Report: communicate the results (20 points)[¶](http://localhost:8888/notebooks/NMF%2BGL/Case4.ipynb#Report:-communicate-the-results-(20-points))

1. (5 points) What is your business proposition?

Our business proposition is to apply web traffic network into forecasting future web traffic for 145,000 Wikipedia articles. The web traffic network can be used to analyze the Wikipedia articles to produce users structures reports and data hotspot results in different time series. Those reports and results are straight-forward to help make a decision or be used for deeper learning.

1. (5 points) Why this topic is interesting or important to you? (Motivations)

Our business proposition is to apply web traffic network into forecasting future web traffic for 145,000 Wikipedia articles. The web traffic network can be used to analyze the Wikipedia articles to produce users structures reports and data hotspot results in different time series. Those reports and results are straight-forward to help make a decision or be used for deeper learning.

(3) (5 points) How did you analyse the data?

Our data comes from kaggle(https://www.kaggle.com/c/web-traffic-time-series-forecasting)

In our case study, we focus on two in particular: cohesive spatial regions (nodes) and relationships between those regions (edges). It makes sense, for the reason that: graphs and networks of the sophisticated systems can bring advances to people’s understanding of complex systems and interactions within them. The web traffic can be viewed as a network structure, in which each Wikipedia articles are organized into multiple homogeneous regions, and complex interactions exist between articles within and across different regions. However, the nodes and edges in web traffic networks are usually not given and should be derived from the brain imaging data. Generally speaking, the discovery of web networks is required before one can conduct web traffic analysis. Therefore, we regard this work as the baseline for further business study, especially for our case.

Previous works in this line usually focus on inferring the edges based on known groups, such as graphic lasso, or inferring the groups by spatial clustering, such as K-Means and spectral clustering. Besides, the independent inference framework infers the nodes and edges of the network separately. These methods, however, are limited due to some reasons.

In order to finish this work and provide some novel views, we propose a combination model of graphical lasso and matrix tri-factorization with complex constraints and spatial regularization. To be more specifically, we consider a derivative model of GLasso, which desires estimation of precision matrix of different groups and a clustering result:

Where is the empirical covariance matrix and penalty is a regularization. The item of is equivalent to K-means clustering, where is an orthogonal nonnegative matrix and be regard as cluster indicator matrix here.

This is an unsupervised machine learning method, which can discover nodes and edges of web traffic network collectively. We show that this formulation works well in controlled experiments with synthetic networks and be able to recover the underlying ground-truth network.

(4) (5 points) How does your analysis support your business proposition? (please include figures or tables in the report, but no source code)

We first evaluate our model using synthetic data, where ground-truth is available. On edge detection, we compare NMFGL with GLasso. On group detection, we compare NMFGL with two clustering methods: k-means and spectral clustering.

To evaluate the quality of edge detection, we define the accuracy of edge detection as , where is the number of true edges detected by the algorithm, is the number of true edges. We control the sample size and the value of beta (internal density of each groups), respectively.

To evaluate the quality of group inference, we use homogeneity score, normalized information (NMI) score and purity score. Higher value of each score indicates better quality of group detection.

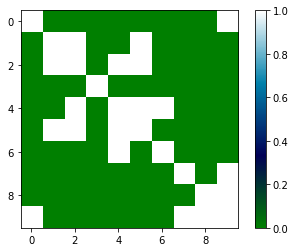


Figure Precision Matrix underlying the toy dataset

Figure 1 shows us the real precision matrix underlying the toy dataset. In the figure, we can figure out the connection among each group. Figure 2,3,4 shows the comparison between NMFGL, K-Means and spectral clustering in terms of different sample size. We can observe from these figures that for our toy dataset, NMFGL achieves better than spectral clustering, however, preform slightly worse than K-Means. But the results are very close. It indicates that NMFGL may not have much advantages compared to K-Means when the components in the network are very isolated, but such case is relatively rare in real-world dataset such as the dataset of web traffic network. Figure 5 shows the comparison between NMFGL and graph lasso in terms of different beta. However, according to the increasing value of beta, the result shows us that Glasso works robuster than our new method, which demonstrates that we need more works to study in the future. Figure 6,7,8 shows us the comparison between NMFGL, K-Means and spectral clustering in terms of different beta. We can observe a similar result on the scenario of comparison on the terms of different sample size.

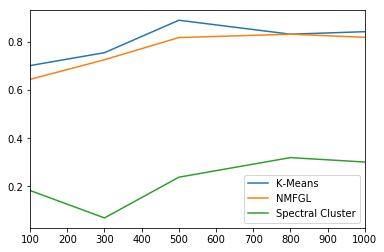


Figure 2 HomoScore on different sample size

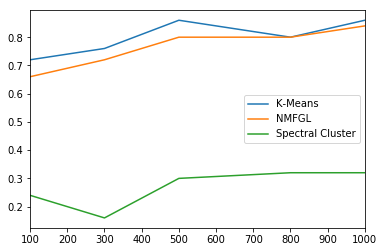


Figure 3 PurityScore on different sample size

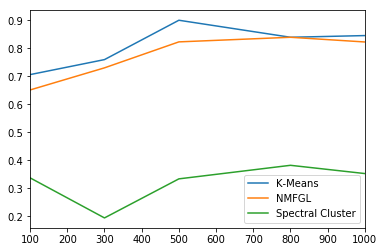


Figure 4 NmiScore on different sample size

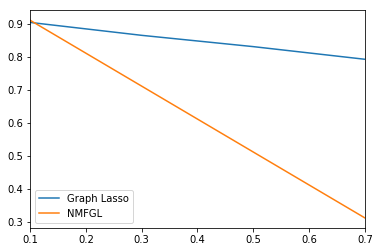


Figure 5 Edge Accuracy on different beta

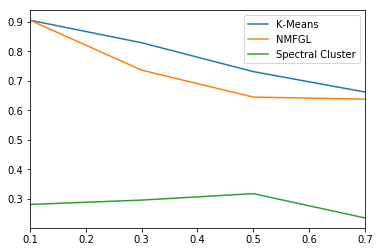


Figure 6 HomoScore on different beta

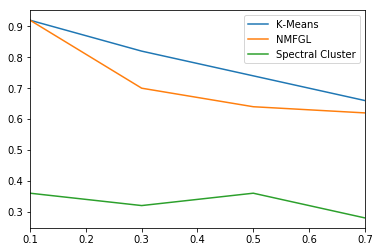


Figure 7 PurityScore on different beta

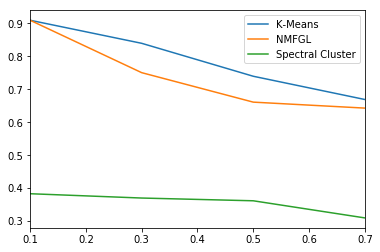


Figure 8 NmiScore on different beta

After evaluation on the toy data, we implement it on the dataset of web traffic in terms of edge detection and group clustering. The Figure 9 shows the result of NMFGL, involving these two things that we focus in particular. In our opinion, we could consider to use this result as the outline to guide our future study., the yellow nodes in the figure means that the groups related to them have direct connection, and the purple ones means that they have no connection or only indirect connection. So the comparison to covariance matrix underlying the dataset, this precision matrix could provide us a more objective perspective for the relation among each node. For example, in the terms of covariance matrix, we may observe that node A and B have a similarity between each other. however, this relation maybe caused by the relation between AC and the relation between BC, which means that the relation between AB is an indirect relation. We believe that a clear precision matrix could bring us a deeper understanding of web traffic network. By finding cluster of nodes and edges among different groups, we can build up a more trustworthy standpoint for prediction analyze in the next step, which hasn’t been finished due to the limited time.

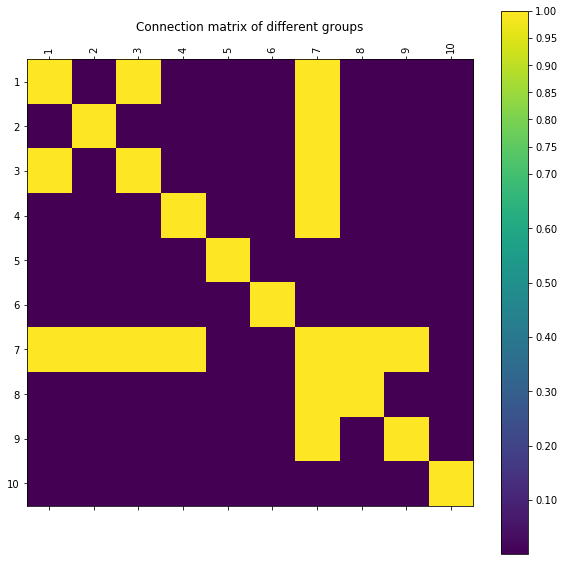


Figure 9 The Estimation of Precision Matrix Underlying the Ground-Truth Data