos (v/c),
## | leamer (v/n), color (v/c), _pos (v/c), width (e/n), color (e/c)
## + edges from 9855ac1 (vertex names):
## [1] METAL FORMING MACHINE TOOLS                    --CONVERTERS,LADLES,INGOT MOULDS AND CASTING MACH
## [2] PRINTING PRESSES                               --OTHER MACH.-TOOLS FOR WORKING METAL OR MET.CARBI
## [3] OTHER FOOD PROCESSING MACHINERY AND PARTS      --PARTS OF THE MACHINERY OF 744.2-
## [4] PRODUCER GAS AND WATER GAS GENERATORS AND PARTS--OTHER PUMPS FOR LIQUIDS & LIQUID ELEVATORS
## [5] PRODUCER GAS AND WATER GAS GENERATORS AND PARTS--CINEMATOGRAPHIC CAMERAS,PROJECTORS,SOUND-REC,PAR
## + ... omitted several edges
```

Questions

1. What is the number of nodes and links?

```r
vcount(graph)
```

```
## [1] 774
```

```r
ecount(graph)
```

```
## [1] 1779
```

There are in total 774 nodes (i.e. economic product) and 1779 links in this network.

```r
components <- components(graph)
head(components$no)
```

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

Our graph is fully connected i.e. it forms one single connected component.

2. What is the average degree in the network? And the standard deviation of the degree?

```r
mean(degree(graph))
```

```
## [1] 4.596899
```

```r
sd(degree(graph))
```

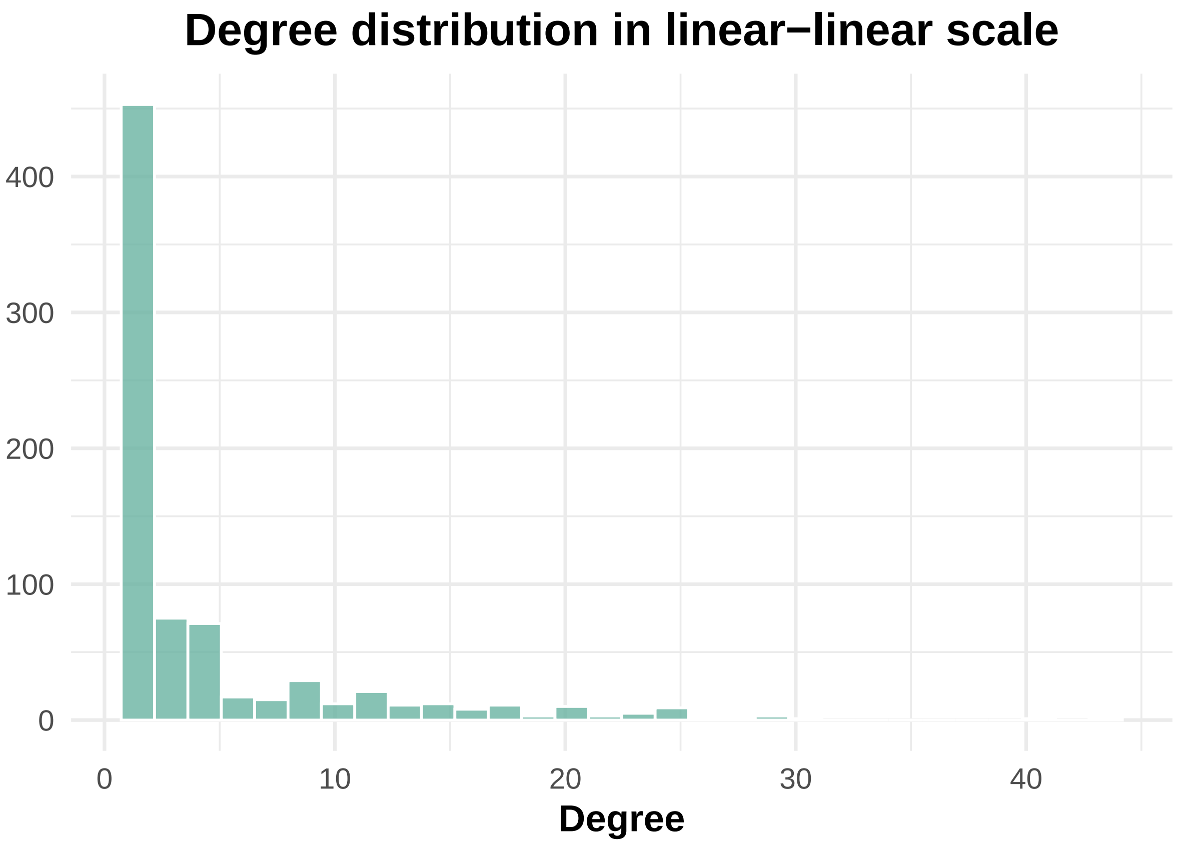
```
## [1] 5.994848
```

The average degree is 4.5969 in this network, with a standard deviation of 5.9948.

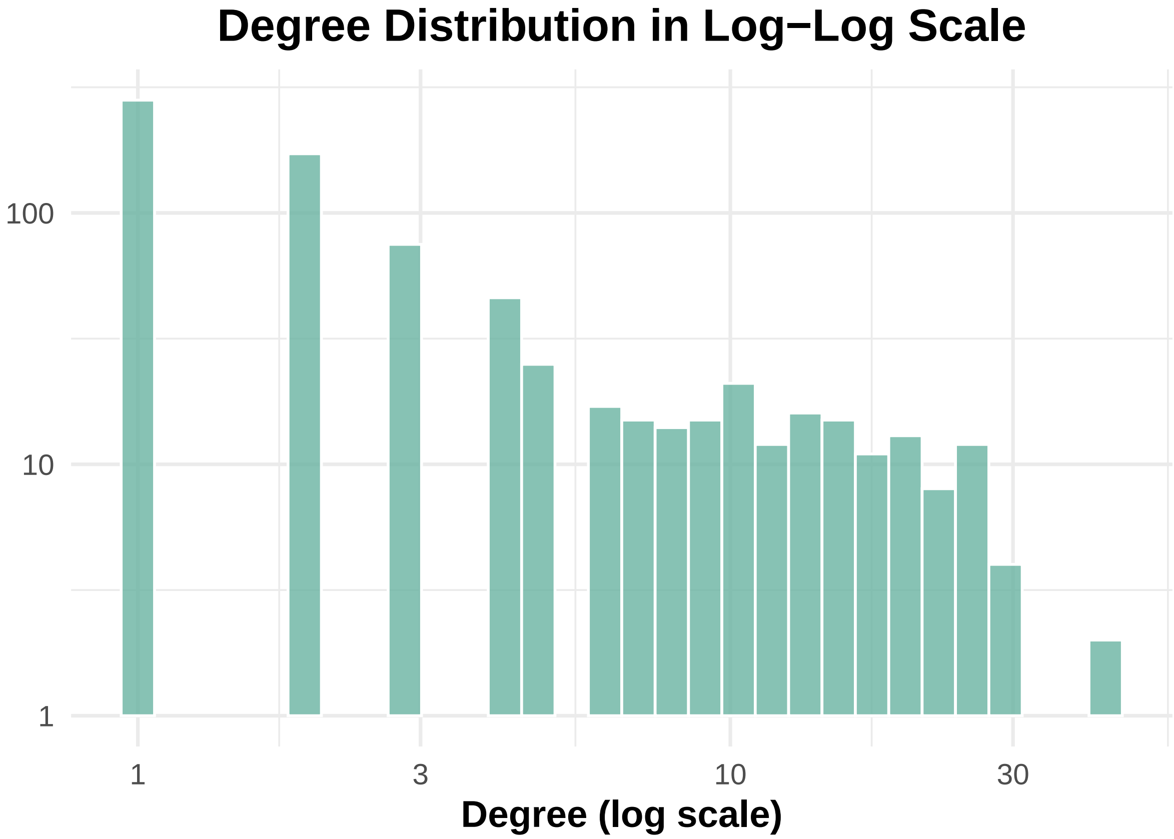
This means that on average every product is connected to 4/5 other products. Standard deviation seems to be high (higher than the mean), therefore indicating that there could be both products with a lot of edges and products with just one edge.

3. Plot the degree distribution in linear-linear scale and in log-log-scale. Does it have a typical connectivity? What is the degree of the most connected node?

```r
# In linear-linear scale
ggplot() +
  geom_histogram(aes(x = degree(graph)),
                 fill = "#69b3a2", color = "white", alpha = 0.8) +
  labs(x = "Degree",
       y = "",
       title = "Degree distribution in linear-linear scale") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )
```



```r
# In log-log scale
ggplot() +
  geom_histogram(aes(x = degree(graph)),
                 fill = "#69b3a2", color = "white", alpha = 0.8) +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    x = "Degree (log scale)",
    y = "",
    title = "Degree Distribution in Log-Log Scale"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )
```



We can observe that this network **does not exhibit typical connectivity**: Its degree distribution is highly skewed and lacks a clear peak. By being its degree distribution so broad there is not a typical connectivity for a node. Most nodes have a very low degree, while a few have very high degree. We observe a power-law-like distribution, rather than a Poisson-like distribution, where most nodes would have approximately the same number of links and no hubs. This is what can usually be observed in social networks, but our product network exhibits the same, indicating that some, but few products are exported together with many different products by many countries, while the majority of products are exported only by a few countries or by highly specialized countries.

```
max_degree(graph)
```

```
## [1] 43
```

The most connected node on our network has a degree of 43.

```
which.max(degree(graph))
```

```
## TRANSMISSION SHAFTS,CRANKS,BEARING HOUSINGS ETC.
##                                               580
```

The node with the highest degree is that of mechanical components, indicating that there is no high specialization required to product this. These are indeed products that are probably going to be used to assemble more sophisticated goods and thus the demand for this product is high everywhere and its supply chain is likely to be very globalized with many countries producing them.

4. What is the clustering coefficient (transitivity) in the network?

```
transitivity(graph, type = "global")
```

```
## [1] 0.429691
```

The global transitivity of this network is 0.4297, which is closer to 0 than to 1, but still indicating a tendency of clustering. A bit less than half of the time, when two nodes share a common neighbor, they will also be directly connected to each other. This could indicate that often, but not always countries tend to specialize on the same products.

5. What is the assortativity (degree) in the network?

```
assortativity_degree(graph)
```

```
## [1] 0.4571059
```

The assortativity coefficient of this network is 0.4571 (greater than 0), indicating a moderate to strong tendency for nodes to connect with others that have a similar degree. In other words, high-degree nodes tend to connect with other high-degree nodes, and low-degree nodes tend to connect with other low-degree nodes. This is a sign of assortative mixing.

6. Using the Louvain method, does the network have a community structure? If so, what is its modularity?

```
louvain_cluster <- cluster_louvain(graph, weights = E(graph)$width)

sizes(louvain_cluster)
```

```
## Community sizes
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20 
##  57 106  10  58 134  66  99  14  45   8  18   4  17  16   3   4   7  13  13  25 
##  21  22  23  24  25  26  27  28  29  30 
##   5   5   6   8   7   9   6   3   5   3
```

```
modularity(louvain_cluster)
```

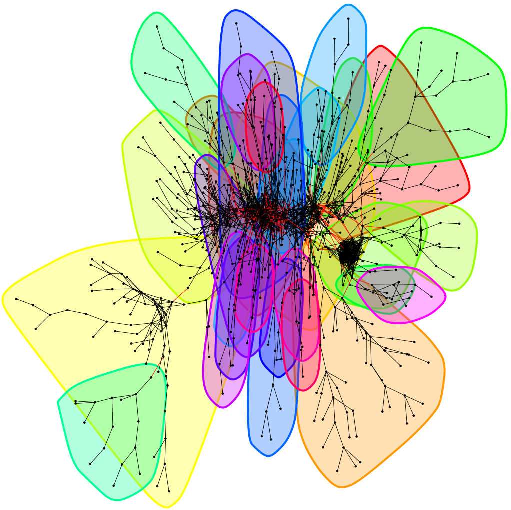
```
## [1] 0.7457563
```

Yes, the network has a clear community structure. Using the Louvain method, the network was partitioned into more than 20 communities. The modularity value is >**0.7**, which is considered high and indicates a strong modular (community) structure within the network.

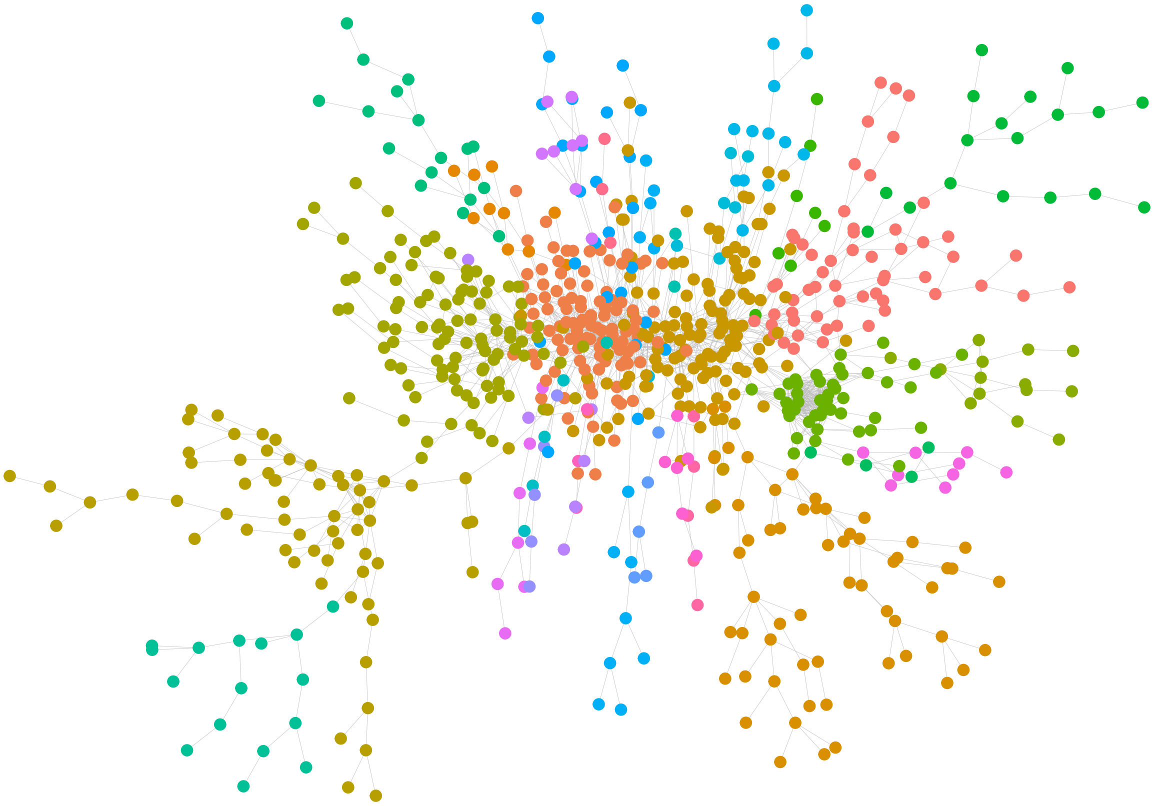
We can also try to visualize these communities:

```
layout <- layout_with_kk(graph)

# Using igraph only
plot(louvain_cluster, graph,
     layout = layout,
     vertex.size = 0.1,
     edge.width = 0.1,
     vertex.label = "")
```



```
# Using ggraph
graph_tbl <- as_tbl_graph(graph)
V(graph_tbl)$community <- membership(louvain_cluster)
ggraph(graph_tbl, layout = "kk") +
  geom_edge_link(width = 0.1, alpha = 0.7, color = "grey") +
  geom_node_point(aes(color = as.factor(community)), size = 1.5) +
  theme_void() +
  theme(legend.position = "none")
```



7. Test that the clustering coefficient in the network cannot be statistically explain by a configuration model in which the nodes have the same degree distribution as the original.

```r
original_clustering <- transitivity(graph, type = "global")
```

Create 1000 random configuration models and register the clustering coefficient for each one of this generated model.

```r
num_simulations <- 1000
# Initializing a vector
simulated_clustering <- numeric(num_simulations)
# Simulating the graphs and saving their transitivity
for (i in 1:num_simulations) {
  config_graph <- sample_degseq(degree(graph), method = "vl")
  simulated_clustering[i] <- transitivity(config_graph, type = "global")
}
```

Here we use the method "vl" (Viger–Latapy) as it is the one suggested for undirected connected graphs. We simulated 1000 graphs to ensure that we can statistically meaningful conclusions.

```
original_clustering
```

```
## [1] 0.429691
```

```
mean(simulated_clustering)
```

```
## [1] 0.03399763
```

The two values seem to be really different from each other, but we want to proceed with a more statistically sound test (t-test), checking whether the clustering coefficients obtained under a random graph with the same degree distribution are significantly lower than the original clustering coefficient.

```
t.test(x = simulated_clustering, mu = original_clustering, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  simulated_clustering
## t = -5004.9, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 0.429691
## 95 percent confidence interval:
##       -Inf 0.0341278
## sample estimates:
##  mean of x
## 0.03399763
```

We tested whether the clustering coefficient of the original network can be explained solely by its degree distribution by comparing it to 1000 configuration model networks with the same degree distribution. The original network's clustering coefficient was **0.42**, while the average clustering coefficient from the configuration models was **0.034**. The result of the t-test showed that the synthetic values are significantly lower than the original value under the null model (p-value < 0.01). This allows us to reject the null hypothesis, and conclude that the clustering structure in the original network **cannot be explained** by degree distribution alone — it has significant non-random structure peculiar to the network. Therefore, the network exhibits a structural clustering that goes beyond what would be expected by random connections among nodes with the same degrees.

One well-known limitation of configuration models is that they tend to produce networks with lower clustering than what we observe in real-world systems. Our test confirmed this in our case as well: the real-world product export network shows a much higher clustering coefficient than the networks generated by the configuration model.

8. Visualize the neighborhood of the node with the largest centrality (closeness)

```
which.max(closeness(graph))
```

```
## SLAG WOOL.ROCK WOOL AND SIMILAR MINERAL WOOLS
##                                           453
```

We discovered that the node with the largest centrality/closeness is "SLAG WOOL.ROCK WOOL AND SIMILAR MINERAL WOOLS". Apparently these are insulating materials coming from minerals. This node is the one that on average is closest to all other nodes. In our context this could be interpreted as a product which does not require particular specialization as from it it is possible to reach all other products in just a few steps.

```
head(neighbors(graph,"SLAG WOOL.ROCK WOOL AND SIMILAR MINERAL WOOLS"))
```

```
## + 6/774 vertices, named, from 9855ac1:
## [1] TRAILERS & SPECIALLY DESIGNED CONTAINERS
## [2] PARTS OF THE MACHINERY OF 723.41 TO 723.46
## [3] MATERIALS OF RUBBER(E.G.,PASTES.PLATES,SHEETS,ETC)
## [4] OTHER VEHICLES,NOT MECHANICALLY PROPELLED,PARTS
## [5] PARTS OF THE MACHINERY OF 744.2-
## [6] MISCELLANEOUS ART.OF MATERIALS OF DIV.58
```

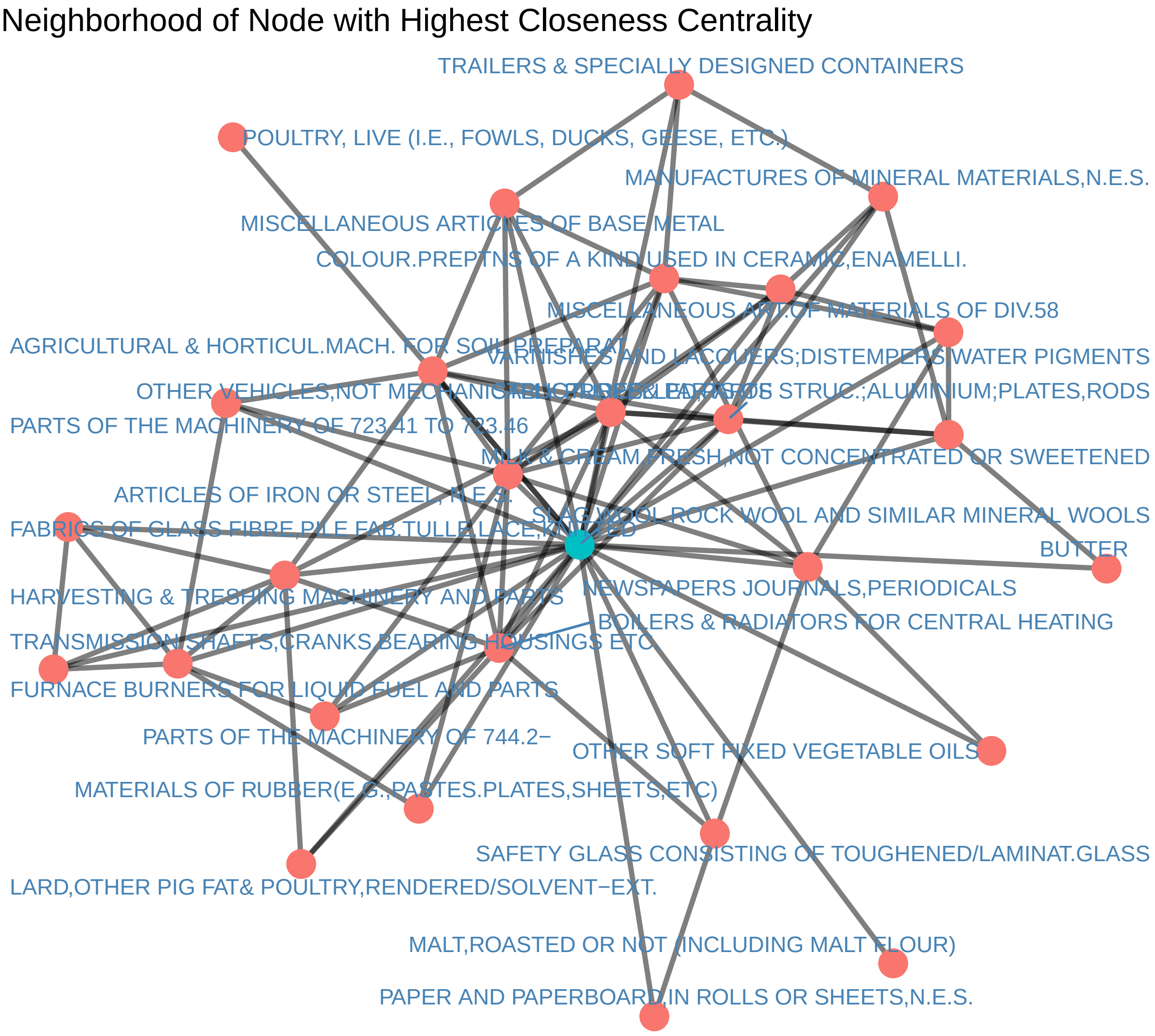
These are some of the neighbors of our node of interest.

We'll now plot its neighbors.

```
neigh_graph <- make_neighborhood_graph(graph,
                                       order = 1,
                                       "SLAG WOOL.ROCK WOOL AND SIMILAR MINERAL WOOLS")[[1]]

g <- as_tbl_graph(neigh_graph)

ggraph(g, layout = 'fr') +
  geom_edge_link(alpha = 0.5, width = 1) +
  # Colouring of a different colour our central node
  geom_node_point(size = 5,
                  aes(color = name == "SLAG WOOL.ROCK WOOL AND SIMILAR MINERAL WOOLS")) +
  geom_node_text(aes(label = name),
                 repel = TRUE,
                 size = 3,
                 color = "steelblue",
                 check_overlap = TRUE) +
  guides(color = "none") +
  theme_void()+
  labs(title = "Neighborhood of Node with Highest Closeness Centrality")
```


Neighborhood of Node with Highest Closeness Centrality
TRAILERS & SPECIALLY DESIGNED CONTAINERS
POULTRY, LIVE (I.E., FOWLS, DUCKS, GEESE, ETC.)
MANUFACTURES OF MINERAL MATERIALS,N.E.S.
MISCELLANEOUS ARTICLES OF BASE METAL
COLOUR.PREPTNS OF A KIND USED IN CERAMIC,ENAMELLI.
MISCELLANEOUS ART.OF MATERIALS OF DIV.58
AGRICULTURAL & HORTICUL.MACH. FOR SOIL PREPARAT.
VARNISHES AND LACQUERS;DISTEMPERS;WATER PIGMENTS
OTHER VEHICLES,NOT MECHANICALLY PROPELLED,PARTS
STRUCTURES & PARTS OF STRUC.;ALUMINIUM;PLATES,RODS
PARTS OF THE MACHINERY OF 723.41 TO 723.46
MILK & CREAM,FRESH,NOT CONCENTRATED OR SWEETENED
ARTICLES OF IRON OR STEEL, N.E.S.
SLAG WOOL,ROCK WOOL AND SIMILAR MINERAL WOOLS
FABRICS OF GLASS FIBRE,PILE FAB.TULLE,LACE,KNITTED
BUTTER
NEWSPAPERS JOURNALS,PERIODICALS
HARVESTING & TRESHING MACHINERY AND PARTS
BOILERS & RADIATORS FOR CENTRAL HEATING
TRANSMISSION SHAFTS,CRANKS,BEARING HOUSINGS ETC.
FURNACE BURNERS FOR LIQUID FUEL AND PARTS
PARTS OF THE MACHINERY OF 744.2−
OTHER SOFT FIXED VEGETABLE OILS
MATERIALS OF RUBBER(E.G., PASTES.PLATES,SHEETS,ETC)
SAFETY GLASS CONSISTING OF TOUGHENED/LAMINAT.GLASS
LARD,OTHER PIG FAT& POULTRY,RENDERED/SOLVENT−EXT.
MALT,ROASTED OR NOT (INCLUDING MALT FLOUR)
PAPER AND PAPERBOARD,IN ROLLS OR SHEETS,N.E.S.

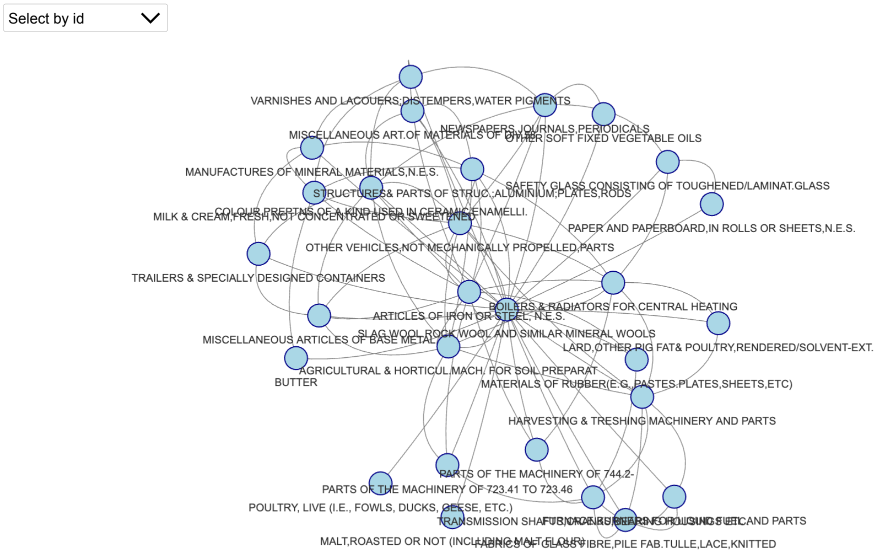
This plot looks a little bit messy, as the labels are quite long and difficult to read. However, if we hide all the labels, we lose a lot of valuable information. Therefore, we decided to use the `visNetwork` package to create an **interactive graph**, which allows us to freely drag nodes around and better explore the names of each node.

It is hard to find any pattern for the neighbors of our central node. Lots of them seem to be agricultural products that do not require particular skills or conditions to be produced, the rest of them seem to be construction materials or other intermediate goods/machinery.

```r
library(visNetwork)
nodes <- data.frame(id = V(g)$name, label = V(g)$name)
edges <- data.frame(from = as.character(ends(g, E(g))[,1]),
                    to = as.character(ends(g, E(g))[,2]))

# NOTE: In the `.pdf` document we are only able to show a screenshot of the graph
visNetwork(nodes, edges) %>%
  visEdges(color = list(color = "gray", hover = "red")) %>%
  visNodes(size = 15, color = list(background = "lightblue", border = "darkblue")) %>%
  visOptions(highlightNearest = TRUE, nodesIdSelection = TRUE) %>%
  visLayout(randomSeed = 123)
```

```
## file:////private/var/folders/5k/zh67w_yx2fj10t7z34l2ct7m0000gn/T/RtmpOUVRcY/file1db537c336f5/widget1
```



Select by id
VARNISHES AND LACOUERS;DISTEMPERS,WATER PIGMENTS
NEWSPAPERS,JOURNALS,PERIODICALS
MISCELLANEOUS ART.OF MATERIALS OF DIV.58
OTHER SOFT FIXED VEGETABLE OILS
MANUFACTURES OF MINERAL MATERIALS,N.E.S.
SAFETY GLASS CONSISTING OF TOUGHENED/LAMINAT.GLASS
STRUCTURES& PARTS OF STRUC.;ALUMINIUM;PLATES,RODS
COLOUR PREPTNS.OF A KIND USED IN CERAMIC,ENAMELLI.
MILK & CREAM,FRESH,NOT CONCENTRATED OR SWEETENED
PAPER AND PAPERBOARD,IN ROLLS OR SHEETS,N.E.S.
OTHER VEHICLES,NOT MECHANICALLY PROPELLED,PARTS
TRAILERS & SPECIALLY DESIGNED CONTAINERS
BOILERS & RADIATORS FOR CENTRAL HEATING
ARTICLES OF IRON OR STEEL, N.E.S.
SLAG WOOL,ROCK WOOL AND SIMILAR MINERAL WOOLS
MISCELLANEOUS ARTICLES OF BASE METAL
LARD,OTHER PIG FAT& POULTRY,RENDERED/SOLVENT-EXT.
AGRICULTURAL & HORTICUL.MACH. FOR SOIL PREPARAT
BUTTER
MATERIALS OF RUBBER(E.G.,PASTES.PLATES,SHEETS,ETC)
HARVESTING & TRESHING MACHINERY AND PARTS
PARTS OF THE MACHINERY OF 744.2-
PARTS OF THE MACHINERY OF 723.41 TO 723.46
POULTRY, LIVE (I.E., FOWLS, DUCKS, GEESE, ETC.)
TRANSMISSION SHAFTS,CRANKS,BEARING HOUSINGS ETC.
FURNACE BURNERS FOR LIQUID FUEL AND PARTS
MALT,ROASTED OR NOT (INCLUDING MALT FLOUR)
FABRICS OF GLASS FIBRE,PILE FAB.TULLE,LACE,KNITTED