Statistical methods for network data

Network Analysis: UK trains

Lorenzo Saracino - Biagio Buono

The dataset

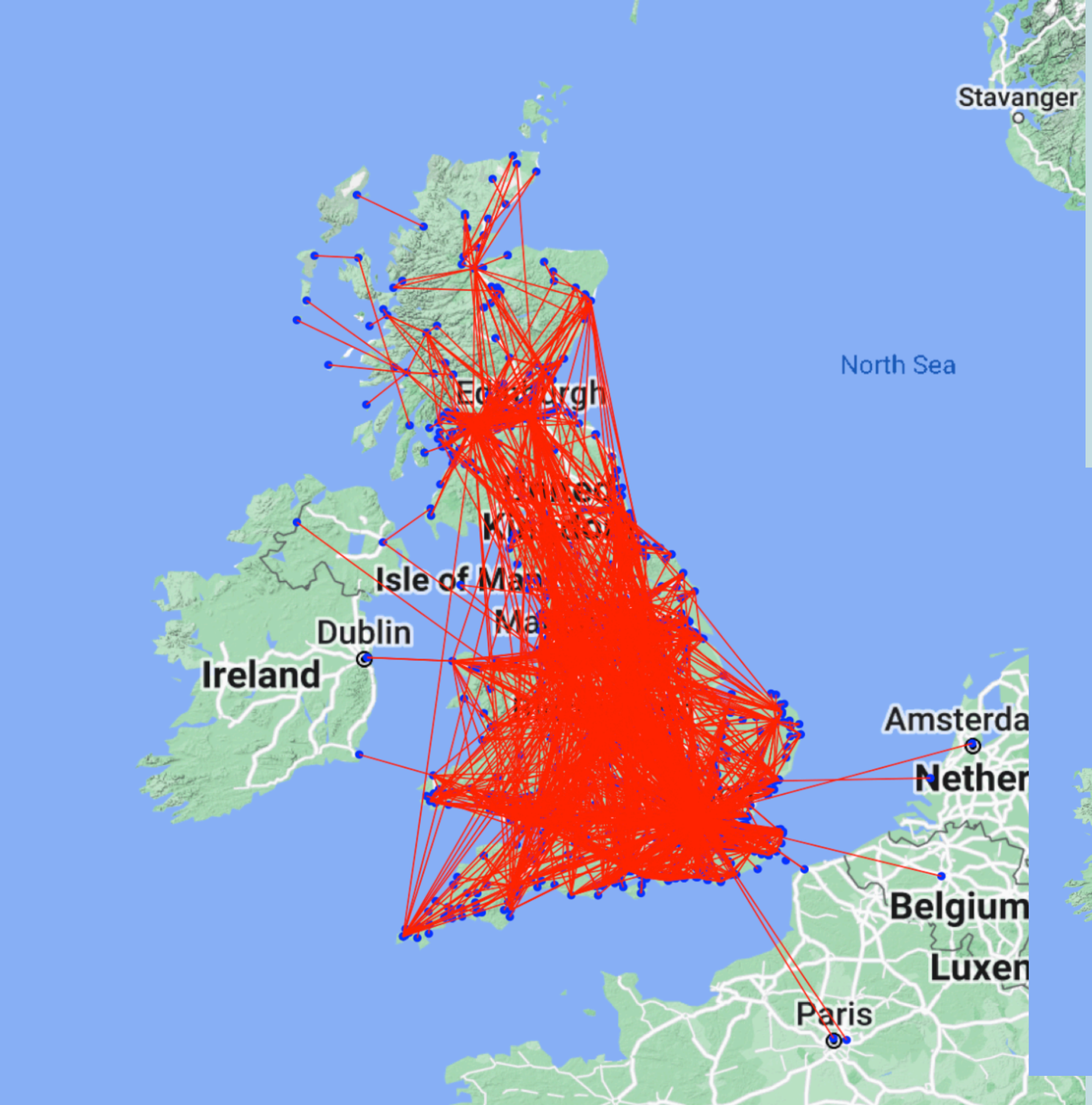
- 41 railway companies
- 2944 railway stations of the United Kingdom
- Nodes: stations
- Edges: routes between stations
- Directed network dataset: routes have direction, represented by ordered pairs (u,v), where (u,v) != (v,u)

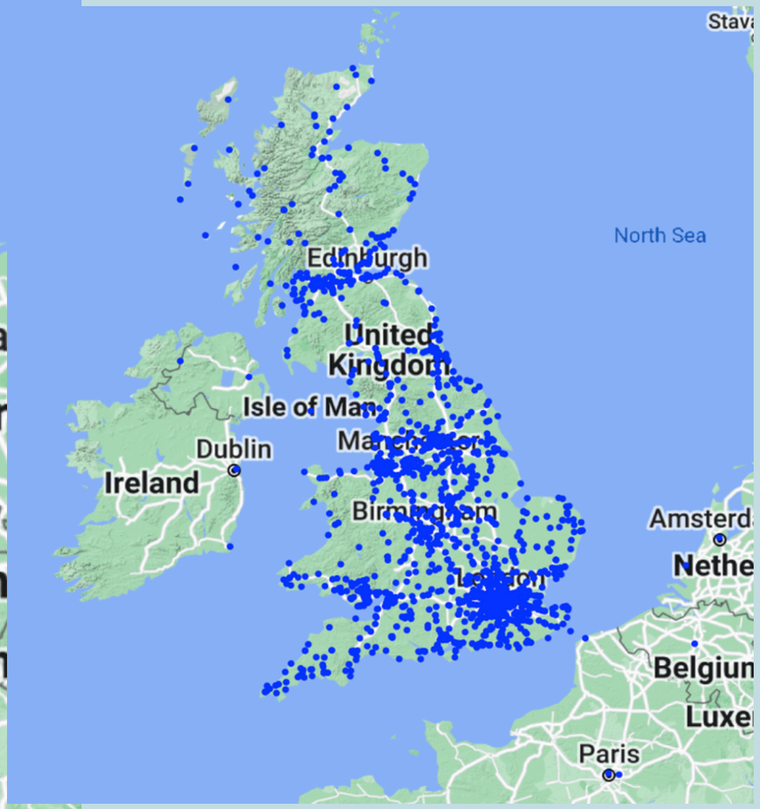
	station
1	3Bdgs Down Thameslink Sdgs
2	Abbey Foregate C.S.
3	Abercynon
4	Aberdare
5	Aberdeen
6	Aberdeen Clayhills Car.M.D

	Source	Target	Company
1	Birmingham International	Chester	Alliance Rail
2	Chester	Wolverhampton	Arriva Trains Wales
3	Cardiff Central Bus Station	Cardiff International APT	Arriva Trains Wales
4	Cardiff Central	Canton C.S.D.	Arriva Trains Wales
5	Birmingham New Street	Aberystwyth	Arriva Trains Wales
6	Milford Haven	Cardiff Central	Arriva Trains Wales

Data Exploration

- Most part of the stations are concentrated in England and Wales
- Only few routes between the bounding stations of the network





Data preparation

- Check for missing values: no NA ✅
- Check for the presence of any duplicated routes in the edges data frame

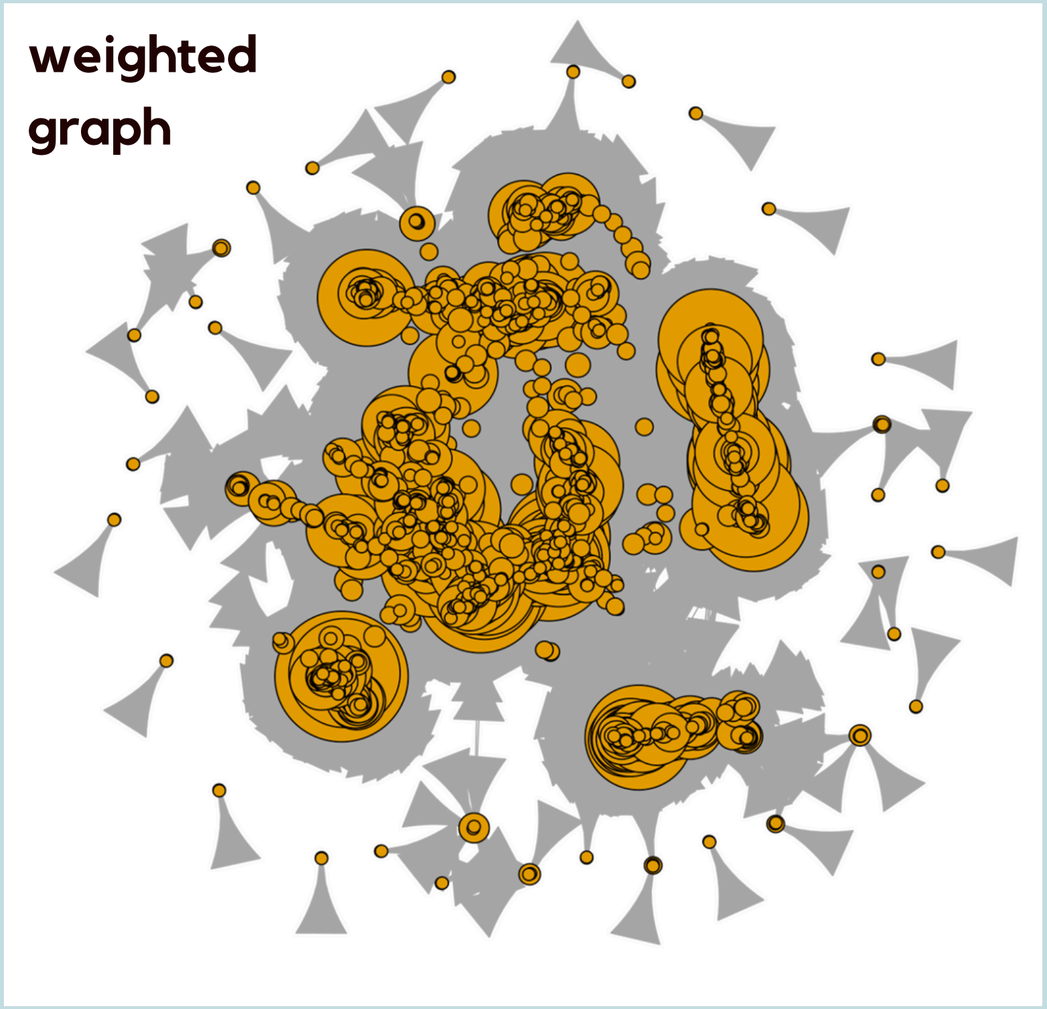
Creation of a dataset containing the companies operating on each route:

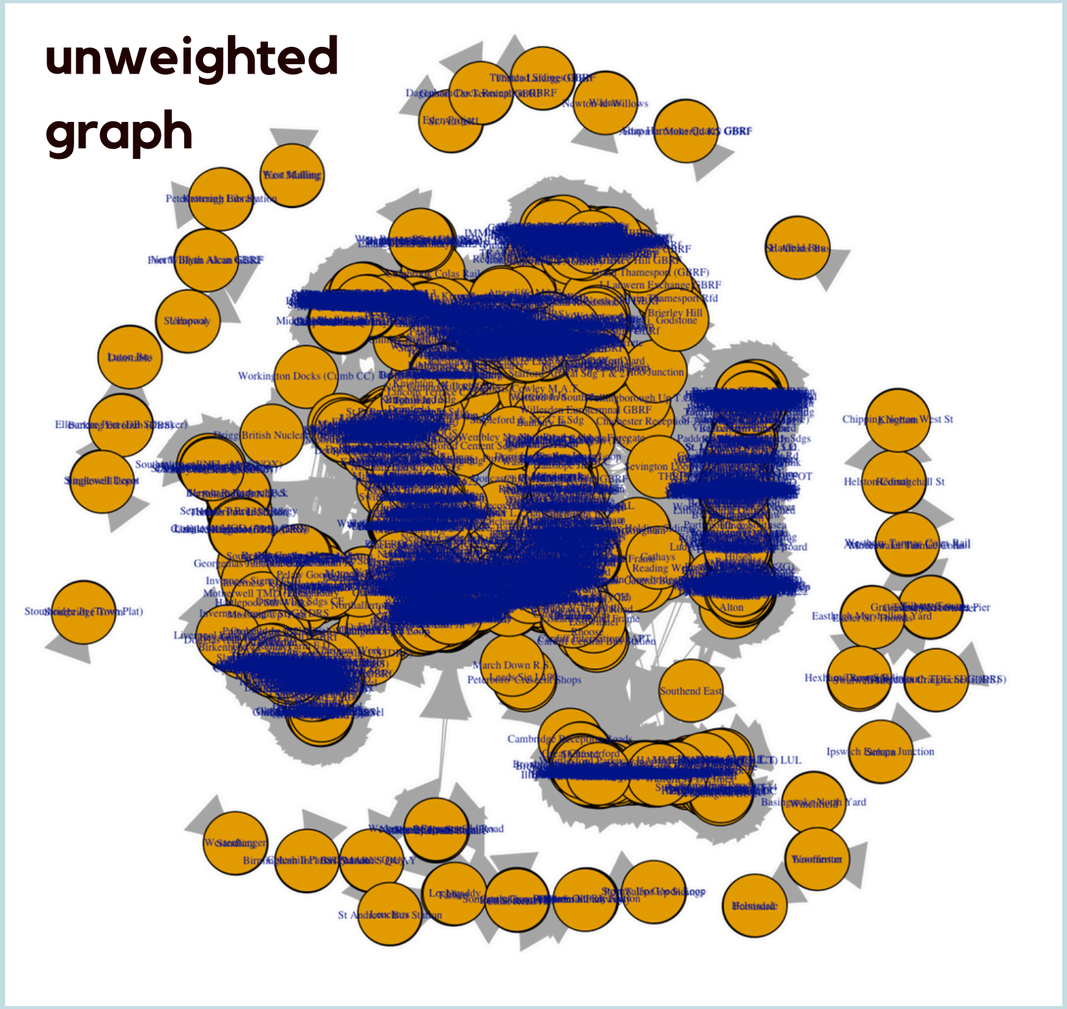
	Source	Target	n
1	Edinburgh	Craigentinny T.& R.S.M.D	5
2	Darlington	York	4
3	Chester	Holyhead	3
4	Crewe	Manchester Piccadilly	3
5	Birmingham New Street	Crewe	3
6	Ilford EMUD	London Liverpool Street	3

⚠️ *Most of routes are covered only by one company*

Graph transformation

- *Better visualization of the network -> graph -> pairs G = (V, E) consisting of sets of vertices and edges*
- Weights -> number of companies operating on each route
- Unconnected graph -> some vertices isolated from the others





Network size

Diameter-> represents the longest distance, i.e. the longest shortest path

```
diameter(g_w)

[1] 18
```

Stations that we pass through by moving along the diameter:

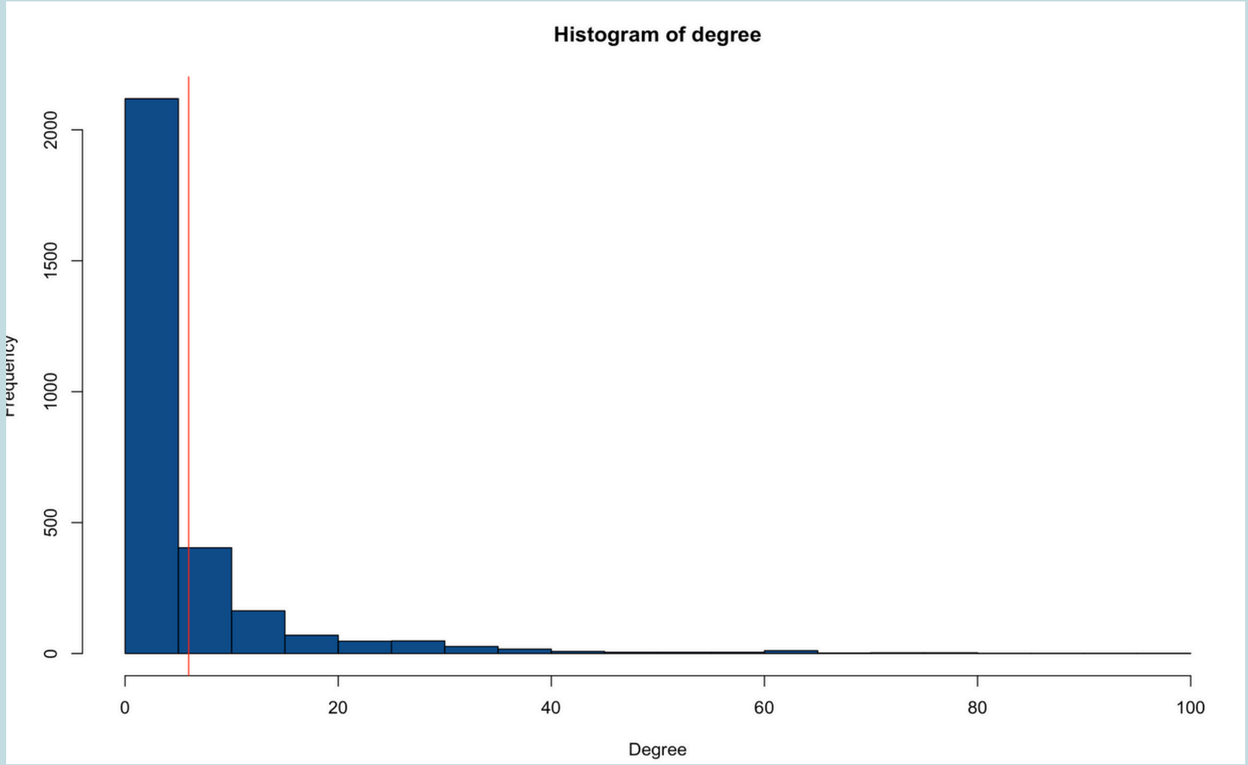
```
+ 19/2941 vertices, named, from 3df131b:
 [1] Kilwinning Up Siding       Lochwinnoch                 Carstairs
C.E.
 [4] Beattock                   Slateford Carriage Sidings Grantshouse
C.E.
 [7] Carlisle High Wapping SDGS Lancaster                   Leeds
[10] London Kings Cross         Cambridge North             Stratford
[13] Willesden Jn Low Level     Kilburn High Road           Stonebridge
Park
[16] Stonebridge Park Depot     Elephant & Castle L.T.      Harrow &
Wealdstone DC
[19] Queens Park A.C.
```

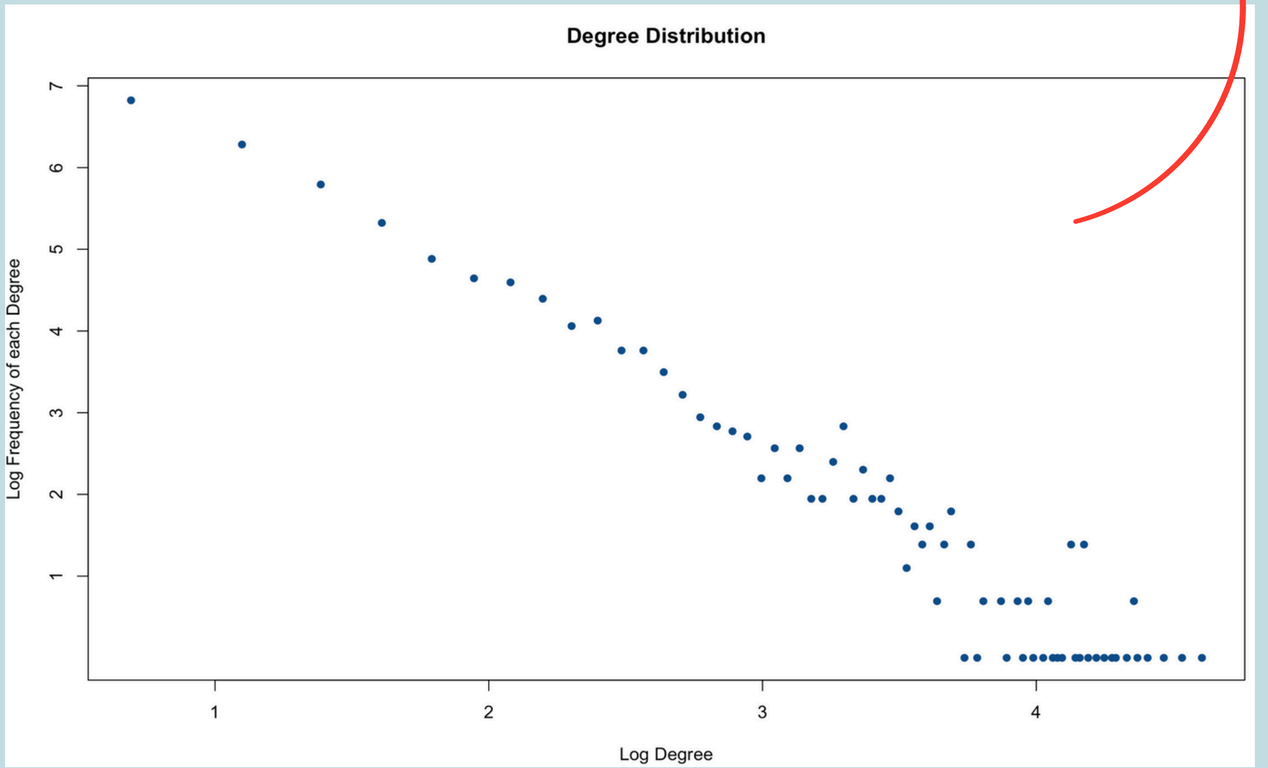
Degree

Power law pattern ->
preferential-attachment (PA) model

Number of edges connected to a vertex.

Examining the degree of each vertex ,we can derive the degree distribution of the network.





More than half of the vertices have a degree of 3 or less; only a few vertices exhibit significantly higher degrees; this suggests a heterogeneous degree distribution characterized by a heavy-tail

Quite linear pattern; for lower degrees, there is minimal noise because most observations fall within 1-3 range;
for higher degree values, there is much more variability due to the significantly fewer observations

Network sampling

Simulation of a series of subgraphs from the full graph to observe how the average degree varies and to gain insight into its variability.

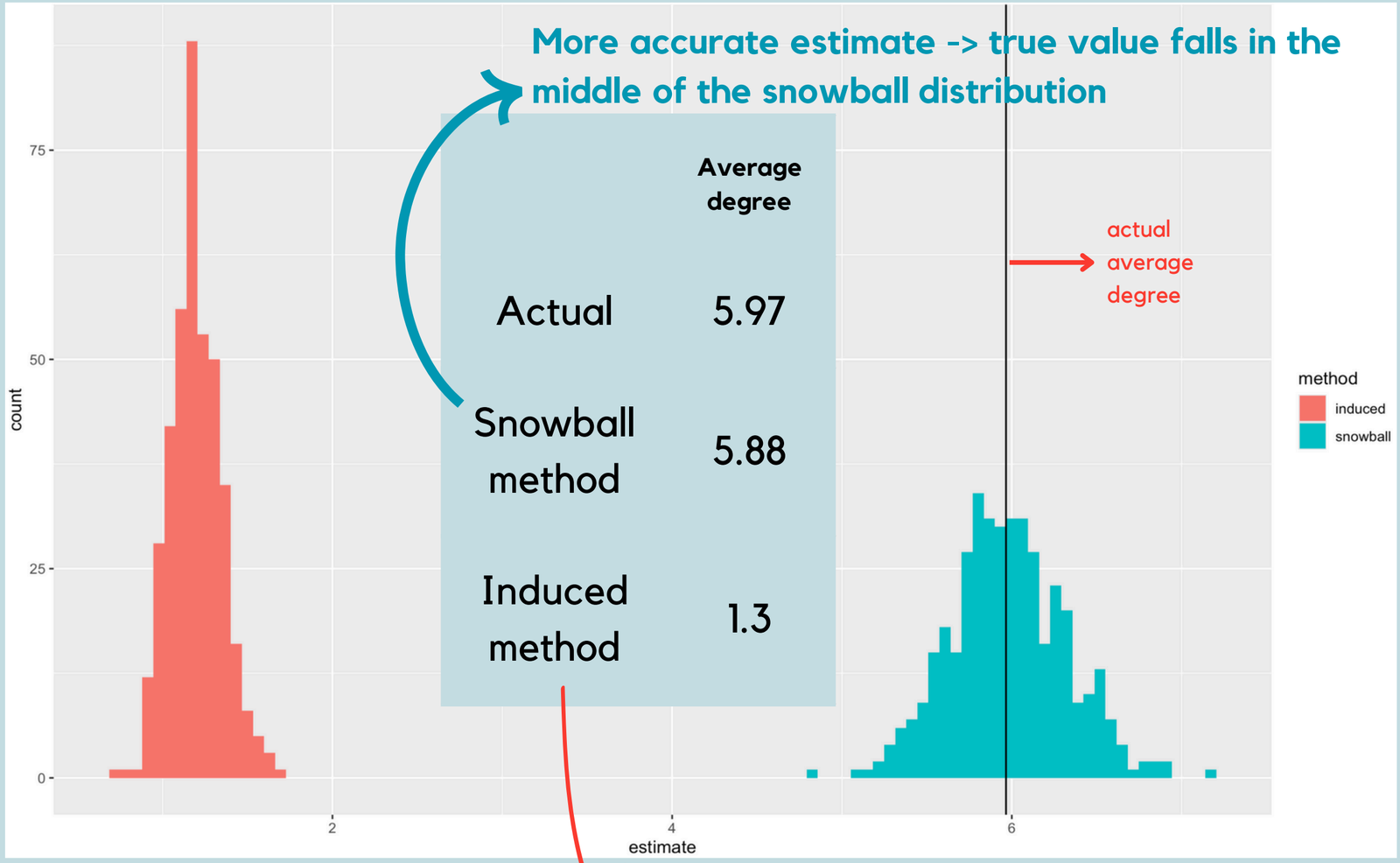
Two different sampling schemes:

- **Snowball**: sample all edges incident to each i ∈ V*
- **Induced**: sample all edges {i, j} such that i, j ∈ V*, i.e. sample G[V**] the subgraph induced from V**

Under each sampling design, we estimate the average degree with:

$$\eta(G^*) = \frac{1}{n} \sum_{i \in V^*} d_i^*$$

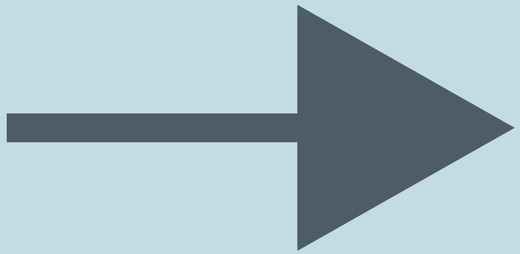
Repeating for 400 times...



Transitivity

- *Measure of the degree to which vertices in a graph tend to cluster together.*
- Defined as the ratio between the number of triangles to the number of triplets (connected triples of vertices) in the network.
- High transitivity -> network contains communities with densely connected nodes.

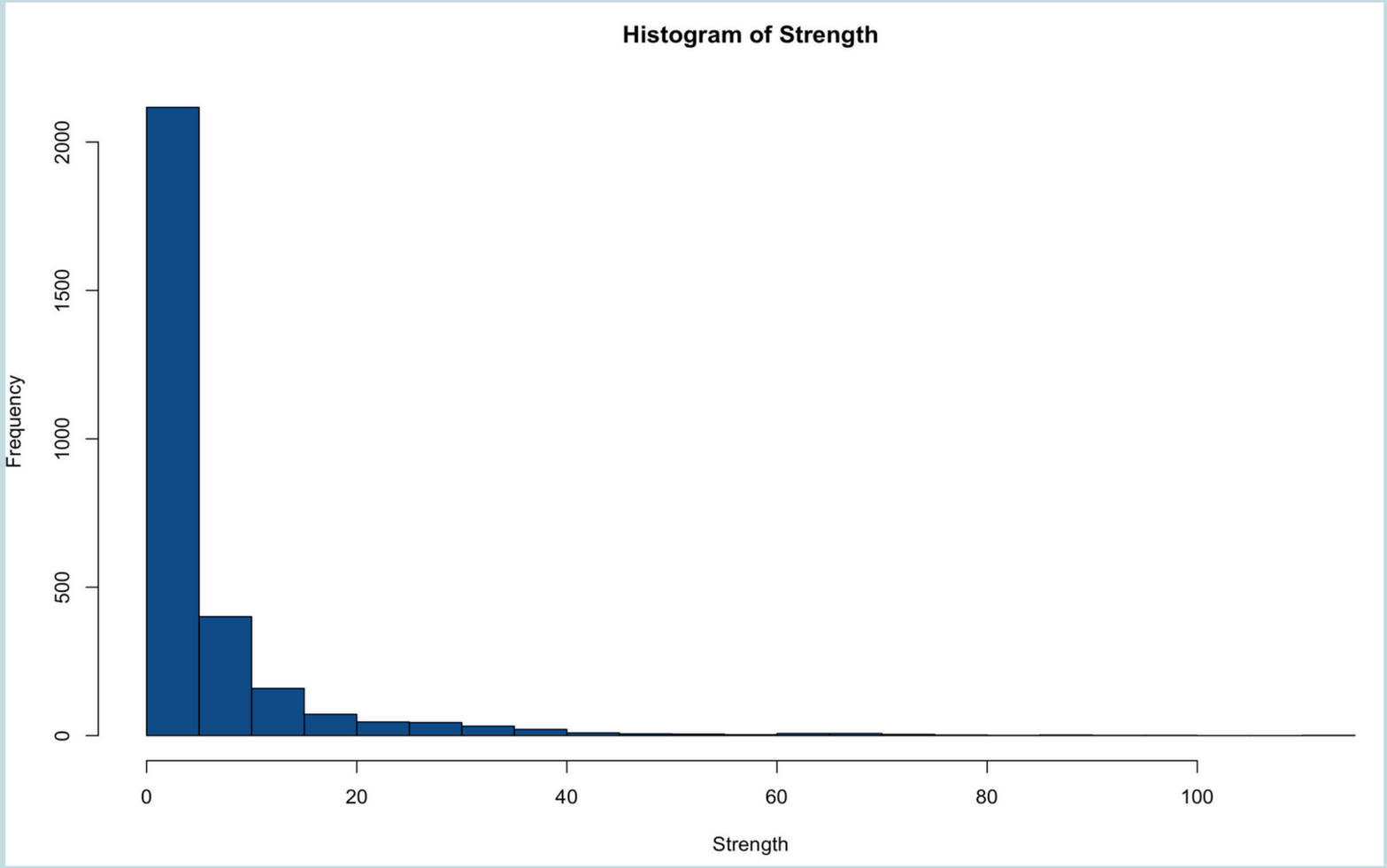
```
transitivity(g_w)

[1] 0.1723685
```



Low clustering coefficient, indicating weaker node connections.

Strength

Sum of the edge weights of the adjacent edges for each vertex.



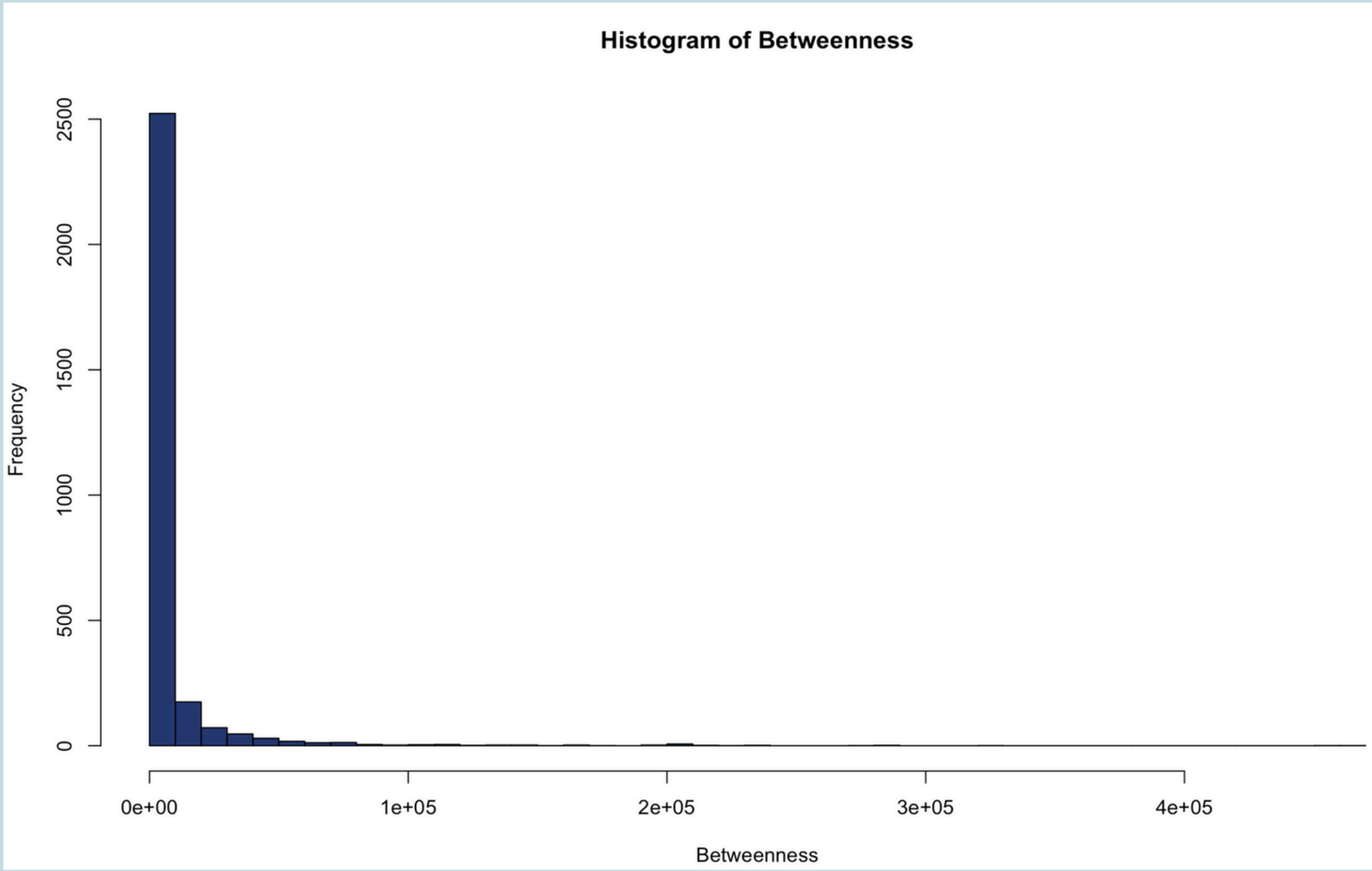

```
s[which.max(s)]

Edinburgh
       115
```

Edinburgh station has the highest strength at 115, the highest degree, and is served by multiple companies.

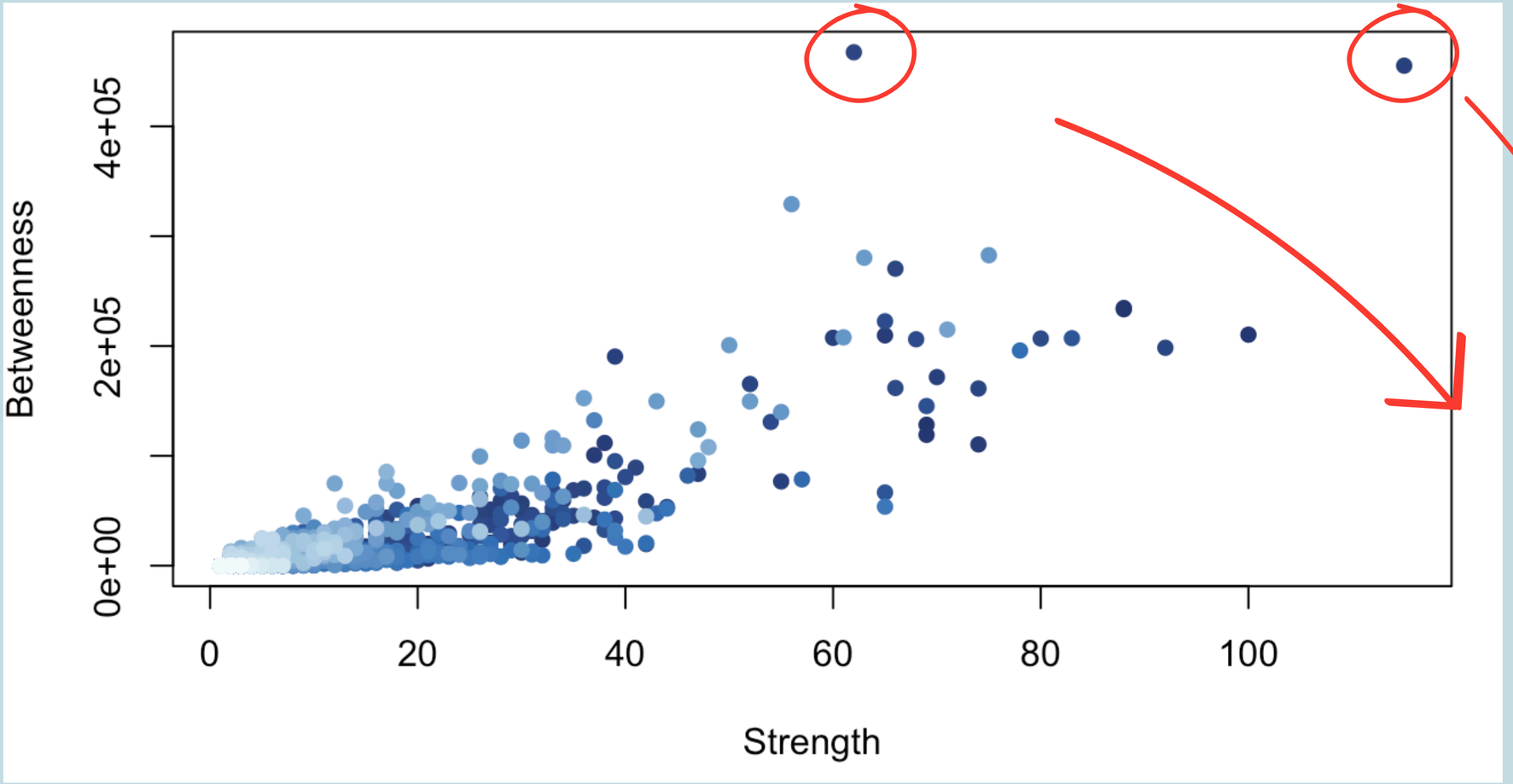
Betweenness

Measures how central or influential a vertex is based on the concept of shortest paths.

Vertices with higher betweenness centrality are more relevant, lying on more shortest paths between other vertex pairs.



Important actors



The two most important actors in the network are **WEMBLEY EUR FRT OPS CNTRE** and **EDINBURGH**

Fairly linear relationship between the two measures;
for strength values below 30, nodes tend to exhibit relatively low betweenness.

Linear model and GLM

Linear model: mod0 <- lm(log(fd) ~ log(d), data=dd)

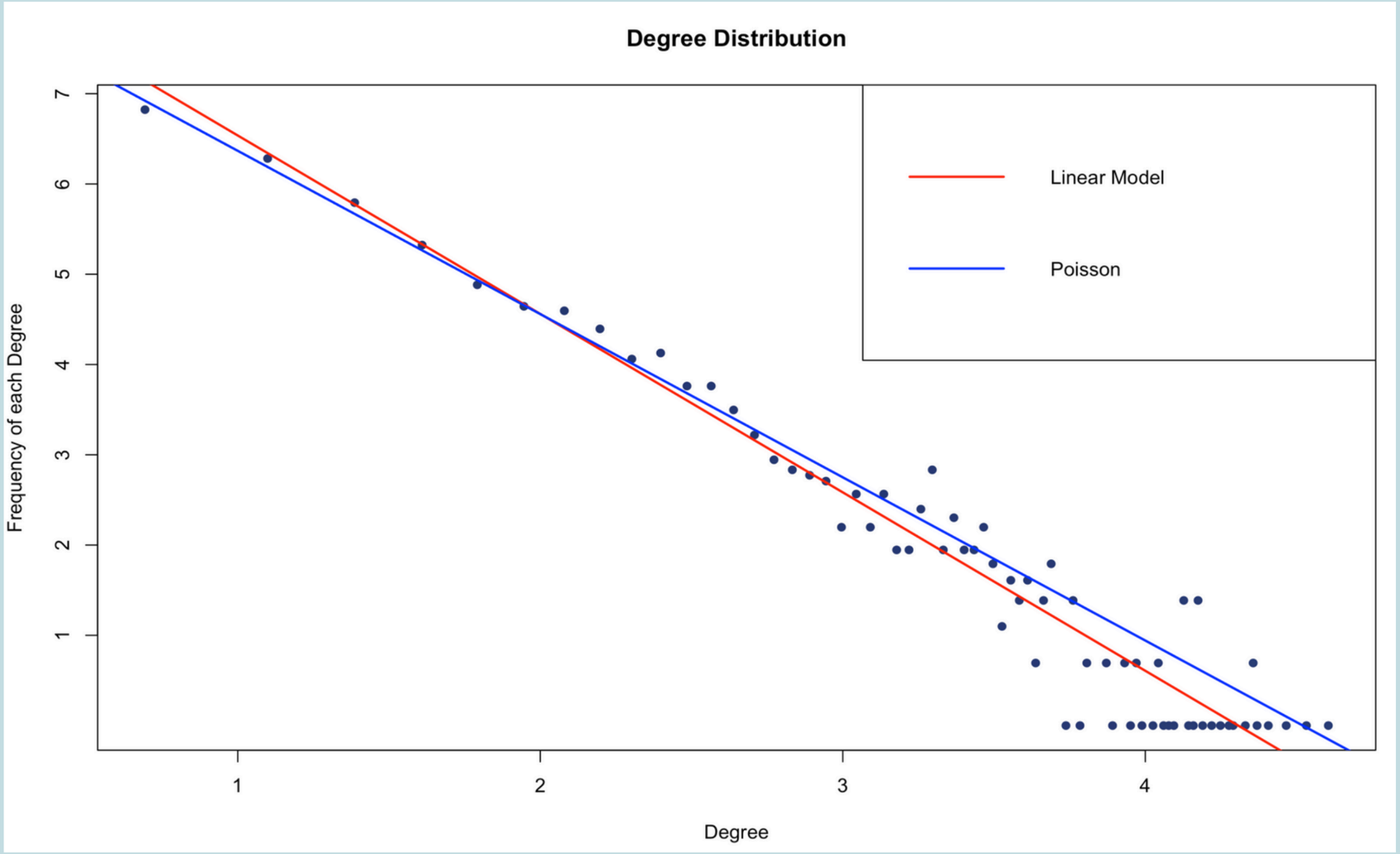
```
Residuals:
    Min      1Q  Median      3Q     Max
-1.1255 -0.2257  0.0167  0.2149  1.1240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.51392    0.20619   41.29   <2e-16 ***
log(d)      -1.97674    0.05872  -33.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4257 on 68 degrees of freedom
Multiple R-squared:  0.9434,    Adjusted R-squared:  0.9426
F-statistic:  1133 on 1 and 68 DF,  p-value: < 2.2e-16
```

GLM: mod1 <- glm(fd ~ log(d), family = poisson, data=dd)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.17672    0.03656   223.7   <2e-16 ***
log(d)      -1.80861    0.02110   -85.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10275.347  on 69  degrees of freedom
Residual deviance:    79.737  on 68  degrees of freedom
AIC: 342.33

Number of Fisher Scoring iterations: 4
```

Linear model and GLM

Both models perform well for smaller log-degree values but for higher degree values less able to properly capture the data structure due to increased variability.



ERGM

mod2 <- ergm(g_ergm ~ edges)

Exponential Random Graph Model (ERGM) analyzes the patterns and the processes within network data, using maximum pseudolikelihood estimation (MPLE) and estimating the impact of the edge predictor on the network structure.

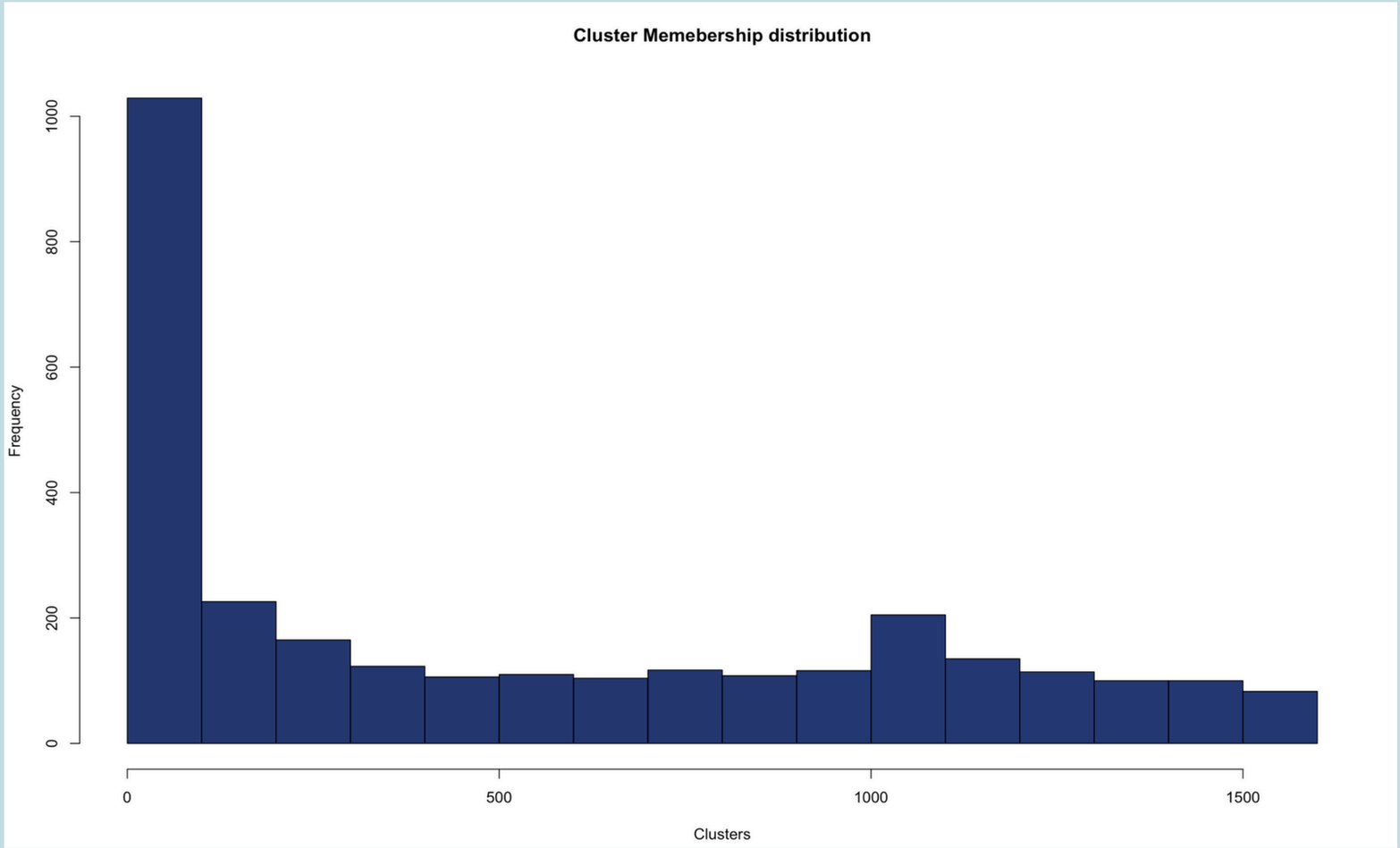
```
Maximum Likelihood Results:

      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -6.89222    0.01068      0  -645.3   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 11986650  on 8646540  degrees of freedom
 Residual Deviance:   138486  on 8646539  degrees of freedom

AIC: 138488  BIC: 138502  (Smaller is better. MC Std. Err. = 0)
```

The negative coefficient suggests that the formation of edges is less likely, indicating a sparse network structure, which is a type of graph where the number of edges is much lower compared to the possible maximum number of edges.

Clustering

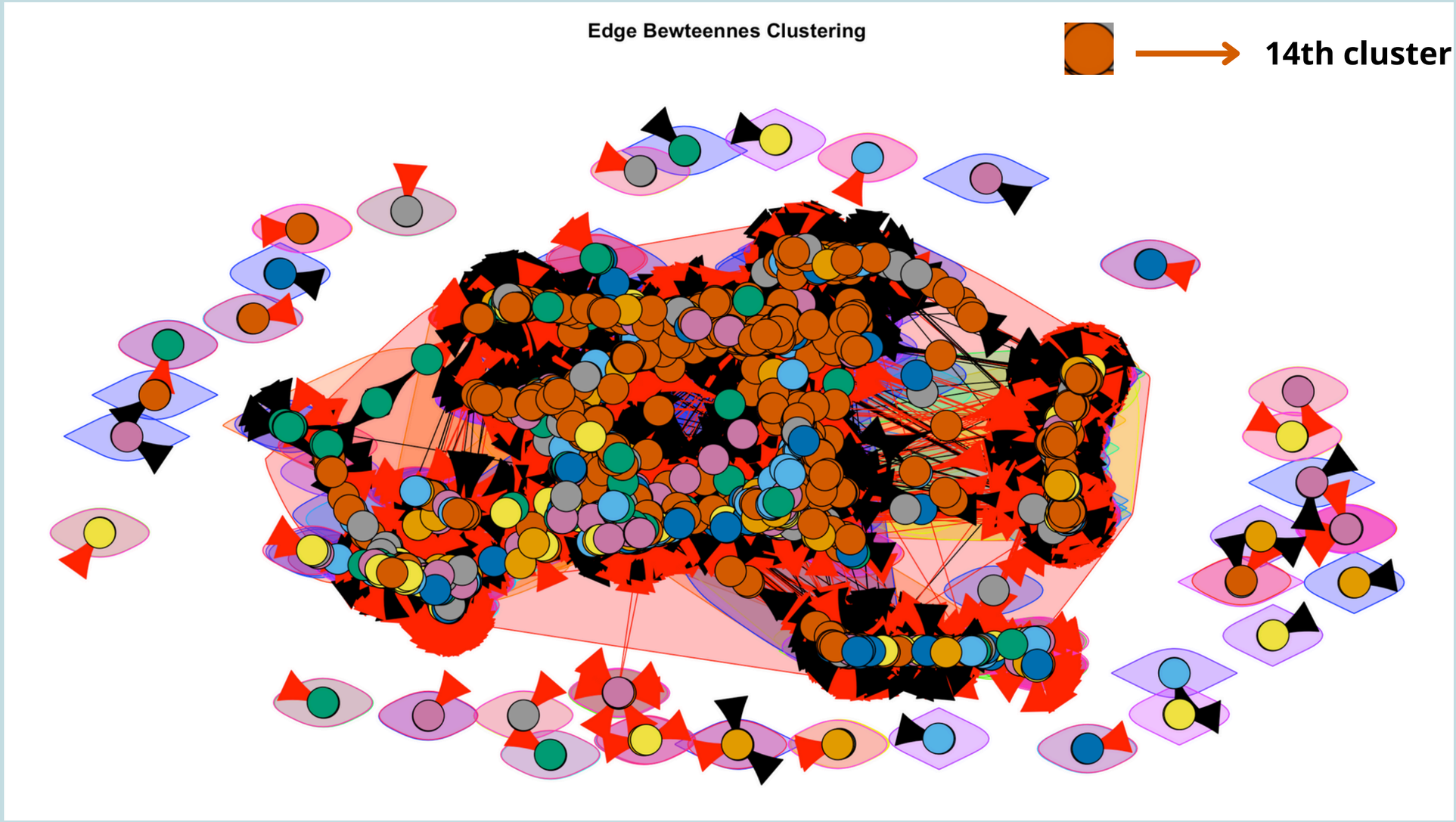
Directed graph -> **EDGE BETWENNESS ALGORITHM:**

- 1583 clusters
- Modularity score = 0.159 -> weak community structure -> network does not have well-defined clusters.

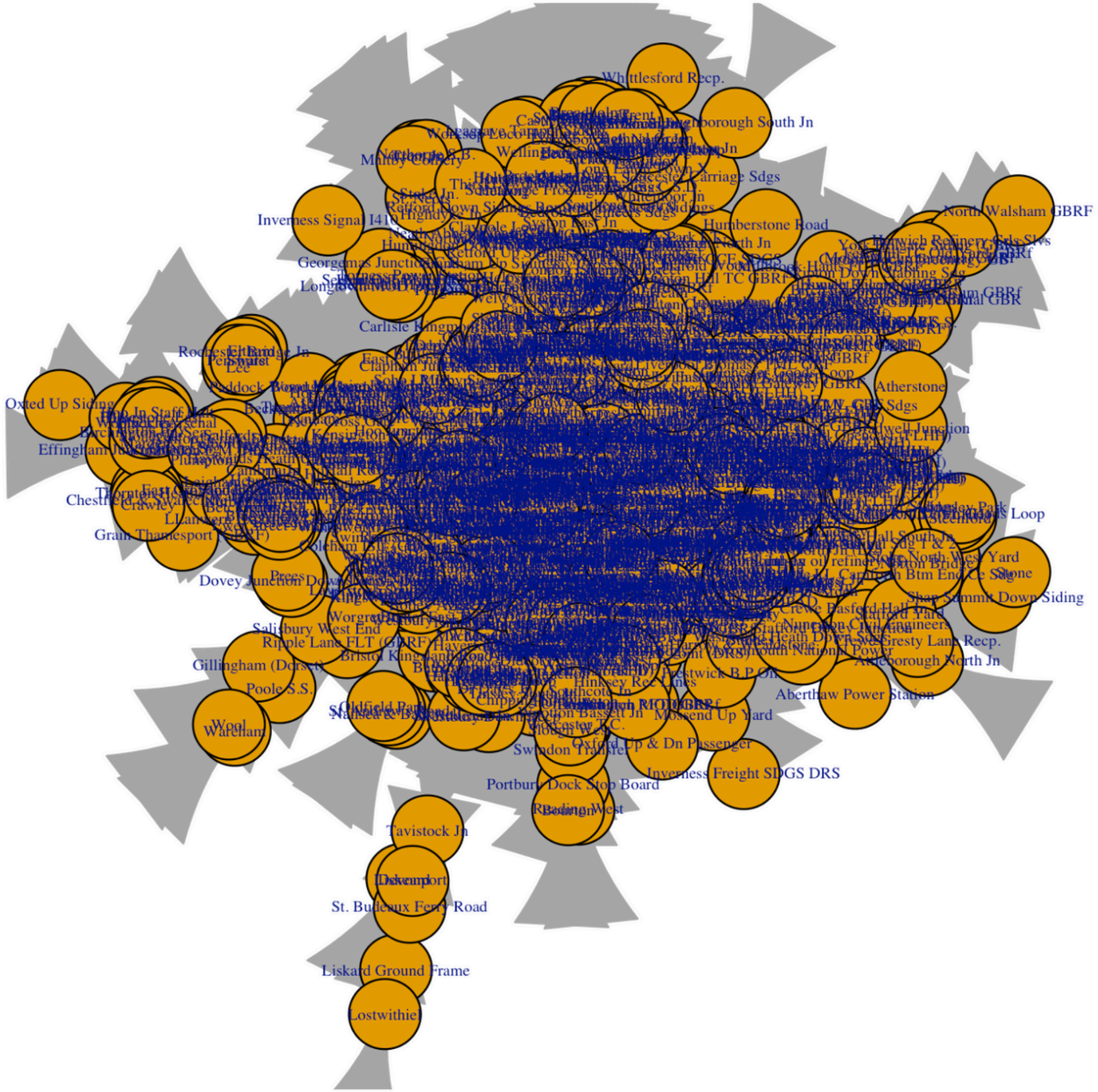


Substantial concentration of nodes in the first clusters, while the others seem to contain just a few nodes.

Sorting the clusters by size in ascending order, the most densely populated cluster is the 14^th, which appears in the first break of the histogram.



Dense and intricate network where nodes are highly interconnected, with significant overlap between different clusters and weak separability between clusters. The presence of many small clusters suggests that the network consists of numerous small communities rather than a few large, well-defined ones.

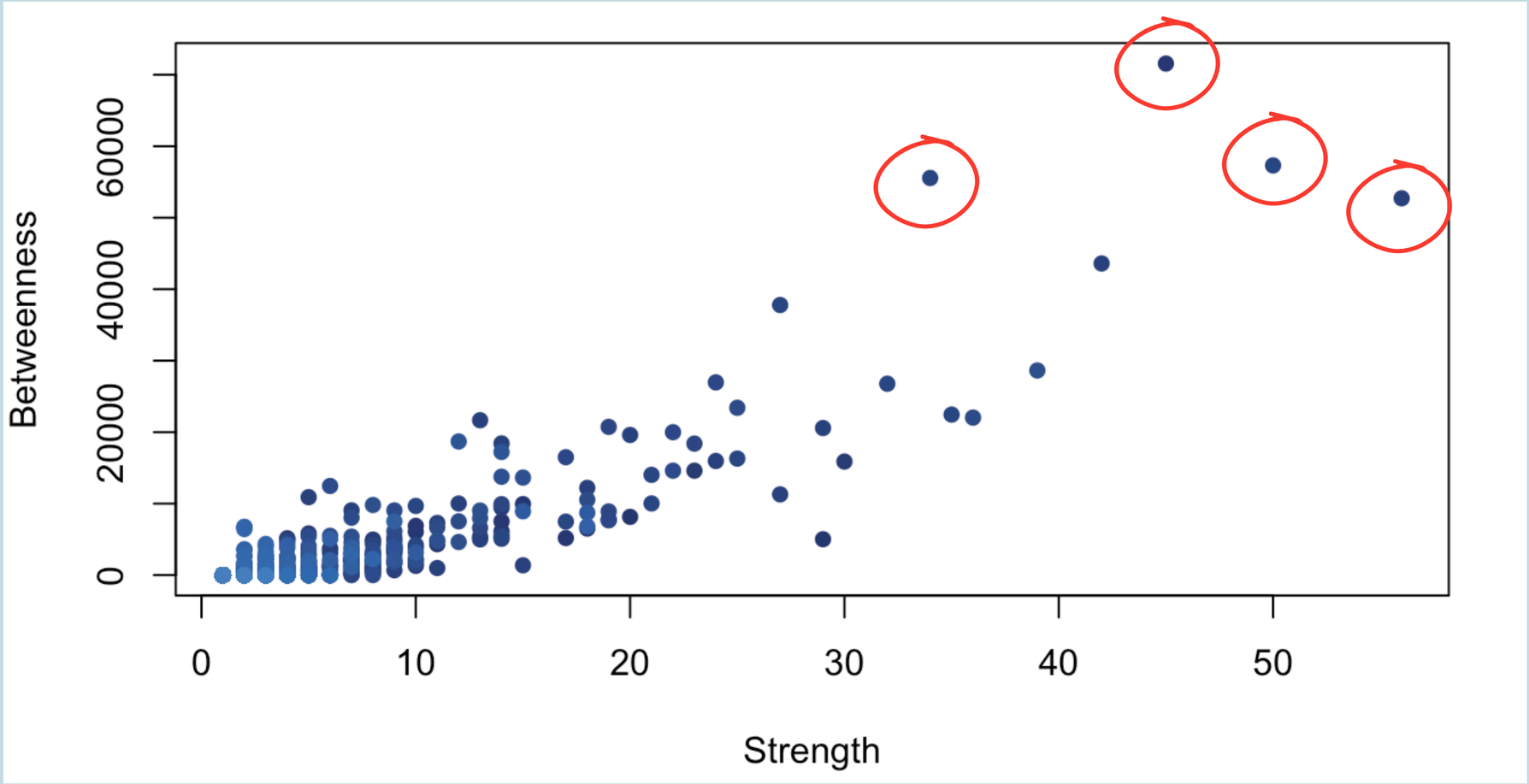


Subgraphing

Let's focus on the 14th cluster since it contains nearly a third of all the network's stations (around 900 nodes), making it a significant and representative subgraph for detailed analysis.

Subgraphing

The most important stations in the subgraph are: **EBBW VALE TOWN** , **CLEETHORPES** , **KETTERING** and **AMERSHAM**



The subgraph appears to share similar properties with the full graph, with most nodes exhibiting small degrees, low strength, and low betweenness values. However, the most important stations within the subgraph differ from those identified in the complete network.

Our Team

Lorenzo
Saracino

Biagio
Buono

Thank you for your attention!