Data science for networks – Project report

Laura Arditti

DISMA – Department of Mathematical Sciences, Politecnico di Torino

Cascading failures in power grids

Propagation and vulnerability analysis

Physical model

In this work I study power grids, focusing on the problem of production and propagation of line outages. The simplest model describing the system is the DC model, which relates θ, the unknown voltage phase angles at nodes, to the active power injection at nodes P (the input data) by means of the Laplacian matrix, that describes both the topology and the electrical properties of lines

$$P{=}L(B)\theta.$$

The notation $L(B)$ indicates that the Laplacian is weighted with weight matrix B, the susceptance matrix of the grid.

The power flow over a line results to be proportional to the difference of the phase angles at its endpoints through the susceptance of the line $b_{(i,j)}$

$$f_{(i,j)} = b_{(i,j)}\left(\theta_i - \theta_j\right),$$

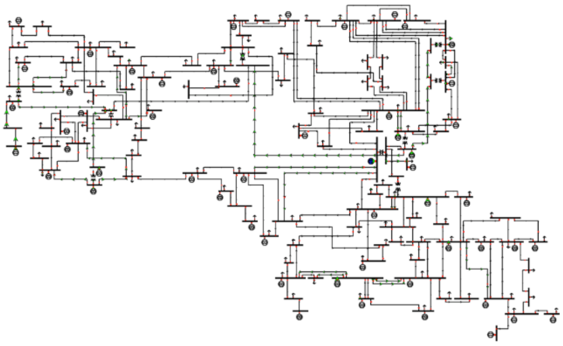
where the unknown θ is obtained inverting the Laplacian matrix (more precisely, by computing the Moore – Penrose pseudo-inverse, since $L(B)$ is a singular matrix)

$$\theta = L(B)^{-1}P.$$

Notice that θ depends linearly on the input P, but it depends non-linearly on the topology and line parameters, through the pseudo-inverse of the Laplacian, and so do f. When the power flow over a line exceeds its capacity, the line breaks and it is removed from the network. This causes the power flow on the failed line to be redistributed among the other edges. The redistribution is non-local, so that it can affect lines very far from the initial failure, and it is costly to compute. Repeated line overloads may turn into a cascade of failures, which is hard to predict.

Detection of vulnerable lines is very important to guarantee system robustness but very difficult to study analytically due to the non-linear dependence of flows on the topology. In this project, I propose a data science approach to the detection of vulnerable lines. It consists in the following procedure:

- Analyze data describing cascade events;
- Construct the influence graph;
- Apply data science techniques to understand the role of lines in the cascade process;
- Exploit available information to predict unobserved dependencies.



In particular, I studied the IEEE 118-Bus System, which is a simple approximation of the American Electric Power system as of 1962. It is represented by a graph with 118 nodes and 177 lines.

Simulation of the cascade process

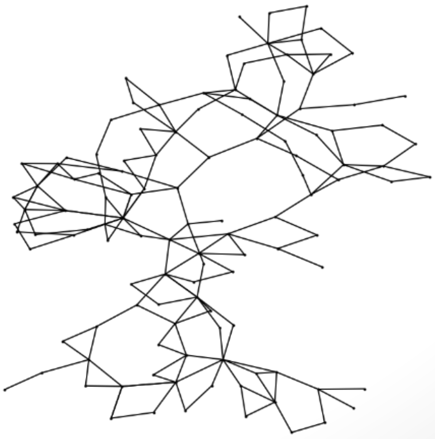
To simulate the evolution of a cascade triggered by the removal of a line from the network I used MATPOWER, a free, open-source Electric Power System Simulation and Optimization Tools for MATLAB and Octave, which implements the DC model.

The simulation of a cascade consists in the following steps:

- Compute the initial state of the grid by solving the power flow equations for nominal values of generators and loads;
- Simulate the failure of a line by removing it from the network;
- Compute the new power flow distribution and check if any power flow exceeds the line capacity;
- If so, remove the line from the graph;
- Iterate until the cascade stops or the graph becomes disconnected.

I repeated the simulation once for each line and I stored the data about the cascade in form of sequences of lines, each sequence corresponding to a different initial failure acting as a triggering event.

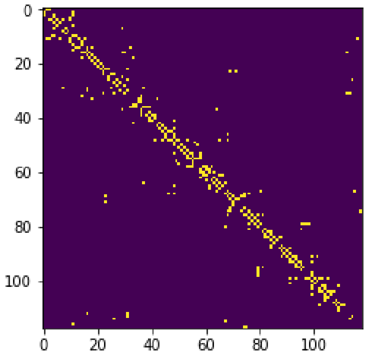
Data analysis



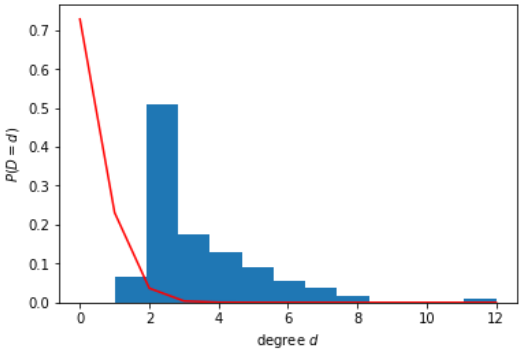
I analyzed these data with Python – NetworkX.

The picture shows the geoGraph, representing the geographical network.

Its degree distribution shows that most nodes have low degree, from 2 to 4. Most non-zero entries of the adjacency matrix are located around the diagonal, as expected from the decentralized nature of the power grid. The network also exhibits little transitivity: the clustering coefficient, i.e. the fraction of all possible triangles present in the graph, is 0.135.

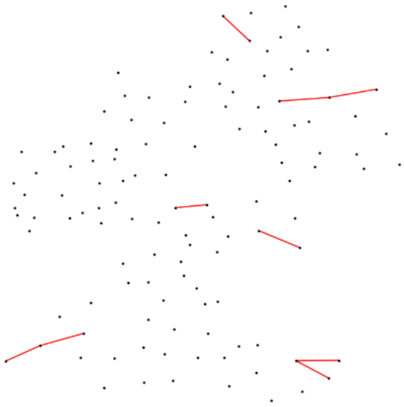


Adjacency matrix

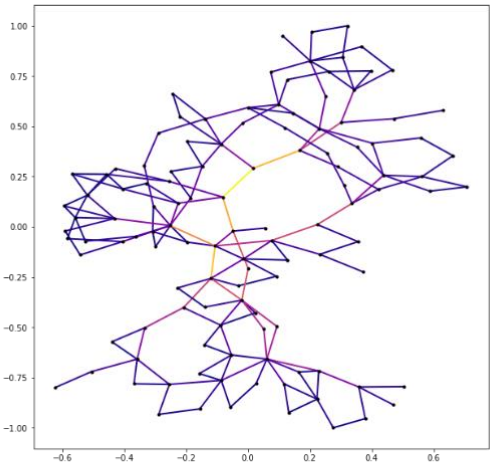


Degree distribution compared with binomial

A bridge is an edge whose removal causes the number of connected components of the graph to increase. The network possesses some bridges, which connect peripherical nodes to the main component. An important quantity related to the redistribution of flows over the network is edge betweenness centrality, but it shows bad performance in the case of power flows. The reason can be understood from its definition: betweenness centrality of an edge is the sum of the fraction of all pairs of shortest paths that pass through it. The length of paths has a central role in many transportation problems, e.g. traffic, making shortest paths "central" in the network, but it's irrelevant for power redistribution and that's why the index fails in identifying critical lines. In the picture, the most central edges are colored in yellow.

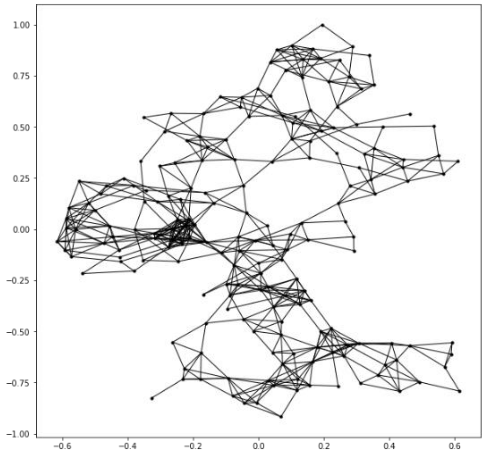


Bridges

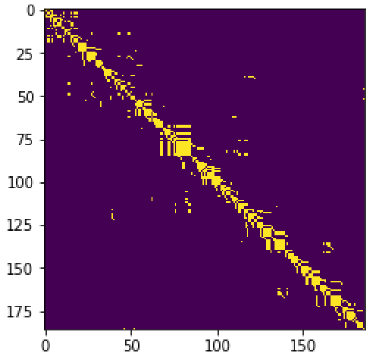


Edge betweenness centrality

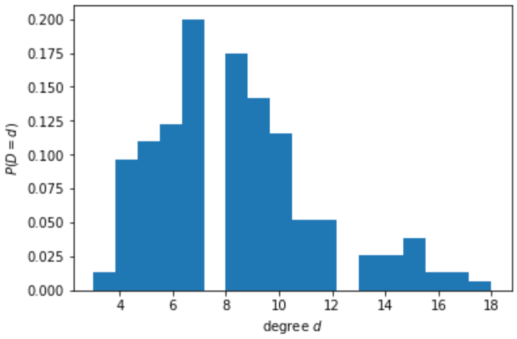
GeoGraph is not the most suitable way to represent the network, since it relates nodes while the flow dynamics happens on edges. A better representation is given by the dual graph of the geographical network which is defined as follows. Nodes of the dual graph are edges of the geoGraph and two nodes in the dual graph are adjacent if and only if the corresponding edges in the geoGraph are incident.



Like the geoGraph, the dualGraph has a mostly diagonal adjacency matrix, since the network is still decentralized, but the degree distribution reaches higher values, since edges are inherently more connected than nodes.

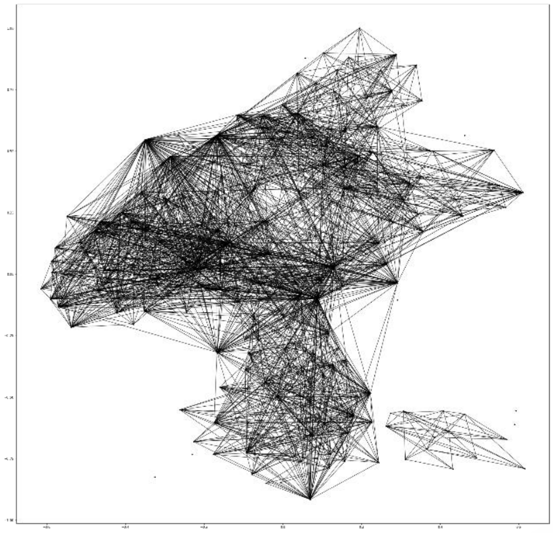


ADJACENCY MATRIX



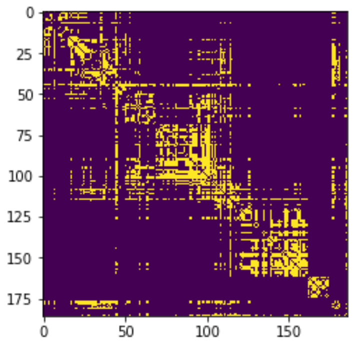
DEGREE DISTRIBUTION

The electrical dependencies are encoded in the influenceGraph, which is built according to the simulation. Again, nodes of the influence graph are edges of the geoGraph. If the failure of a line i causes the failure of a line j, an edge (i,j) is present in the influence graph. Notice that with this construction the influenceGraph is undirected so that causality relation is not represented but a wider class of algorithms can be applied.

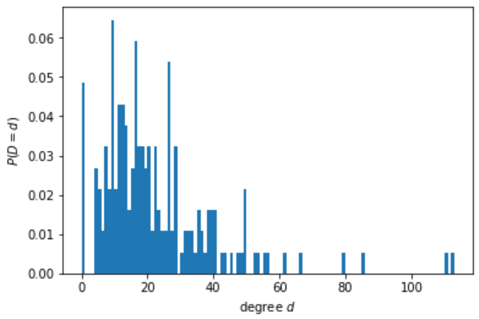


This corresponds to consider equally "central" an edge whose failure triggers the failure of many others and an edge which is very fragile, i.e., whose failure is a consequence of many distinct events.

Non-locality of the cascade process is express by the fact that nodes distant in the dual graph are connected in the influence graph. Also lines far from each other are electrically connected, and this is clearly visible from the adjacency matrix: non-zero entries are no more concentrated around the diagonal. The clustering coefficient is now higher: the fraction of all possible triangles present in the graph is 0.416. The degree distribution is more spread and there are nodes with high degree that represent lines frequently involved in cascades.

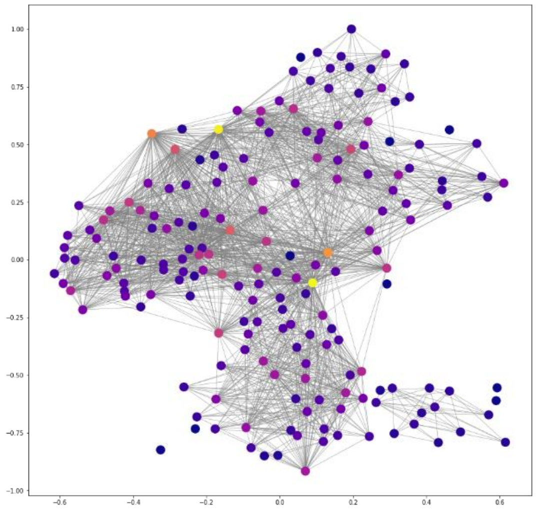


ADJACENCY MATRIX

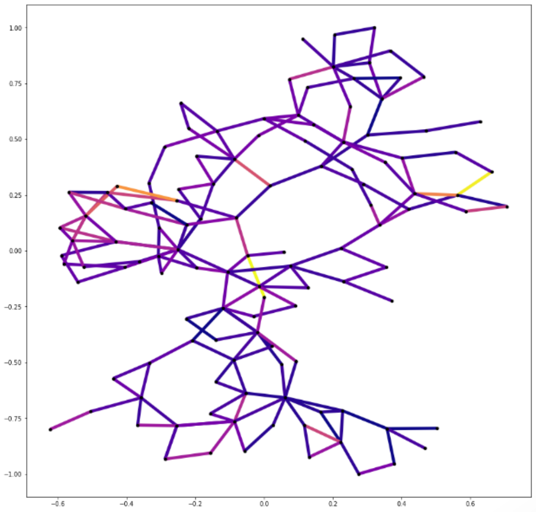


DEGREE DISTRIBUTION

These lines are identified by computing the degree centrality of nodes in the influence graph, which for a node v is the fraction of nodes it is connected to. The result is transferred to the groGraph by coloring edges according to the centrality of corresponding nodes in the influence graph.

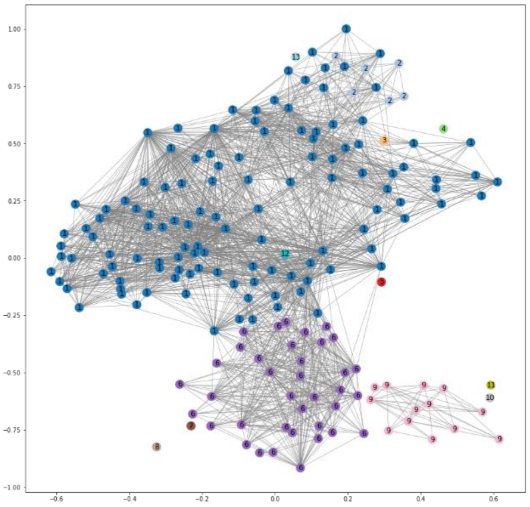


NODE DEGREE CENTRALITY OF THE INFLUENCE GRAPH

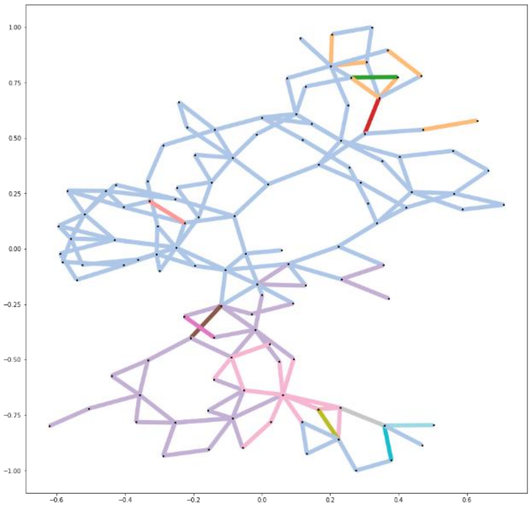


«EDGE» DEGREE CENTRALITY OF GEOGRAPH

Communities in the influence graph represent groups of edges tightly correlated. The communities were detected with the Girvan Newman algorithm that detects communities by progressively removing edges from the original graph. The algorithm removes the “most valuable” edge, the edge with the highest betweenness centrality, at each step. As the graph breaks down into pieces, the tightly knit community structure is exposed and the result can be depicted as a dendrogram. The pictures show 13 communities and correspond to the second level of the dendrogram, i.e., the second iteration.

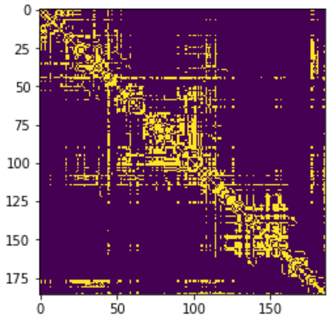


Communities of the influence graph



«Edge» communities of geoGraph

The previous analysis underlines that even if the dual and the influence graph share the same nodes they have different edges, since one represents the real system topology while the other represents the electrical dependency. The difference of their adjacency matrices is full, showing that geographical and electrical correlation are mostly unrelated.

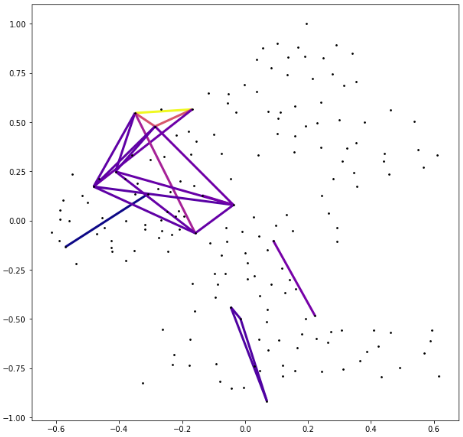


Link prediction on the influence graph

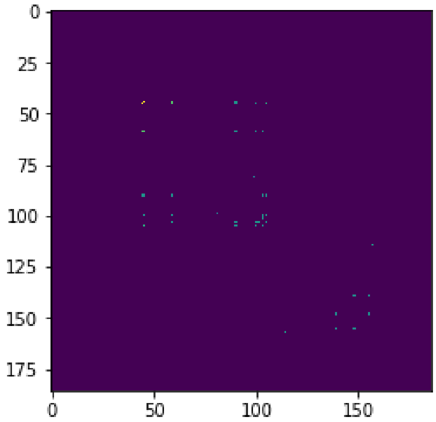
Link prediction on the influence graph can reveal dependencies that were not observed, for example because the simulation did not consider cascades triggered by simultaneous multiple initial failures or because the DC model is an approximation of the real system behavior. To perform link prediction on the influence graph, different techniques, both local and global, were first evaluated and then applied. The evaluation metrics is based on twofold validation with ROC score and average precision. The set of edges was split into 2 even parts to form train and test sets. Only links whose prediction index were greater than 0.4 were kept. The total of 2031 edges of the influence graph was partitioned as follows: 1016 training edges (positive), 1015 test edges (positive), 1015 test edges (negative).

In all cases, the predictions seem reliable, since the ROC and AP scores are high. The predictions look a bit different, but this is mostly due to the fact that different techniques give different distributions of scores along the range, i.e., a different ratio among average and maximum score. Apart from this, they agree on the link with the highest score, which is the brightest one. Among local indices, resource allocation achieves the best result: this makes sense, since the resource allocation index is defined in terms of flow distribution. Katz similarity index achieves an even better result, as expected from a global index. The rest of this section is devoted to show the results obtained with each method.

Adamic – Adar: Test ROC score: 0.908, Test AP score: 0.887.

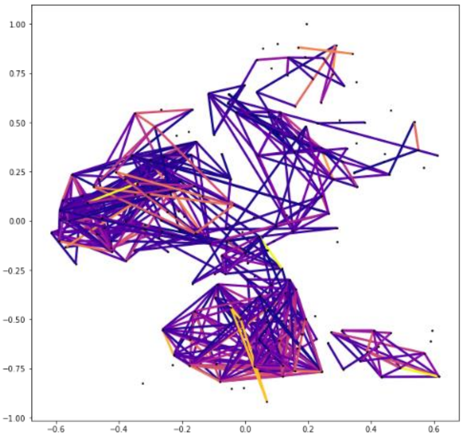


PREDICTED ADDITIONAL EDGES

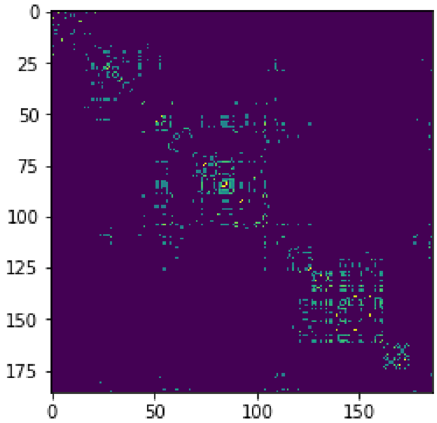


ADJACENCY MATRIX OF PREDICTED ADDITIONAL EDGES

Jaccard coefficient: Test ROC score: 0.837, Test AP score: 0.819.

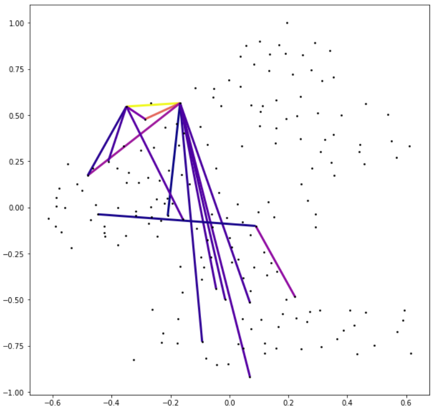


PREDICTED ADDITIONAL EDGES

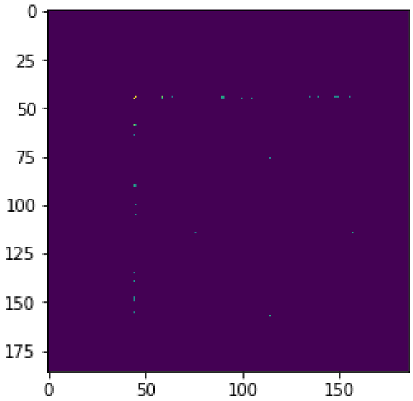


ADJACENCY MATRIX OF PREDICTED ADDITIONAL EDGES

Preferential attachment: Test ROC score: 0.806, Test AP score: 0.797.

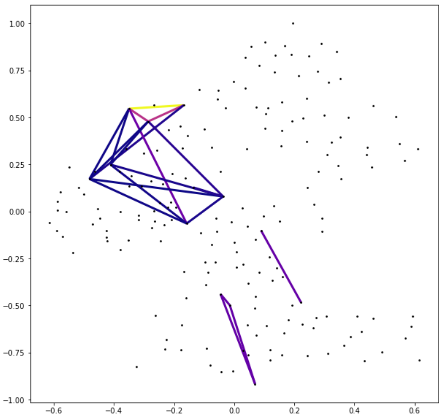


Predicted additional edges

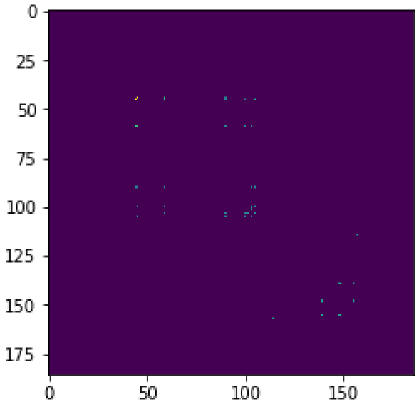


Adjacency matrix of predicted additional edges

Resource allocation index: Test ROC score: 0.913, Test AP score: 0.896.

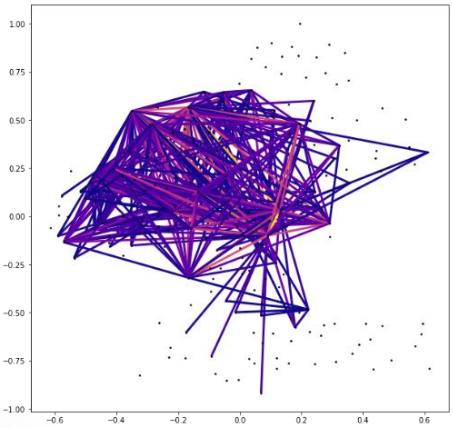


Predicted additional edges

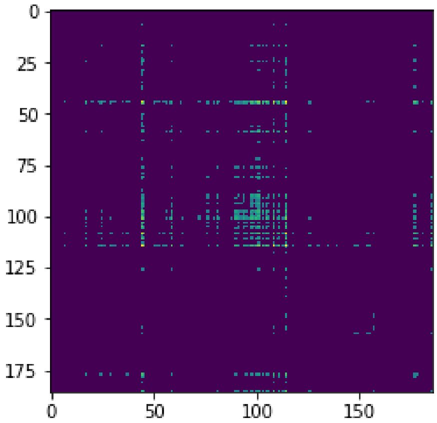


Adjacency matrix of predicted additional edges

Katz similarity index: Test ROC score: 0.974, Test AP score: 0.949.



Predicted additional edges



Adjacency matrix of predicted additional edges

Conclusions

In this work I proposed a data science approach to the analysis of power grids. I aimed to show that it could be preferable compared to an analytical approach, since it could be based on empirical data, it does not require expensive simulations and it might overcome the limitations arising from a simplified model.

The analysis performed on the influence graph has a prescriptive nature in the following sense. It can help identifying vulnerable lines so that the limited available resources can be devoted to their reinforcement, making the action of the system operator more effective in achieving network robustness.

On the other hand, link prediction analysis has a descriptive nature. The predicted additional edges are not necessarily vulnerable ones, then the outcome of this analysis should not be regarded as a way to direct the efforts to strengthen the network. Instead, these edges represent information on the electrical dependencies that is probably missing: this information can be useful to justify the observed behavior of the system, that is often unexplained or misunderstood.