UNIVERSITY OF SYDNEY

MATH3888

PROJECTS IN MATHEMATICS

Principal Component Analysis for Network Centrality Analysis

JONATHAN ORON

November 5, 2023

Contents

1 Introduction

1.1 Biological Background

The process of developing treatments for different genetic disorders is extremely complex, and requires considering the interactions between different proteins. Since protein-protein interactions are modelled by complex networks or graphs, biochemists often enlist mathematicians to assist with the analysis of these networks. The challenge presented to mathematicians involves simplifying the network and identifying nodes that carry some sort of relevance or importance.

1.2 Mathematical Theory

The accepted method for identifying these target nodes consists of two key steps:

1. Partitioning the network into communities, where communities consist of nodes that interact closely with other members.
2. Using centrality measures to select the most interesting or important nodes from relevant communities.

This report will focus on the second step by introducing a method to consider multiple different centrality measures in a combined index to capture as much information about the importance of different nodes as possible.

1.3 Centrality Measures

Centrality measures are indices or scores used to rank or quantify how central a given node is in relation to other nodes in a network. Different centrality measures may consider how many neighbours a node has, how close a node is to other nodes, the importance of a node's neighbours as well as many other metrics [2].
Some of the important centrality measures this report will consider include [1]:

- Degree Centrality: a measure of the number of edges leading into a given node.
- Closeness Centrality: a measure of how close a given node is to all other nodes in the graph.
- Eigenvector Centrality: a measure of prestige, where the prestige of a node is obtained as the sum of the prestige scores of its neighbours.
- Betweenness Centrality: a measure of the importance of a node in terms of connecting other nodes on the network. For example, nodes that are part of a bridge will have very high betweenness centrality.
- Subgraph Centrality: a measure characterising the participation of a node in all possible sub-graphs of the general network [3].

- Information Centrality: is a measure of the average amount of "information" stored in the paths between a given node and all other nodes. Here information is defined as the inverse of the path length between two nodes [4].

1.4 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method most commonly used to reduce the dimensionality of complex data. This method involves combining the information from many variables into a smaller subset of variables called the principal components. The quality of the representation of information by the principal components is measured by the total amount of the original variables' variance that can be explained [5].

Mathematically, the general process of PCA is as follows:

1. Form an $n \times d$ data matrix where each of the n rows represents a node of the network, and each of the d columns represents a centrality measure.
2. Compute the correlations of the centrality measures to determine whether some can be dropped, and remove those which are not required. Variables that are highly correlated with other variables can be predicted using other measures, and therefore can be dropped.
3. Standardise the data. This is important because the magnitude of the different centrality measures can vary largely. This ensures that the measures are comparable in their contribution to overall variance. The preferred method of standardisation is through computing Z-scores (subtracting the mean of each measure, and then dividing by the standard deviation).
4. Compute the eigenvalues of the matrix. For each eigenvalue compute the corresponding eigenvector.
5. Select the k largest eigenvalues. Here, k should correspond to the new number of dimensions desired. For each of the k eigenvalues select their corresponding eigenvectors. These eigenvectors are the new components.
6. These k eigenvectors now form a new subspace, and the information for each of the centrality measures is projected onto this new subspace.

1.5 Objective

This report will investigate how Principal Component Analysis can be used to construct an index composed from a linear combination of other centrality measures, which captures different elements of the importance or centrality of a given node. The goal is to create a score that incorporates many different centrality measures to measure cumulative importance. Furthermore, this report will explore a brief analysis of potentially interesting nodes by analysing the relationship between the first and second principal components obtained through the PCA.

2 Approaches and Results

2.1 Standardising the Data

The first step to conduct the PCA involves preparing the data. Using the NetworkX package in Python, the centrality measures for the yeast protein-protein interaction network were calculated. Using the Pandas and Numpy packages, the centrality measure vectors were standardised and stored in a data-frame. (Refer to Appendix 1 for GitHub link containing source code)

The process of standardising the data is paramount since each of the centrality measures essentially incorporates a different feature of the network. This means that the scales of each of the measures is different, making it difficult to compare the different centralities. To remedy this, the data is standardised using Z-scores. In other words, subtracting the mean of each centrality measure from each observation, and then dividing by the standard deviation.

The reason for choosing to standardise the data in this way as opposed to other commonly used methods such as dividing by the $L2$ norm is to retain the shape of the distributions whilst centering the data consistently. Since PCA involves analysing the variance of each measure, the standardisation should not influence the distributions of the centrality measures, and instead only re-scale them such that they can be compared fairly.

2.2 Visualising Relationships and Centrality Selection

Before conducting the PCA, it is important to ensure that all of the centrality measures considered are appropriate. From Figure 1 below, it is clear that some of the centrality measures are highly correlated, whilst others appear to have little relationship to one another. For the Principal Component Analysis, it is important that only measures that contain new information that is not contained in the other centralities are included. This ensures that the weights attributed to each of the measures depends on how much information they carry independently.

To determine which variables to include, the pairwise plots were considered, but also the correlation between each of the centrality measures (refer to Appendix 2 for correlation matrix). Since it was observed that the Eigenvector and Subgraph centrality measures were almost perfectly correlated (0.9675), only one of the measures needed to be included in the PCA. In practice, the decision of which measure to drop is arbitrary, however the argument can be made that since Subgraph centrality is more computationally expensive to calculate, it is sensible to drop it rather than Eigenvector centrality.

Whilst not directly important at this stage, it is interesting to note the relationship between Betweenness centrality and all of the other measures. Both in terms of correlation and the plots in Figure 1, it appears that Betweenness centrality has no discernible relationship with any of the other measures.

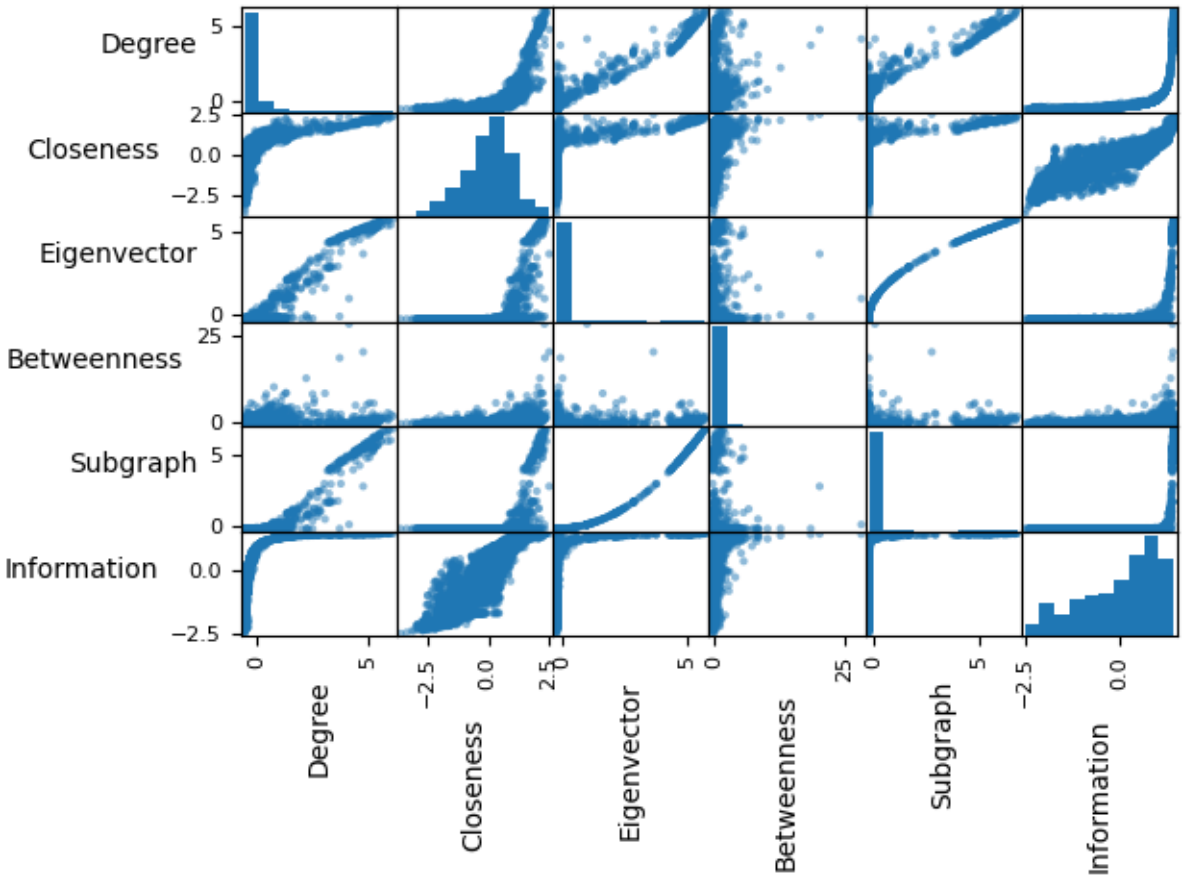

Figure 1: Pairwise Plots of Centrality Measures

2.3 Conducting Principal Component Analysis

The Principal Component Analysis was conducted computationally using the Python sklearn package. The first step is to form a matrix out of the data, where each column of the matrix represents a centrality measure, and each row of the matrix represents a node in the network. Using python, the eigenvalues of this data matrix and their corresponding eigenvectors are found to be:

λ_1	3.043	[0.520, 0.498, 0.445, 0.272, 0.458]
λ_2	1.014	[-0.384, 0.318, -0.610, 0.461, 0.408]
λ_3	0.797	[0.131, -0.307, 0.110, 0.837, -0.419]
λ_4	0.123	[-0.148, 0.746, 0.0244, -0.031, -0.648]
λ_5	0.024	[0.738, 0.0189, -0.646, -0.107, -0.166]

Now that the eigenvalues and eigenvectors have been computed, the proportion of variance ex-

plained by each component can be computed. The computation is as follows:

$$\text{variance explained by component } i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

For the centrality measure data, this is found to be:

Component	1	2	3	4	5
Proportion of Variance Explained	0.608	0.203	0.159	0.025	0.005

The first component explains $\approx$ 61% of the variance. Whilst this is not perfect, a substantial proportion of the information from the whole data set is contained in the PCA. The next step is to select the k largest eigenvalues and their corresponding eigenvectors. Since the goal is to create an index, only the largest eigenvalue and the corresponding eigenvector are required. To calculate the weights of each centrality measure contained in the first component, the eigenvector is divided by its $L1$ norm:

$$\vec{w} = \frac{\vec{v}}{\sum_i |v_i|}$$

Applying to the first component found earlier:

$$\vec{w} = \begin{bmatrix} 0.237 & 0.227 & 0.203 & 0.124 & 0.209 \end{bmatrix}$$

This weights vector is interpreted as the weight of each centrality measure in the index. In other words, these are the weights for the weighted average of the centrality measures, where the weights correspond to the proportion of variance of the first principal component that each measure explains. From this vector, the importance score of a node can be calculated as:

$$\text{Importance Score} = 0.237 \times \text{Degree} + 0.227 \times \text{Closeness} + 0.203 \times \text{Eigenvector} + 0.124 \times \text{Betweenness} + 0.209 \times \text{Information}$$

2.4 Extension to Second Principal Component

One of the consequences of considering only the first principal component is that only approximately 60% of the variance is explained. Furthermore, it is difficult to take into account other interesting aspects of the data. This is because the importance score index calculated from the first principal component only takes into account those nodes with the largest cumulative score of centrality measures. For this reason, this report will provide a brief exploration into how the second principal component can be incorporated into identifying nodes that are potentially interesting.

Now, the same process as before is repeated, but the two largest eigenvalues and their corresponding eigenvectors are selected. The benefit of considering both the first and the second components is that an even greater proportion ($\approx$ 80%) of the variance can be explained.

Whilst in theory, considering the second component in addition to the first should provide more insight into selecting potentially important, or at the very least interesting nodes, often it is difficult to interpret the importance of the different components. Figure 2 below, is a plot of all of the centrality measures projected onto the first two principal components. Based only off of this plot, it is extremely difficult to determine which groups of nodes could potentially be of interest.

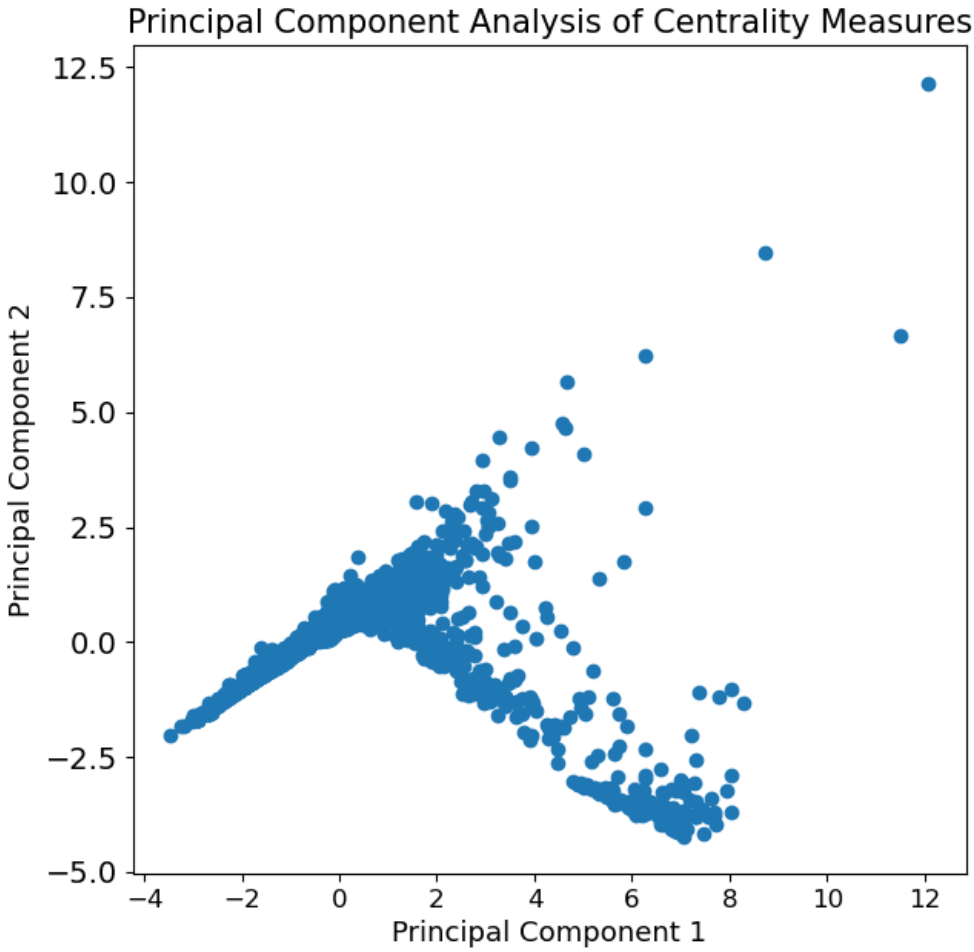

Figure 2: Plot of Data Projected onto the First Two Principal Components

To try and better understand what is being shown in Figure 2, a special focus was paid to betweenness centrality. Previously it was noted that betweenness has very weak correlation to the other centrality measures. As a result, it was hypothesised that some important nodes could have some dependence on betweenness.

To try and analyse the impact of betweenness, nodes were classified as either low or high betweeness.

- Nodes with a betweenness centrality less than the mean betweenness centrality of all the nodes were classified as "Low".
- Nodes with an above average betweenness centrality were classified as "High".

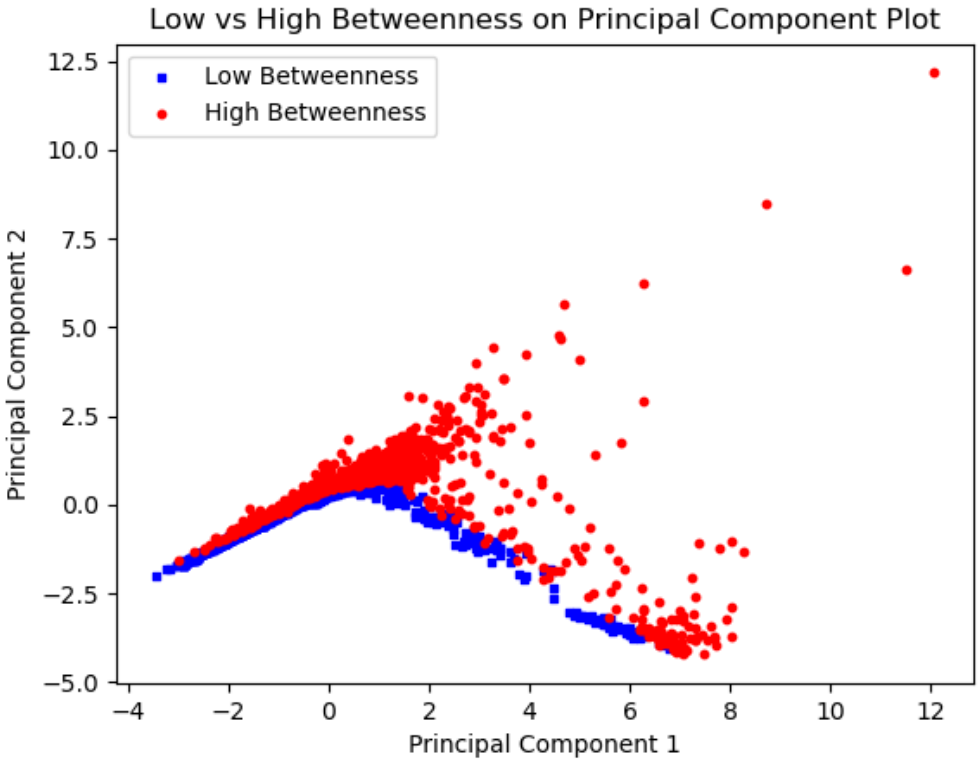

Figure 3: Data Projected onto First Two Principal Components by Betweenness Centrality Value

It was interesting to see from Figure 3 that the second principal component essentially determined whether a node had high or low betweenness centrality. From this plot, a potential idea for finding interesting nodes could involve filtering for only low betweenness nodes, and then selecting those that correspond to high values on the first principal component.

When applying the index created using only the first principal component to a biological network to suggest potentially important nodes to biochemists, the feedback was that in general the nodes identified were "too important". Biologically, the nodes were too vital to the basic functioning of yeast to be potential targets to be knocked out. As a result, many of the nodes that were identified could not be used for experiments.

Using the new findings presented in this report, perhaps using the index derived from the first principal component only on those nodes that score low in the second principal component (low betweenness nodes) would identify nodes that are potentially significant for biochemists, but are also not too important that there removal would ultimately lead to the death of the cell.

3 Discussion

3.1 Evaluation of Methods

This report has focused on applying the statistical method of Principal Component Analysis to protein-protein interaction networks. The key advantage of this method is that it allows mathematicians to analyse the importance of nodes by taking into account many different centrality measures simultaneously as opposed to focusing only on a single measure

at a given time. This is beneficial as it allows mathematicians to easily rank the significance of nodes, without having to debate which different aspects of a node contribute to its importance. By forming a weighted score for nodes, one is able to compare nodes which scored highly in different centrality measures.

Whilst the importance score index derived from the first principal component of the PCA is extremely useful in forming a ranking of nodes using the combined information from many different centrality measures, there one very major limitation. This limitation is related to the context of the problem. Since the aim of this method is to identify biologically important proteins, it would be useful if the score generated from the index had some biological interpretation. Individually, many centrality measures can be interpreted to represent some sort of biological process or function. However, when taking a linear combination of these different measures it is difficult to provide a biological meaning to the obtained result.

When incorporating the second principal component, a significantly larger proportion of the variance can be explained. Essentially this means that when incorporating the second component into the process of predicting the importance of a node, a larger amount of the underlying information can be included. It would be interesting to compare the results obtained when only considering the nodes deemed interesting by the second principal component (low betweenness nodes) to those where only the index score is used. Perhaps the information contained in the second component is useful for identifying nodes that are important but not essential to the regular functioning of yeast.

3.2 Potential Future Research

Potential avenues for future research in this field are abundant. Researchers may wish to consider components beyond the first two, to try and identify further aspects of centrality measures that correspond to the importance of proteins. Additionally, future research may seek to follow the same procedure but include more or different centrality measures in the Principal Component Analysis. This report only investigated some of the more generic measures that had biological relevance supported by literature, however there may be other interesting methods to determine the importance of nodes that could yield significant results.

Another suggested direction where further research could be applied involves finding rotation matrices that could potentially make the vector of centrality measure weights scarce. This would be interesting as the existence of such matrices could imply that the information from some centrality measures could be attributed to or contained in other measures. This report did not consider this concept since the weights determined by the PCA were relatively equal, suggesting that applying such a rotation matrix could be inappropriate. However, other networks could be better suited to such analysis.

4 Conclusion

It has been found that the importance of a node in a network can be summarised by an index containing information from a collection of centrality measures. By applying Principal Component Analysis to a data set containing the centrality scores for centrality measures across different nodes, an index comprised of the weighted sum of different centrality measures can be derived. Furthermore, by considering the second principal component, different underlying trends in the data can be identified, and criteria for node selection to improve the process of identifying significant or interesting nodes can be suggested. This report has explored and analysed the process of applying Principal Component Analysis to a protein-protein interaction network, and has presented methods to apply PCA to identify potenitally important nodes.

5 Appendices

5.1 Appendix 1: GitHub Link

https://github.com/yonoron/MATH3888-Individual

5.2 Appendix 2: Correlation Map

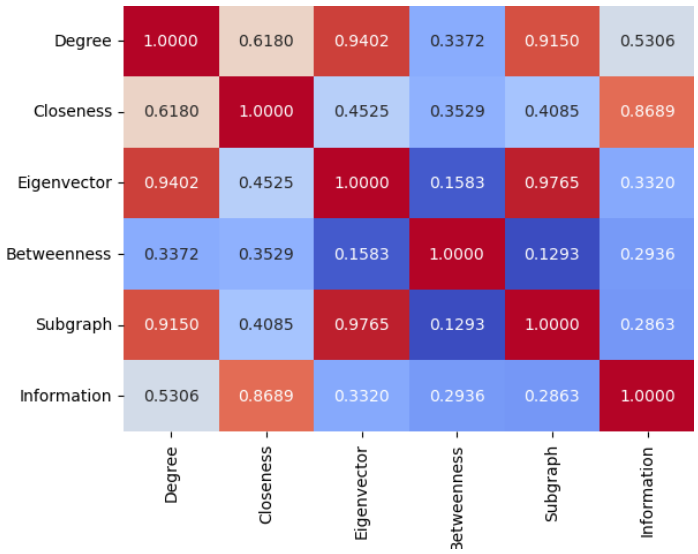

Figure 4: Pairwise Plots of Centrality Measures

From Figure 4 it is noticeable that Eigenvector centrality and Subgraph centrality are essentially interchangeable due to their high correlations. Furthermore, Betweenness centrality can be identified as a potentially interesting metric due to its relatively low correlation to all other centrality measures.

References

[1] F. Bloch, M. O. Jackson, and P. Tebaldi. Centrality measures in networks. *Social Choice and Welfare*, pages 1–41, 2023.

[2] L. F. Bringmann, T. Elmer, S. Epskamp, R. W. Krause, D. Schoch, M. Wichers, J. T. Wigman, and E. Snippe. What do centrality measures measure in psychological networks? *Journal of abnormal psychology*, 128(8):892, 2019.

[3] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5): 056103, 2005.

[4] K. Fitch and N. E. Leonard. Information centrality and optimal leader selection in noisy networks. In *52nd IEEE Conference on Decision and Control*, pages 7510–7515, 2013. doi: 10.1109/CDC.2013.6761082.

[5] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.