Summer Project 2020 Report

During July and August 2020, I researched techniques for clustering points in large signed networks. I was supervised by Dr Mihai Cucuringu, who has cowritten many papers on this subject, and the project gave me the opportunity to get a much better grasp of what cutting edge research in data science can look like. Clustering techniques are briefly introduced in a first year mathematics module and I enjoyed this part of the course, so this project allowed me to delve much deeper into the subject beyond the scope of the course.

Networks are ubiquitous in the world today, and this means that efficient algorithms for analysing networks are of interdisciplinary interest. In a technological setting, the internet has allowed a wide variety of large networks to form, such as networks of members of an online forum, or networks of customers and items bought on an online shopping site. Analysing the structure of these networks is interesting both to academics and to companies who wish to maximise profit. For example, a company might use the information from the network analysis to improve user experience on an online forum or to provide more appropriate targeted adverts to online shoppers. In addition, many interesting networks can be formed by processing time series data. For instance, I was able to perform analysis on graphs from time series data as varied as the level of rainfall across Australia, the value of stocks, and foreign exchange rates. This type of analysis can be extremely useful to climate scientists, investors and economists.

Initially the project was focused around reading introductory papers to understand the principles of clustering in graphs and signed graphs. I learnt that spectral graph theory is a powerful tool for analysing networks – that is, using the eigenvectors and eigenvalues of some matrix associated with the graph to gain structural information about the graph. In particular, the signed Laplacian matrix can be used to generate a useful embedding of a signed graph into a low number of dimensions. Next, I wrote a Python program to generate synthetic datasets (random networks), perform the signed Laplacian embedding and perform a standard clustering algorithm to the points in 2D. After testing it on the synthetic datasets, I ran the program on several real-world datasets that appeared in recent papers.

In the second stage of the project, I explored algorithms for a subtler problem: finding "bipartite motifs" in signed networks. In a signed network, we call a graph bipartite if it can be separated into two sets of vertices with only positive edges inside each set, and only negative edges between the two sets. A bipartite motif is a subset of vertices and connected edges that are mostly bipartite (although in practice, there may be some noise). The successful algorithms that appear in the literature include deterministic and non-deterministic spectral algorithms, and matrix completion algorithms. I also attempted to adapt the signed-Laplacian approach to this context. My approach was successful on certain types of low-noise synthetic datasets, although I had to conclude that this approach was fundamentally unsuited to finding small bipartite motifs in large noisy networks (a type of network that often occurs in the real world).

Undertaking a summer project has been a valuable experience because it has allowed me to explore a subject not directly on the course syllabus whilst gaining a much better understanding of the research process. It also helped me to make better informed module choices for the next year. I decided that I would like to submit another project as an assessed module, something that I had not really considered before doing the summer project. The content of the projects will be fairly distinct, but I will start the second project with confidence that I have already practised highly relevant skills such as reading academic papers and writing mathematics.