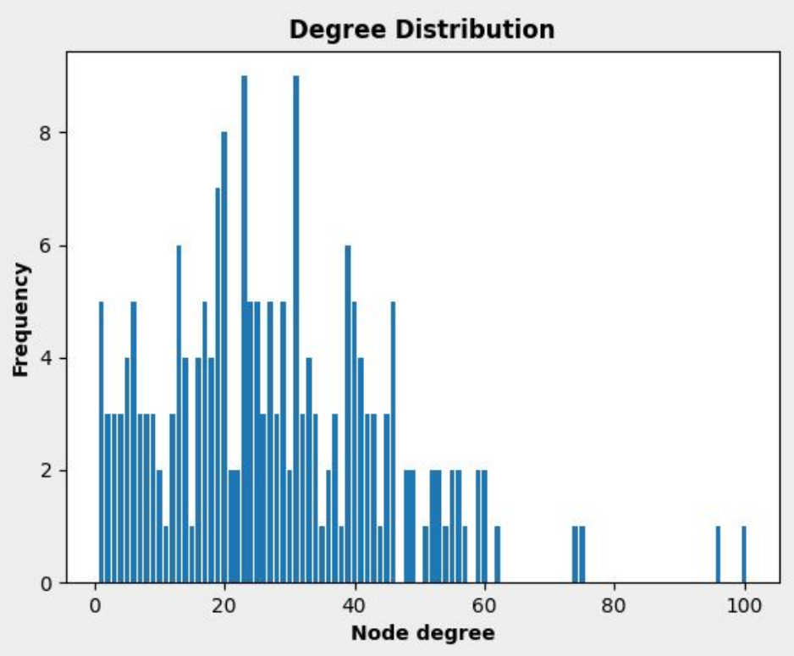
Similarity Indices for Link Prediction

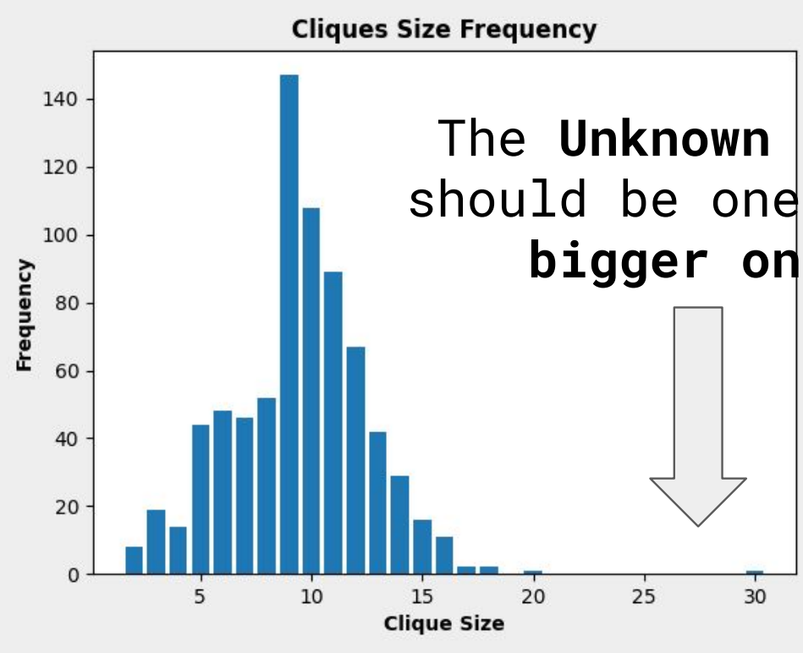
Group 3: Maarten van Sluijs; Roëlle Bänffer; Andrea Mangrella

Problem formalization

Dataset:
The Jazz Musician dataset consists of jazz bands as nodes. Bands sharing the same musician(s) are connected with an edge. **198 nodes, 2742 edges.**

Small world property:
the average distance between vertices is small, while the clustering vertices remains high. (degree distribution P(k) is skewed)





The **Unknown Clique** should be one of the **bigger ones.**

Unknown Musician:in the original database an unknown band member is transformed into the unknown member (that is always considered as the same person).

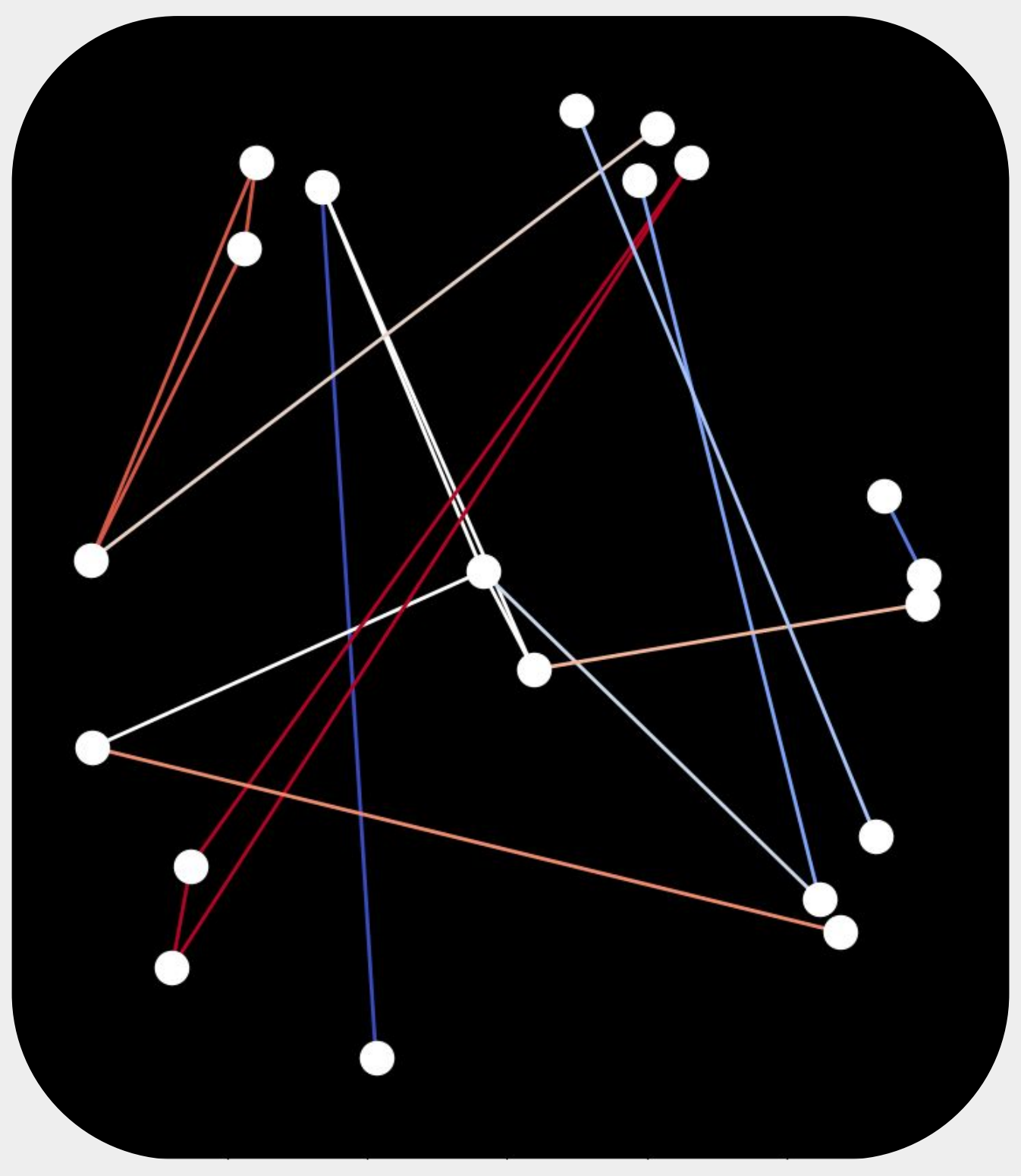
Two bands that contain an Unknown are connected, but the connection is spurious (**should not exist**).

Our Idea:

We suppose that **the Unknowns form together a clique**, and that the edges in this clique should be **detected as the most Spurious** (the ones with the lowest similarity score).

Visualization of some of the smaller Cliques of the network.

We **kept** the edges between two Cliques and colored in **white**.



Possible Solutions

Algorithm step to find the most Spurious links:

1. Find the biggest clique in the graph;
2. Remove all of the internal edges of the biggest clique from the **Training Set** and turn them into the **Probe Set**;
3. Using existing **(dis)**similarity indices to find the edges with the lowest values;
4. Compare these edges with the **Probe Set**;

Similarity-based Algorithm:

The simplest form of link prediction methods is the, where each pair of nodes, x and y, is assigned a score sxy, which is directly defined as the similarity between x and y. The higher the similarity, the more likely that the edge exists.

Leicht-Holme-Newman Index (already existing method):
The LHN_2 index in global form checks if two nodes are similar if either of them has a neighbor which is similar to the other node.

It's a **global index**:

$$S = D^{-1} * (I - \frac{\phi A}{\lambda_1})^{-1} * D^{-1}$$

- D = degree matrix
- A = adjacency matrix
- ϕ = free parameter
- λ_1 = maximum eigenvalue of matrix A
- I = identity matrix

Spectral Comparison (our method):
We want to exploit the **Community Structure** of the graph to get how likely two nodes are in the same community, and then use this value as a Similarity Index.
The index denominator is the **Euclidean Distance** between the two x and y nodes **Eigen Vectors** of the Standard **Graph Laplacian**.

first get $L = D - A$

and then solve $L * v = \lambda * D * v$

$$s^{SS}_{xy} = \frac{1}{\sqrt{\sum_{k=1}^{n}(v^x_k - v^y_k)^2}}$$

Ranked Solutions

Leicht-Holme-Newman Index2:

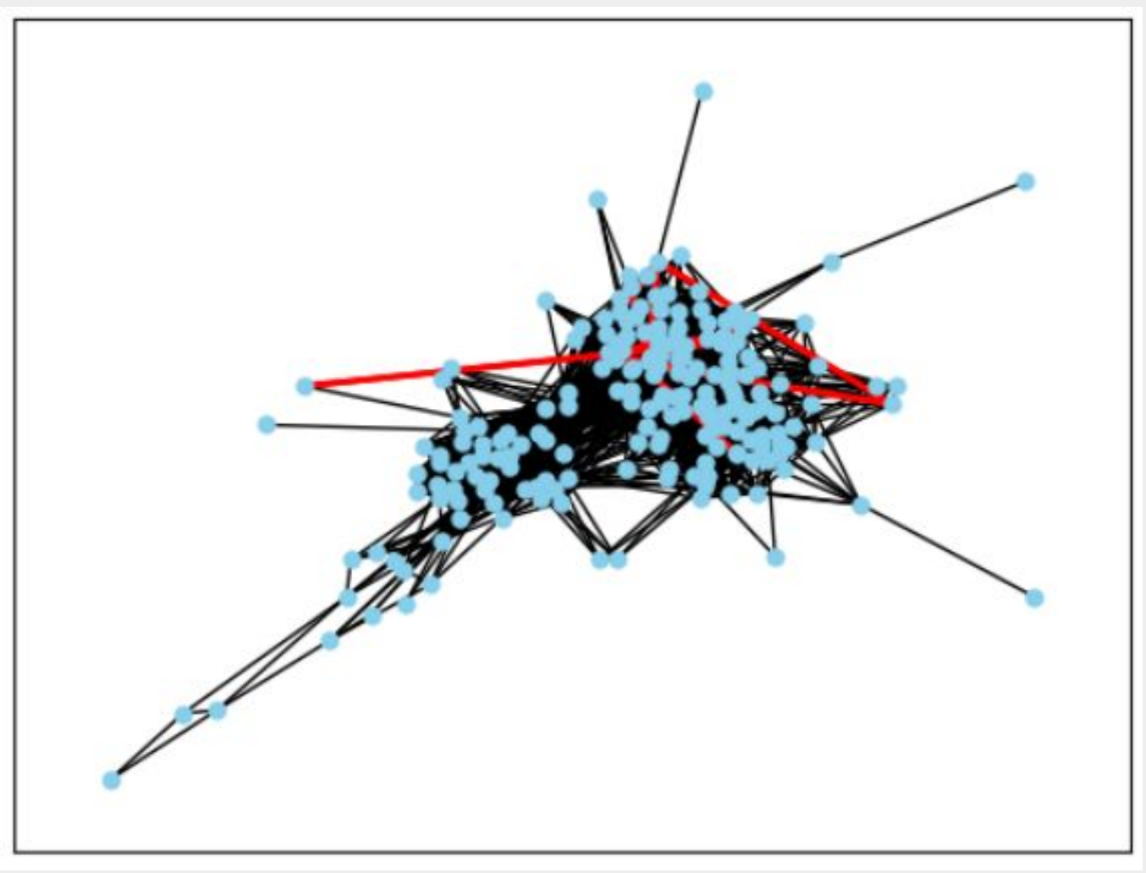
Pros:
- Can utilizes both local (node degrees) and global information (network structure)
- Since the jazz musicians networks has a skewed degree distribution, LHN_2 index can correct for the possible high degree bias that other indices have
- Good for detecting spurious links

Cons:
- Slow with large network size
- Assumes random network as null model

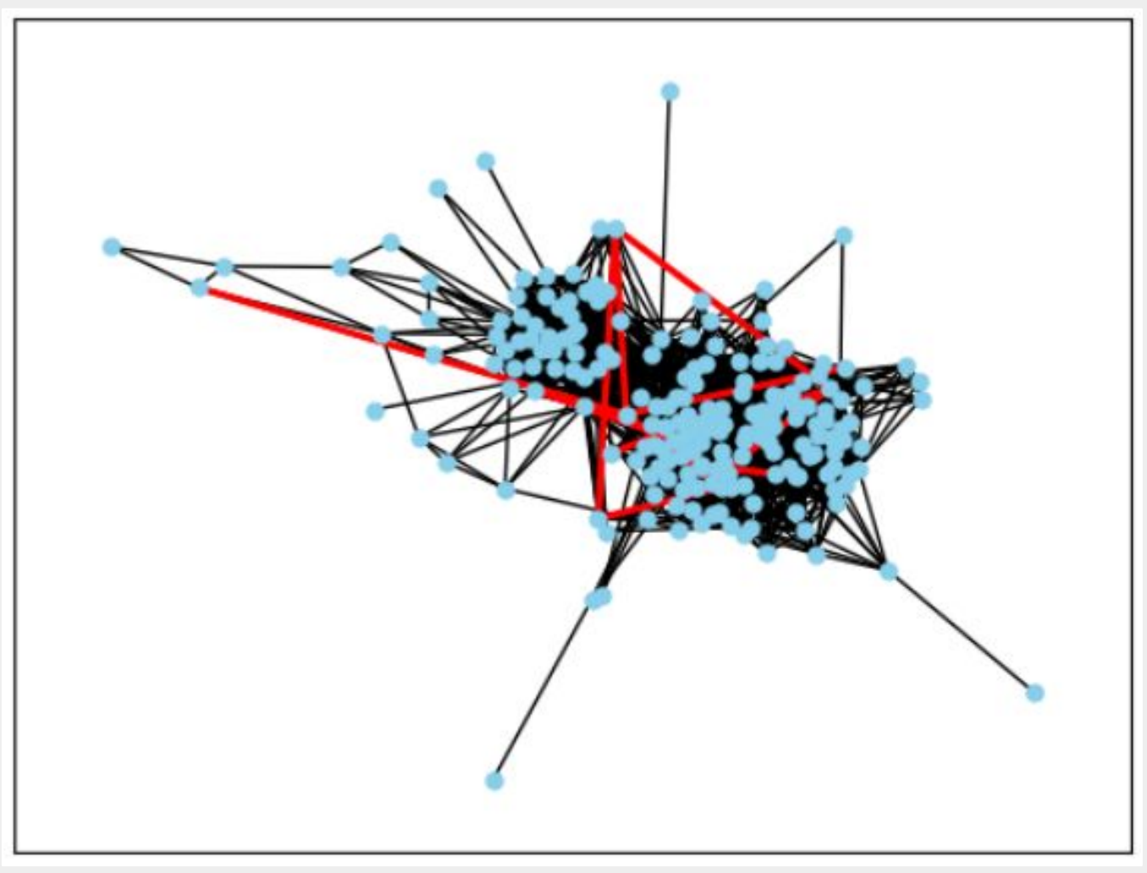
Spectral Comparison Index:

Pros:
- Easy to understand index
- Works with continuous features
- Robust to different scaling
- Make use of more dimensions than other indices

Cons:
- Assumes linear relationship between nodes
- Can not deal with missing data
- Does not make use of network structure
- Can be influenced by scaling if the nodes distances are very different
- Sensitive to outliers



Example graph with the 25 most spurious links marked in RED with Spectral Comparison index



Example graph with the 25 most spurious links marked in RED with LHN2 index
(Free parameter = 0.9)

Applied Solution

Testing Metrics: $Precision = \frac{l}{L}$ $AUC = \frac{n' + 0.5n''}{n}$

Also all of the edges that we use as the **probe set** (for this problem) are the edges of the Clique we are studying as spurious.

The testing metrics are run over a list of the indices **from smallest to bigger**, for this reason **the Metrics indicate spuriousness of the clique**. If the accuracy is high it means that most of the selected clique edges are part of the top ***L*** lowest scored edges. If the AUC score is high it means that the clique is **most likely NOT spurious**.

Results with Biggest Clique:

LHN			Spectral		
Edge	Value	Path	Edge	Value	Path
(108, 109)	-4.96 e-5	2	(32, 179)	6.34 e-4	2
(106, 107)	-1.87 e-5	2	(33, 179)	6.38 e-4	2
(66, 131)	-1.64 e-5	2	(35, 179)	6.47 e-4	2
(44, 108)	-1.54 e-5	2	(40, 179)	6.71 e-4	2
(122, 123)	-1.40 e-5	2	(32, 168)	6.85 e-4	2

Results with The second Biggest Clique:

LHN			Spectral		
Edge	Value	Path	Edge	Value	Path
(132, 178)	-4.58 e-5	2	(43, 197)	5.55 e-4	2
(106, 107)	-1.50 e-5	2	(43, 194)	5.66 e-4	2
(66, 131)	-5.92 e-6	2	(43, 182)	6.15 e-4	2
(44, 108)	-4.16 e-5	2	(43, 178)	6.33 e-4	2
(122, 123)	-3.75 e-5	2	(43, 174)	6.71 e-4	2

Also we can see that, for both cliques, the top lowest scores are from **completely different edges** for both techniques.

The data is not supervised, so we **don't have a direct comparison**.

Instead **we assumed that LHN is a appropriate index** for our data and measured how good our measure approximation is, by using the ***Cosine Similarity***:

$$p = \frac{s_{lhn} \cdot s_{spectral}}{||s_{lhn}|| \cdot ||s_{spectral}||}$$

In for both cliques the Similarity results are around **-0.3 ∈ [0,1]**.

Also the AUC and Accuracy results are underwhelming: Scoring a **0% Accuracy** in all of the cases and **more than 0.5 AUC** (that with this problem setting is a bad result).

Solution Reflection

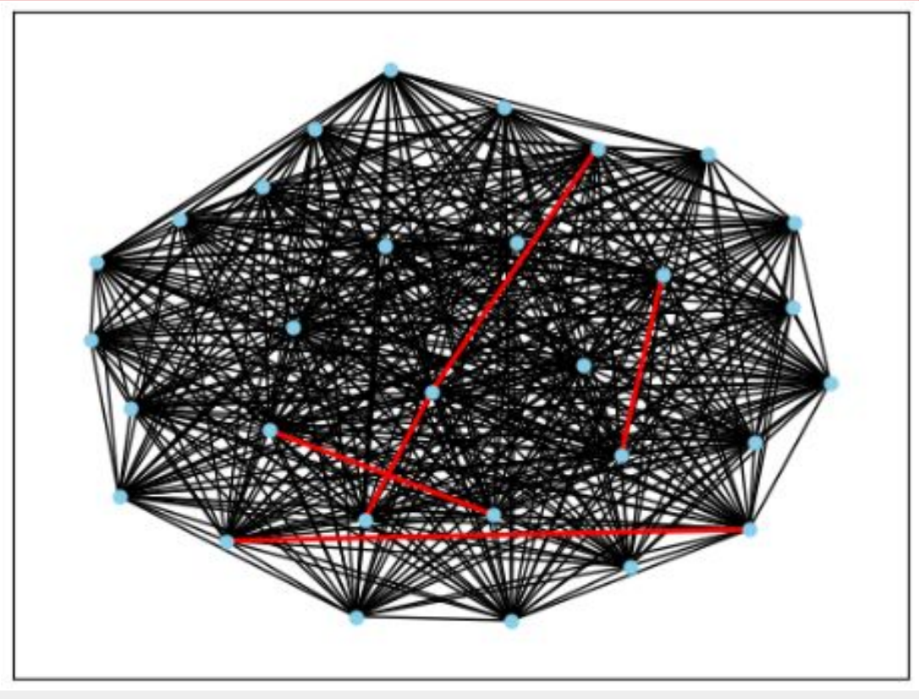
Reflection on results:
- The spectral index does not work as good as we had hoped.
- The LHN index is less suited for our task than initially thought.
- The unknowns are in fact not present leading to a poor signal to noise ratio
- The number of unknowns in the network is not known. The biggest or second biggest clique does not consists of unknown/spurious links.

Spectral index does no work as good as we hoped:
- The spectral index sees many edges which are close together as spurious
- The spectral index might be influenced by the few nodes with a high degree amount
- The spectral index selects many spurious edges which are related to the same node

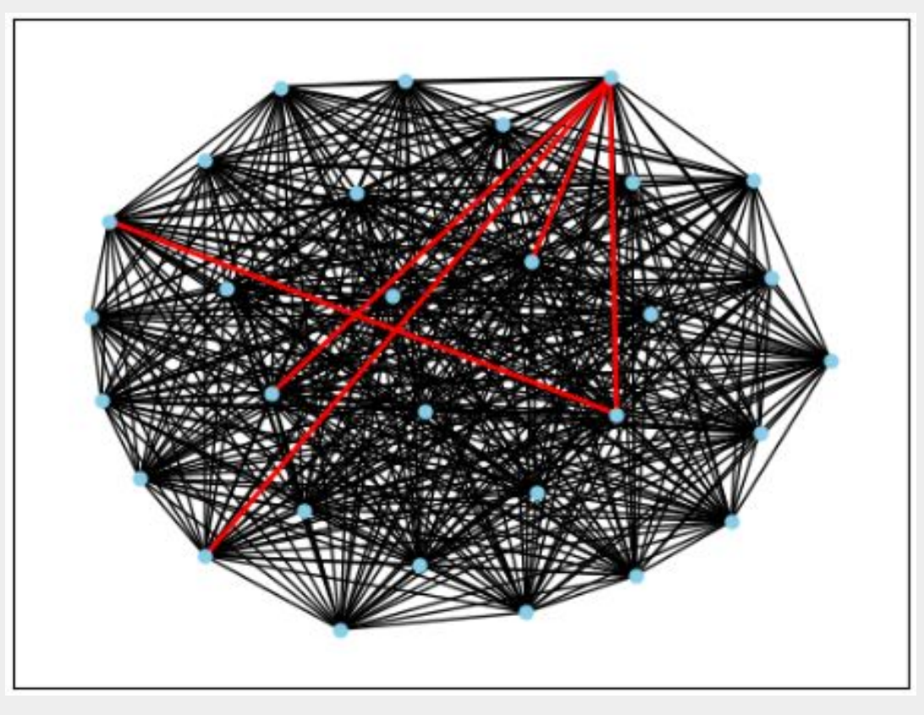
The LHN index is less suited for our task:
- The LHN index selects different edges than the spectral index
- The LHN index does give negative values to some links that could be spurious
- The LHN index finds spurious edges which are not in the biggest cluster

The unknowns are in fact not present leading to a poor signal to noise ratio:
- Both indices get very different results
- Unknown could be filtered out beforehand

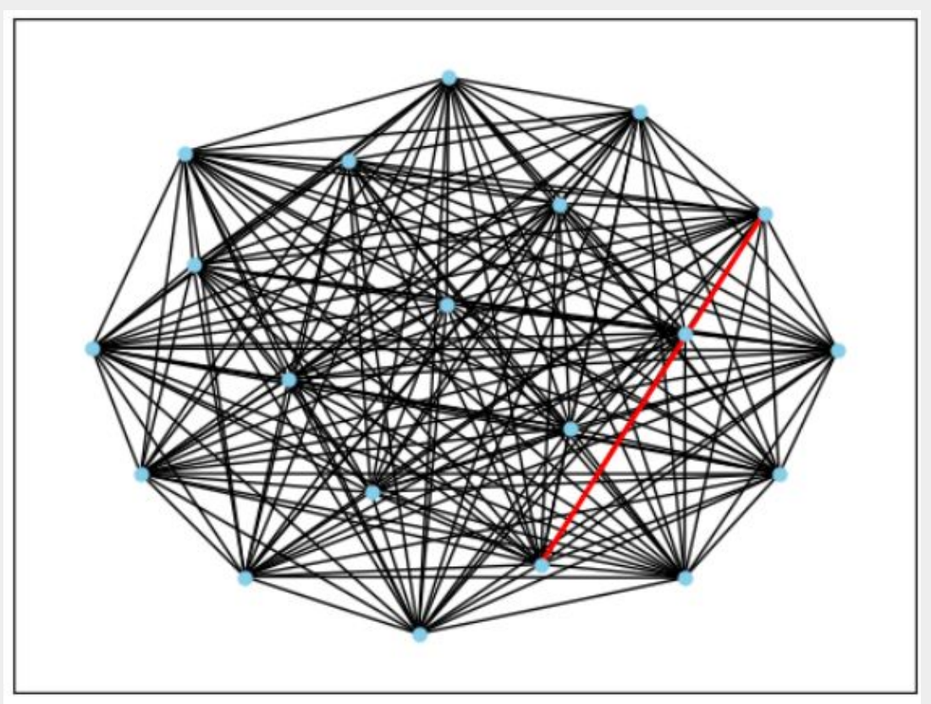
The number of unknowns in the network is not known. The biggest or second biggest clique does not consists of unknown/spurious links:
- The most spurious edges of the indices are not in the biggest clusters



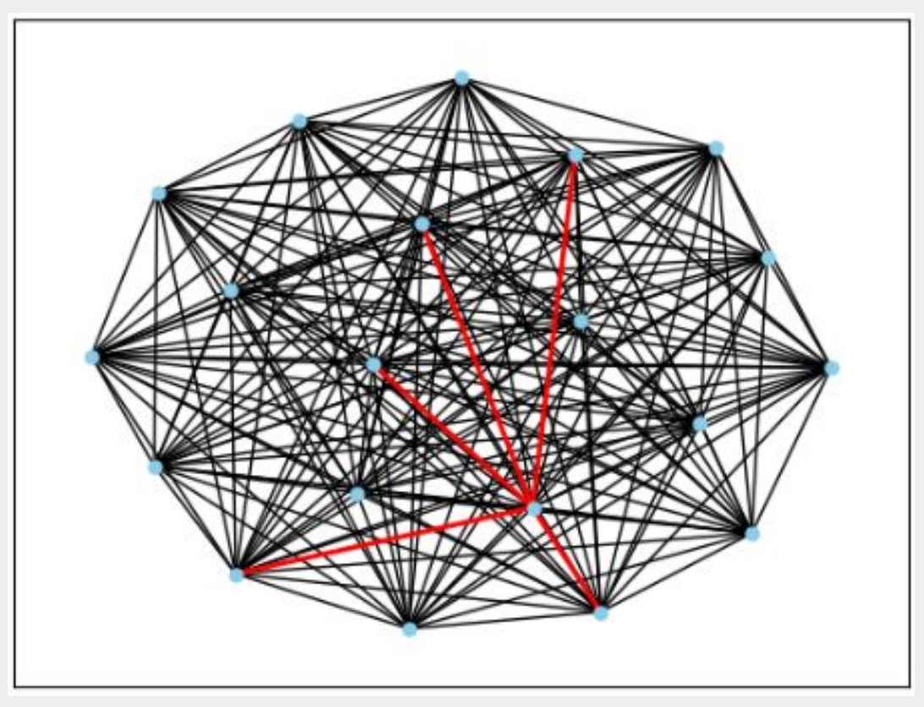
LHN2: Biggest cluster with the 5 most spurious links marked in RED



Spectral Comparison: Biggest cluster with the 5 most spurious links marked in RED



LHN2: Second biggest cluster with the 5 most spurious links marked in RED



Spectral Comparison: Second biggest cluster with the 5 most spurious links marked in RED

References

[1] **Community Strucutre in Jazz,** Pablo M., Gleisler and Leon Danon.

[2] **Link prediction techniques, applications, and performance: A survey,** Ajay Kumar , Shashank Sheshar Singh, Kuldeep Singh, Bhaskar Biswas

[3] **Link prediction in complex networks: A survey,** Linyuan Lü a,b,c, Tao Zhou a,d,