Visualizing and analyzing the “Fernverkehrs-Ausbauschritt 2035 (STEP 2035[1/2])” of the Swiss federal railway as a complex network

Name: Florian Ackermann

Course: Data Science & Complex Networks

Semester: Fall 2023

Introduction

At the early age of artificial intelligence where free-to-use AI systems such as ChatGPT become more accessible, complex networks and distributed autonomous systems are rising in public interest. Network theory started with Euler’s problem of the “Bridges of Königsberg”. It became subject to mathematical graph theory and in Sociology, social networks have been studied for a long time (Mitchell 2006, pp. 1196). Nowadays, there are many scientific disciplines in which network related problems arise and the science of networks is a frequent topic in journals of physics, mathematics, computer science, biology, economics, and sociology (Watts 2004, pp. 243). A network is defined as a collection of nodes and links between nodes, where the links can be directed or undirected, and weighted or unweighted. Often-mentioned examples of complex systems in nature and society include the brain, the immune system, biological cells, metabolic networks, ant colonies, the Internet, economic markets, and human social connections (Mitchell 2006, pp. 1195). Some types of networks include Small-World networks (Watts & Strogatz 1998), Scale-Free Networks (Barabasi & Albert 1999), and Random networks. Each of these differ in the way in which networks are created, and by several statistics, such as degree distribution, average path length, and degree of clustering (Mitchell 2006, pp. 1197). To create complex networks in a visual form can be a powerful tool to situate and visualize emergent collective behavior (Mitchell 2006, pp. 1195), but it can sometimes be challenging to design a complex network. Data processing and practicing both visualizing methods and techniques are innate to mapping out a network. Here, I focus on a recently published plan of the Swiss federal railway SBB for re-organizing their passenger express rail system by 2035[1/2]. Evaluating passenger railway systems is essential for detecting the flow and distribution of passenger travelling, and for detecting points of heavy and sparse railway loadings. Recent research emphasized the complexity of railway systems, and classified edges based on their importance to the railway network, e.g. the failure of a “rich-rich” connection would damage the system more severely than a “poor-poor” connection (Xu et al. 2020). The purpose of this project is to recreate the network and to classify its position and relevance for network theory. The former is carried out via Python-coding using NetworkX (Version 3.2.1)[3] and NumPy (Version 1.25.2)[4] libraries and via Gephi software. Whereas Python is frequently used for programming and data processing, it can also be used for creating visual networks. However, Gephi is a more in-depth graph and network visualization technique. The latter will be done by discussing the network based on network theory, going into detail concerning the degree distribution, the betweenness and closeness centrality, and the degree of clustering as well as the different types of networks. While analyzing the Swiss express rail system network (STEP 2035[1/2]), and discussing network theory, advantages and hindrances of different graph visualization methods will be highlighted.

Method

The data used here stem from the Swiss federal railway SBB[1/2] and are free for access. The STEP 2035 is a collection of express rail lines consisting of the train stations as nodes and railway routes as edges between the train stations. Table 1 & 2 show the STEP 2035 passenger railway plan. The edges are flagged in different colors corresponding to railway routes. Figure 1 provides a picture of the network including all railway lines halts as nodes, whereas in figure 2, only those train stations are displayed that are either a cross-point of two or more routes or a terminal station to one or more routes, only involving Intercity lines.

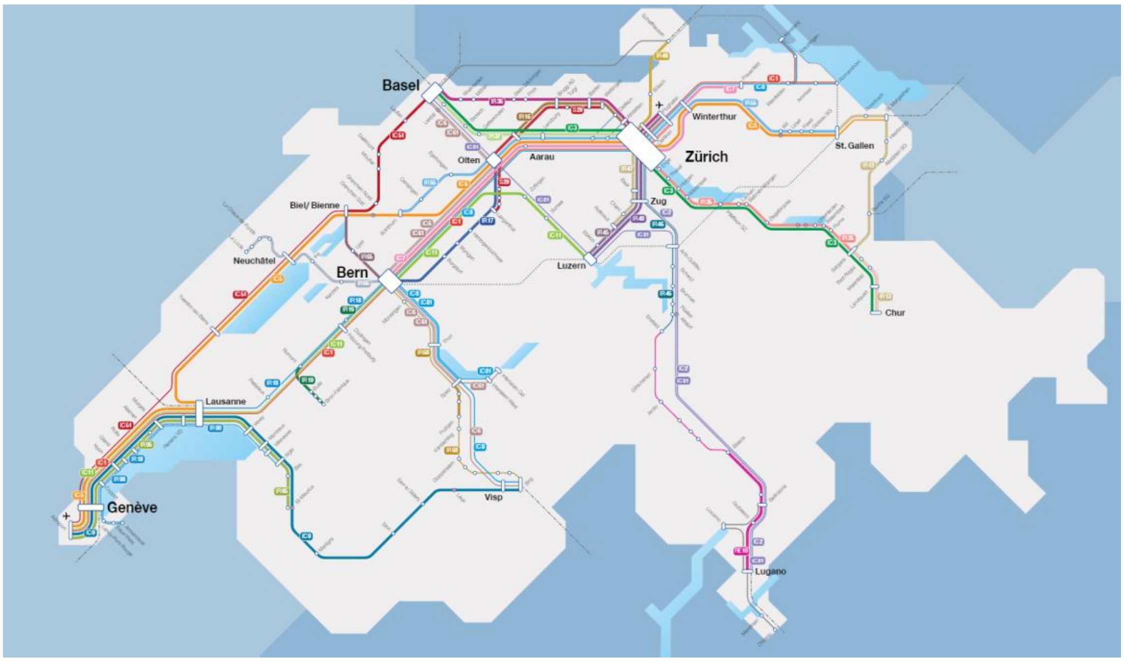

Figure 1: Detailed STEP 2035 express railway routes[1]

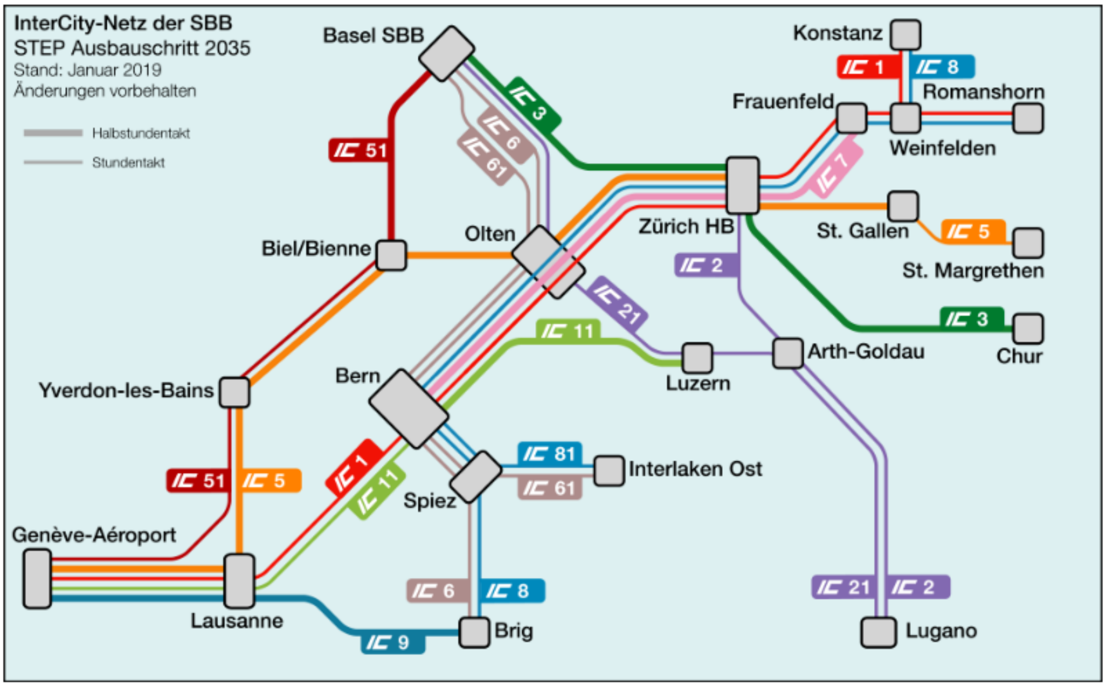

Figure 2: Simplified image of STEP 2035, excluded nodes that are non-Intercity halts, non-cross-points nor terminal stations[2]

The data analysis procedure required to define the nodes of the system and the edges. Because the detailed version of STEP 2035 (Figure 1) would be too time-consuming to recreate, I focused on the simplified version (Figure 2). The NetworkX package allowed me to add nodes and edges to a graph. Here, the edges were undirected and weighted with the weight representing the total number of express passenger rail connections between two nodes in each direction per hour. This resulted in an adjacency matrix of 21 nodes and 28 edges (Table 1).

```
0;0;0;0;0;2;3;0;0;0;0;2;0;0;0;0;0;0;0;0;0
0;0;6;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
0;6;0;2;2;0;0;2;0;0;0;0;0;0;0;0;0;0;0;0;0
0;0;2;0;0;0;0;0;2;0;0;0;0;0;0;0;0;0;0;0;0
0;1;2;0;0;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
2;0;0;0;3;0;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0
3;0;0;0;0;2;0;2;0;0;1;2;0;0;0;0;0;0;0;0;0
0;0;2;0;0;0;2;0;4;0;2;4;0;0;0;0;0;0;0;0;0
0;0;0;2;0;0;0;4;0;2;0;0;0;0;0;0;0;0;0;0;0
0;0;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0;0;0;0;0
0;0;0;0;0;0;1;2;0;0;0;0;1;0;0;0;0;0;0;0;0
2;0;0;0;0;0;2;4;0;0;0;0;1;0;2;2;0;4;0;0;0
0;0;0;0;0;0;0;0;0;0;1;1;0;2;0;0;0;0;0;0;0
0;0;0;0;0;0;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0
0;0;0;0;0;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0;0
0;0;0;0;0;0;0;0;0;0;0;2;0;0;0;0;1;0;0;0;0
0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0
0;0;0;0;0;0;0;0;0;0;0;4;0;0;0;0;0;0;2;0;0
0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;0;1;1
0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0
0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0
```

Table 1: Adjacency matrix of the simplified STEP 2035 network

In a next step, the data have been mapped out and tested in their responsiveness to statistical measures, such as degree distribution, graph density, cluster coefficient, betweenness centrality and closeness centrality. The degree distribution provides information about the prevalence of nodes n with different degrees k in the network. In other words, it is the probability distribution that describes the likelihood of a node having a certain number of edges. Preferential attachment leads to a power law degree distribution, where a minority of nodes have a very high degree distribution and the majority has low degree distribution (Chung 2010, pp. 727).

$$P(k) = \frac{n_k}{N} \qquad 1 = \sum_k P(k)$$

The graph density D is a measure that quantifies the proportion of edges E in a graph relative to the total number of possible edges. It provides an indication of how well-connected the nodes in the graph are. The density of a graph is a value between 0 and 1. A graph has a value 1 when all nodes are connected to each other.

$$D = \frac{2(E)}{N(N-1)}$$

In graph theory, the cluster coefficient is an important tool to measure the degree to which the nodes tend to form clusters. It is based on a local cluster coefficient for each node, which is the ratio of the number of links among the nodes to the maximum number of links in the neighborhood of a node. The local cluster coefficient C_i calculation involves the number of triangles connected to a node i, and the edge k_i of node i. The global clustering coefficient is the average of the local clustering coefficients for all nodes in the graph.

$$C_i = \frac{2T_i}{k_i(k_i - 1)}$$

Some centrality measurements are the betweenness centrality and closeness centrality. Centrality measurements rely on the shortest path. The normalized betweenness centrality $B_{norm(v)}$ quantifies the number of times a node v acts as a bridge along the shortest path between two other nodes i and j. In our case, a node with high betweenness centrality has a significant influence on the flow of passenger rail operations through the network.

$$B_{norm(v)} = \frac{\sum \sigma_{i,j}(v)}{(\mathrm{N}-1)(\mathrm{N}-2)/2}$$

The access of a node i to other nodes is quantified by the normalized closeness centrality $C_{norm(i)}$. It measures the average length of the shortest paths $d_{i,j}$ to all other nodes in the network from the given node.

$$C_{norm(i)} = \frac{N-1}{\sum_{j=1}^{N} d_{i,j}}$$

The STEP 2035 network has been evaluated based on these statistical measures using Python and Gephi. These measures offer a variety of possibilities to evaluate the network. Each measurement has been calculated as described by the respective formula. In addition, I carried out visualizations of the network as an assistance for understanding the spatial extent of node- and edge-distribution.

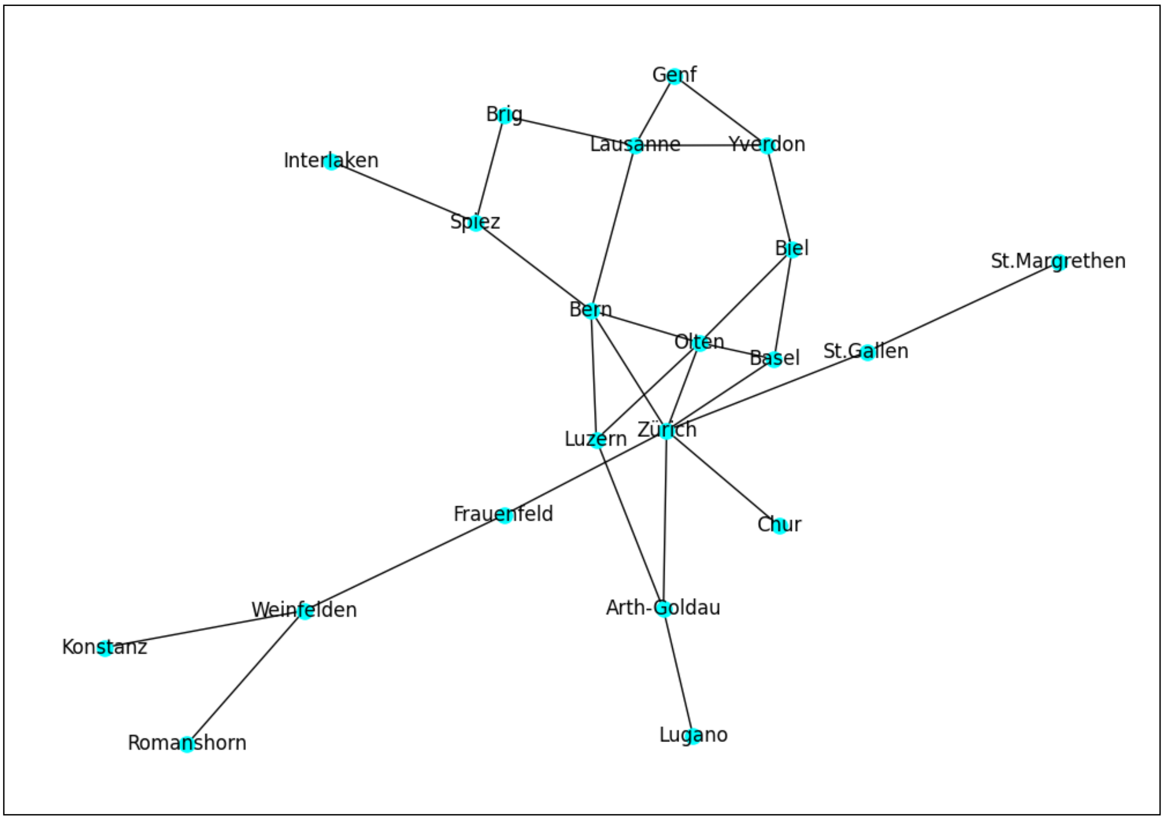

Figure 3: Network visualization with Network:X Unweighted edges, random layout

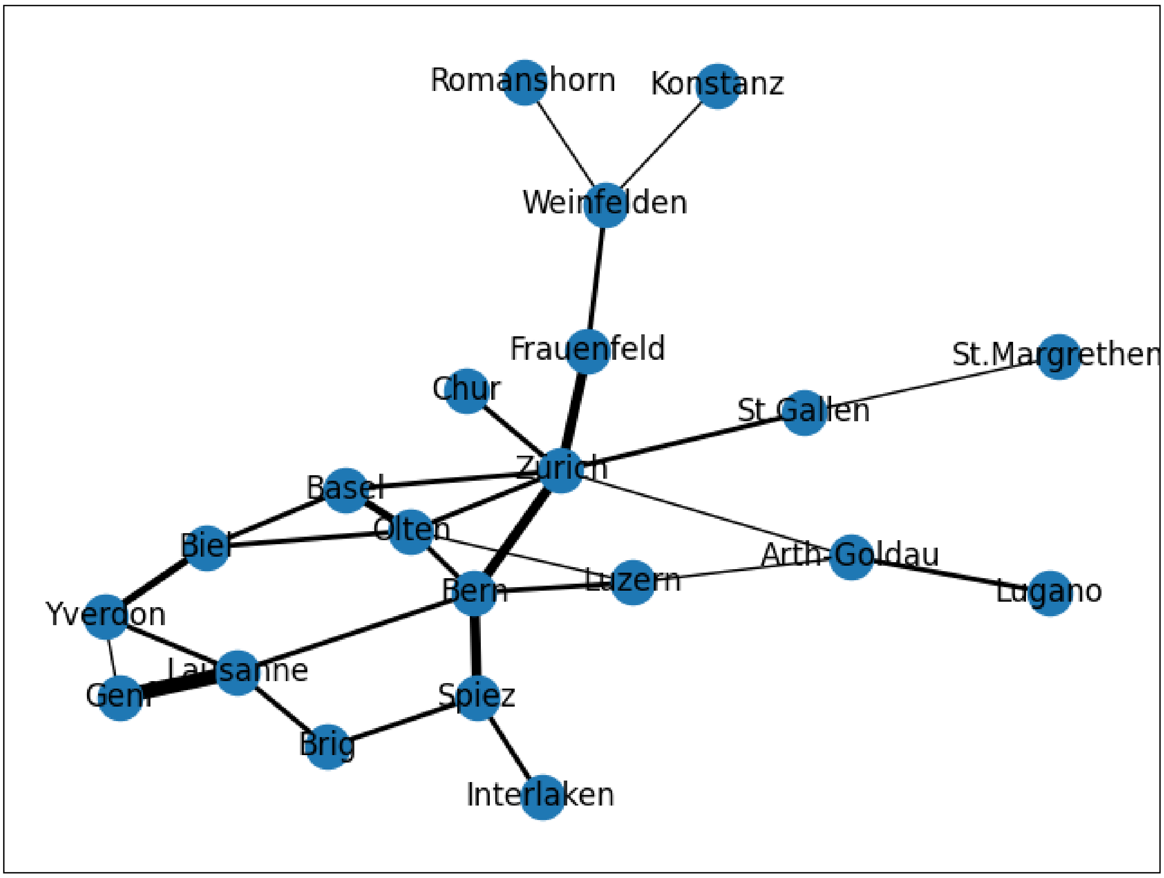

Figure 4: Network visualization with NetworkX: Weighted edges, random layout

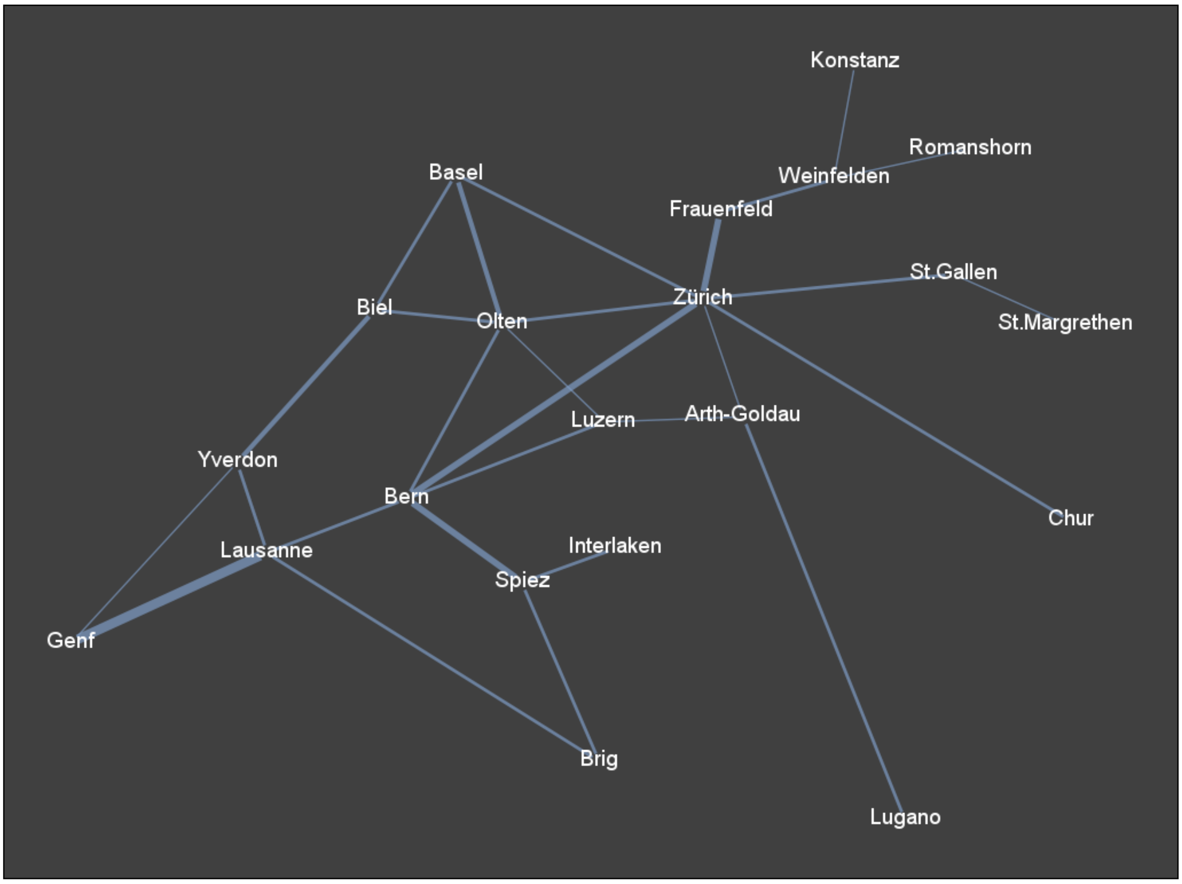

Figure 5: Network visualization with Gephi: Weighted edges, geographically more accurate layout

Average cluster Coefficient:	0.168
Graph density:	0.133

	Degree	Cluster Coefficient	Betweenness Centrality	Closeness Centrality
Zürich	7	0.095	0.625	0.513
Bern	5	0.200	0.392	0.476
Olten	5	0.400	0.101	0.444
Lausanne	4	0.166	0.166	0.370
Basel	3	0.667	0.040	0.392
Luzern	3	0.333	0.044	0.370
Arth-Goldau	3	0	0.114	0.370
Biel	3	0.333	0.058	0.351
Spiez	3	0	0.136	0.351
Yverdon	3	0.333	0.026	0.299
Weinfelden	3	0	0.194	0.299
Frauenfeld	2	0	0.268	0.385
St.Gallen	2	0	0.100	0.357
Brig	2	0	0.016	0.290
Genf	2	1	0	0.286
Chur	1	0	0	0.344
Lugano	1	0	0	0.274

St.Margrethen	1	0	0	0.267
Interlaken	1	0	0	0.263
Romanshorn	1	0	0	0.233
Konstanz	1	0	0	0.233

Table 2: Overview of each node's metrics

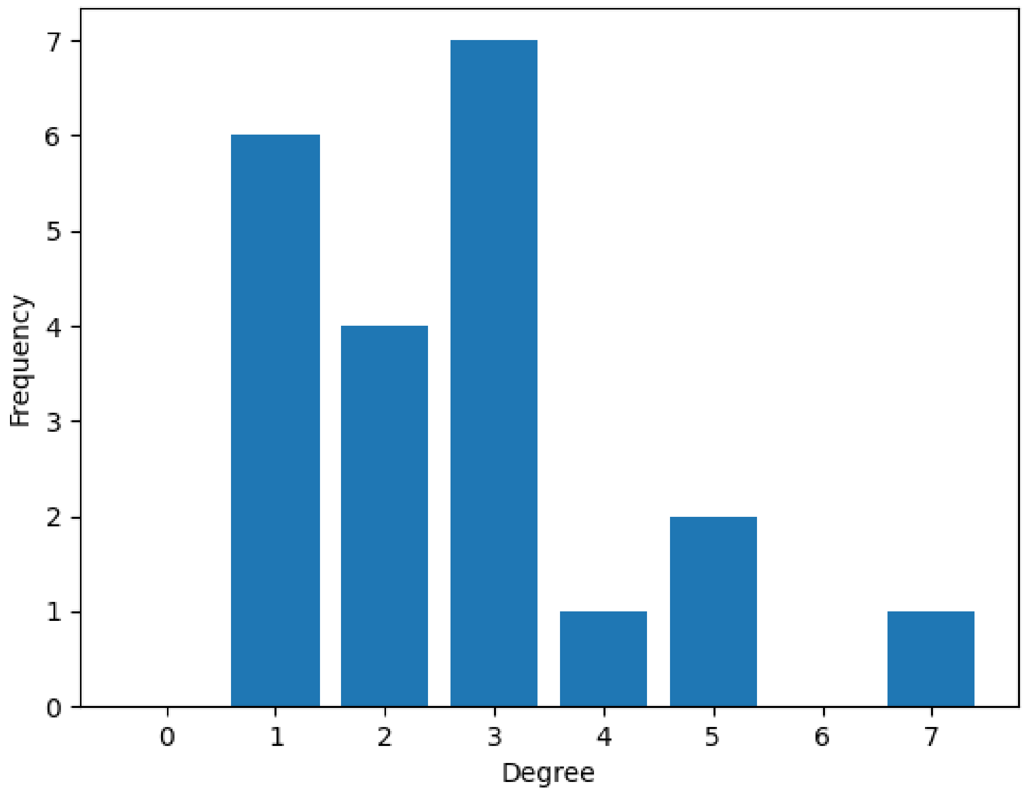

Figure 6: Degree distribution, via Python coding

Results

Zürich has the highest degree of all nodes in the network with total 7 different edges to surrounding nodes. Next are Bern and Olten with 5 and Lausanne with 4 connected edges in total. Most nodes have between 1 and 3 edges, resulting in a degree distribution graph that shows a tendency to tilt downwards with an increased number of edges for a node (Figure 6). Table 2 provides an overview of degree, cluster coefficient, betweenness centrality and closeness centrality for each node. The graph has an average cluster coefficient of 0.168, however it varies strongly dependent on the geographic location of the node. Genf is the only node that has a cluster coefficient of 1 and Basel is another node with a high cluster coefficient of 0.667. In general, nodes with a non-zero cluster coefficient are to be found in the western part of the Swiss railway system. Figure 5 shows a geographically arranged layout of nodes, made with Gephi. The uneven distribution of clusters is composed of clustered nodes in the West and poorly clustered nodes in the east. While many flows in the east seem to go through Zürich, the eastern nodes themselves are not well-connected. A high degree does not necessarily predict a high cluster coefficient. As expected from the uneven geographically distributed cluster coefficient and from the picture that Zürich is an access location for many eastern nodes, it has the highest betweenness centrality of 0.625. Almost all paths that end or start in the eastern nodes, such as Chur, St. Gallen, St. Margrethen, Romanshorn, Konstanz, Weinfelden and Frauenfeld lead through Zürich. Both centrality measures align, with some exceptions (Weinfelden, Frauenfeld, St. Gallen), linearly with the degree distribution, although the magnitude of closeness centrality is weaker compared to betweenness centrality. A graph density of 0.133 is moderately low, indicating that a graph has few connections and is not highly connected. However, having another graph of a different Express passenger system would be useful as an object of comparison since the graph density is hard to interpret here. For visualizations of the graph, there are three examples (Figure 3, 4, 5). Figure 3 and 4 have been created with Python, with figure 4 being composed of weighted edges. Figure 5 has weighted edges too and is laid out in a geographically more accurate manner, using Gephi.

Discussion

Python is a programming language which can image networks and graphs with additional packages such as NetworkX and NumPy. To calculate statistical measurements for a graph is also possible with Python, but it is important to have the necessary coding knowledge resource. On the other hand, Gephi provides an all-round package for network conceptualization, and directly provides the user with various graph metrics, such as clustering coefficient, centrality measures, degrees, eccentricity, and authority. It gives a graph image window with the possibility to rearrange the depicted network spatially by a hold-and-drag mechanism. Smaller networks can be rearranged like in this example. Yet bigger, or more complex geographic networks should be visualized using tools such as GIS or Python with additional packages. For this network, centrality measures do not account for realism. Because the weighted edges represent the frequency of express rail connections, and not the travelling time, we cannot assume that both betweenness centrality or closeness centrality are accurate at all. The reason for integrating them is to present a variety of statistical measures that can be applied to a graph as well as to exercise and demonstrate new Python coding concepts. Another observation is that the probability for a node to have a high cluster coefficient increases with decreasing edges attached to that node, assuming that the sum of edges is not 0 or 1. This is why a high cluster coefficient does not always correspond to a high degree. Zürich for example, which has the highest degree, betweenness centrality and closeness centrality, shows a lower-than-average cluster coefficient of 0.095. The average cluster coefficient of 0.168 is rather low, which makes the network have a commonality with random networks (Mitchell 2006, pp. 1197). The degree distribution of a random networks is approximately Gaussian (or Poisson), while the degree distribution of a Scale-free network follows the power-law. Scale-free networks therefore have a "heavy tail" of low-degree nodes and very few nodes with a high degree (Watts 2004, pp. 250). The simplified STEP 2035 follows the power-law but note: This is only true for the simplified STEP 2035. The original version would include many nodes with a degree of 2, giving the degree distribution a more Gaussian shape. This would agree with recent research, stating that the indegree and outdegree of the China Railway System heavily accumulate at degree 2 (Xu et al. 2020, pp. 1497). Future research should investigate the STEP 2035 and other railway systems in detail, which becomes essential for classifying weak-points in railway systems and determining the mechanics that give a railway system its specific shape (such as power-law).

Conclusion

The topology of express passenger railways follows the demands of passenger travel. Here, I assessed an imaging of the STEP 2035 Swiss railway project with weighted edges, and I therefore do not claim accuracy for closeness and betweenness centrality measures. In the STEP 2035, the edges are distributed heterogeneously. Edges with a heavy weight as well as nodes with high degree and centrality metrics amplify the magnitude of disturbances to the whole system, while light edges and eccentric nodes have a relatively low impact on the network flow. Zürich is the key node to access eastern Swiss cities and is the most impactful node. The STEP 2035 has similarities with scale-free networks and random networks, that makes it hard to classify into one category. Various software products can carry out illustrations and metrics analyses of graphs, such as Python or Gephi. While Gephi is a good choice for small networks, Python with additional packages appeals to large-scale complex networks.

Sources

[1] https://de.wikipedia.org/wiki/InterCity_(Schweiz)

[2] https://company.sbb.ch/de/ueber-die-sbb/projekte/nationale-programme/step-as-2035.html

[3]Aric Hagberg, Pieter J. Swart, Daniel A. Schult (2008): Exploring Network Structure, Dynamics, and Function using NetworkX. United States, DOI: https://www.osti.gov/servlets/purl/960616

[4]Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, Travis E. Oliphant (2020): Array Programming with NumPy. Nature, 585, pp. 357–362, DOI : https://doi.org/10.1038/s41586-020-2649-2

Literature

Albert-Laszlo Barabasi, Réka Albert (1999): Emergence of Scaling in Random Networks. Science, 286, pp. 509-512, DOI: 10.1126/science.286.5439.509

Fan Chung (2010): Graph Theory in the Information Age. Notices of the AMS, 57, pp. 726-732

Melanie Mitchell (2006): Complex Systems: Network Thinking. Artificial Intelligence, 170, pp. 1194-1212, DOI: https://doi.org/10.1016/j.artint.2006.10.002

Duncan J. Watts (2004): The “New” Science of Networks. Annual Review of Sociology, 30, pp. 243-270, DOI: https://doi.org/10.1146/annurev.soc.30.020404.104342

Duncan J. Watts, Steven H. Strogatz (1998): Collective Dynamics of ‘Small-World’ Networks. Nature, 393, pp. 440-442, DOI: https://doi.org/10.1038/30918

Zhongzhi Xu, Qingpeng Zhang, Dingjun Chen, Yuxin He (2019): Characterizing the Connectivity of Railway Networks. IEEE Transactions on Intelligent Transportation Systems, 21, pp. 1491-1502, doi: 10.1109/TITS.2019.2909120