Graph Clustering through Spectral Decomposition

Christopher Gaw Gonzalo – christophercgaw@gmail.com
Diamond Bar High School, Class of 2022
USC Viterbi Department of Computer Science, SHINE 2021

Introduction

Graph clustering is the process of grouping nodes based on the connections between the two nodes. In SHINE, I worked on implementing a cutting-edge clustering algorithm to cluster graphs that follow the Stochastic Block Model. In this model, graphs follow a set of patterns. First, the model graph has a number N of vertices. These vertices are then grouped randomly into K groups. Then the model generates edges between these vertices based on the groups. If the two vertices are from the same group, the two vertices are connected by a chance P. Otherwise, they are connected by a chance Q.

Objective & Importance of Professor's Research

Professor Zhang's research focuses on theoretical computer science. His research deals with optimizing algorithms and calculating the time complexity of these algorithms through theoretical analysis. Recently, Professor Zhang has been working on improving Van Vu's clustering algorithm. Van Vu was the first to come up with a simple spectral decomposition algorithm.

Optimizing graph clustering algorithms is important for two main reasons. First, graph clustering is considered an NP-Hard problem which takes a long time to solve using typical algorithms. Second, applications such as YouTube, Netflix, Google, and TikTok can benefit by improving the accuracy and speed of their recommendation systems.

How it Works

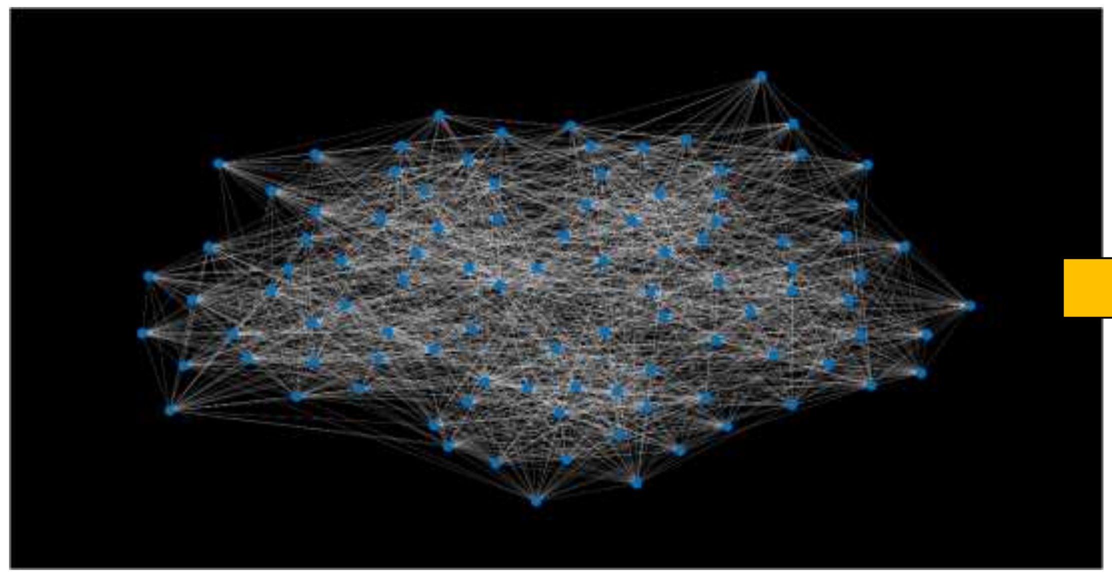


Figure 1: Graph of Raw Data (Not Clustered)
PC: Chris Gaw through MATLAB

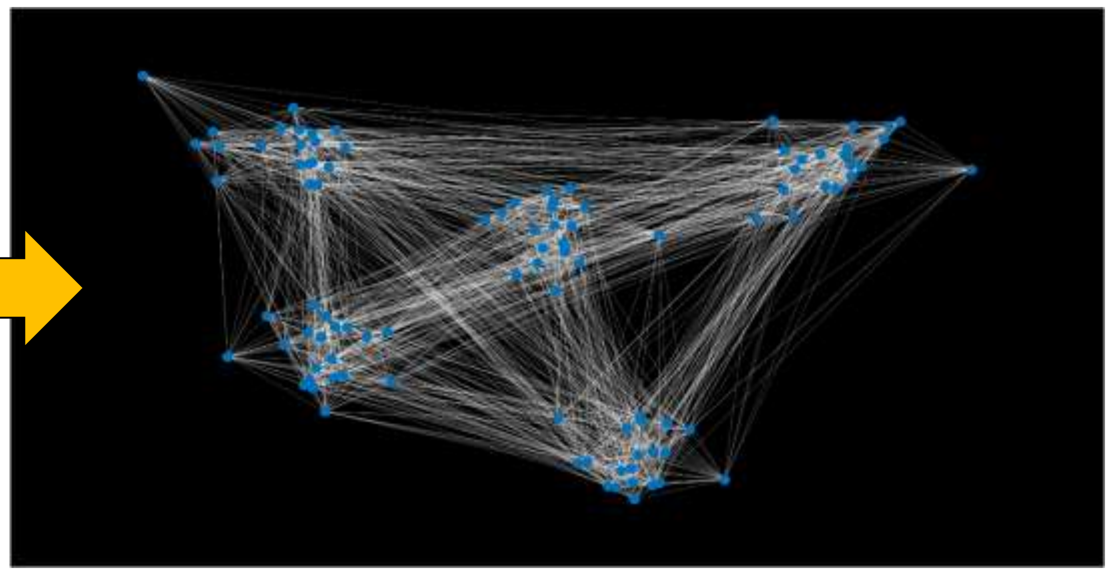


Figure 2: Raw Data projection graph
PC: Chris Gaw through MATLAB

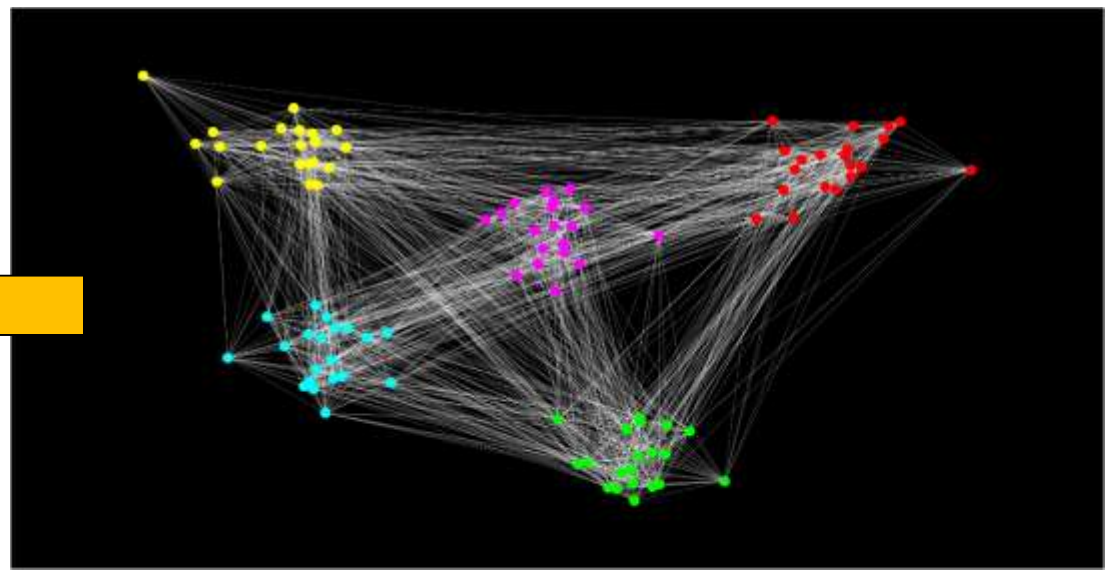


Figure 3: Graph is clustered based on distance
PC: Chris Gaw through MATLAB

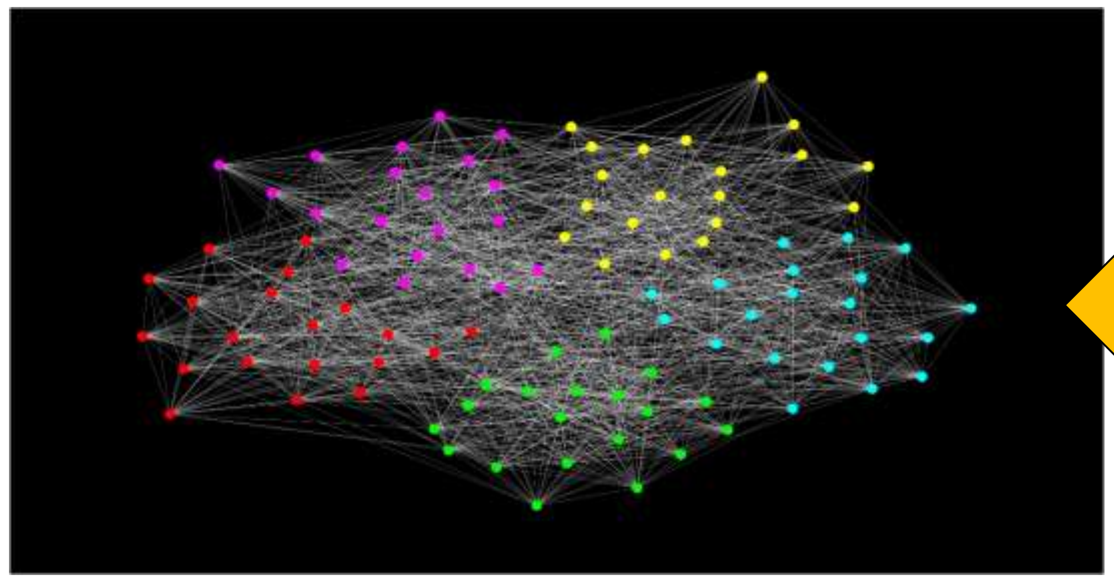
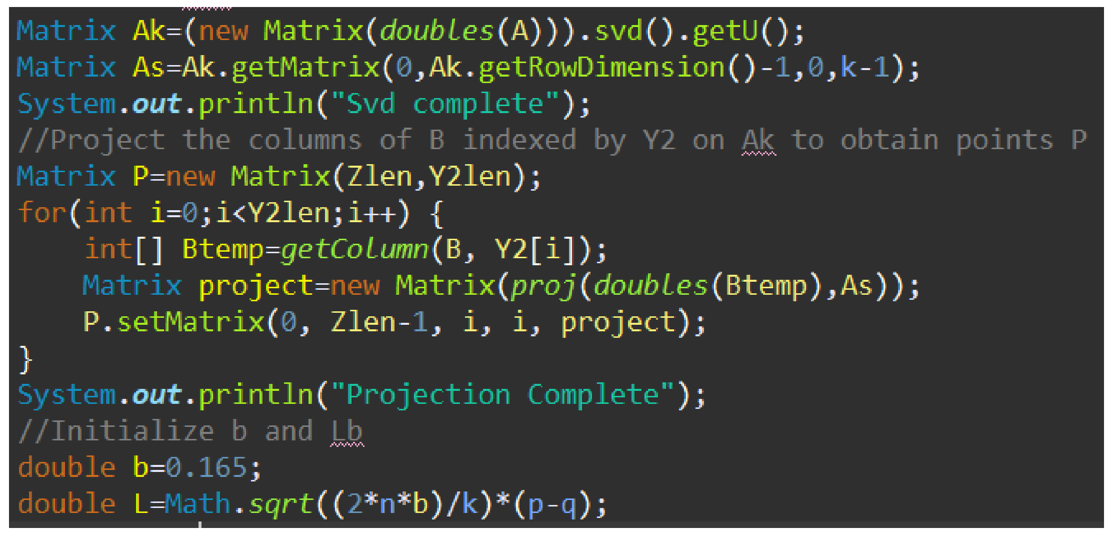


Figure 4: Final Graph representing Clustered Data
PC: Chris Gaw through MATLAB

How This Relates to Your STEM Coursework

SHINE helped me see how essential mathematics is in both understanding algorithms and optimizing them. In particular, SHINE made me realize the importance of linear algebra, set theory, and statistics in data analysis. It also helped me realize how many different mathematical concepts from drastically different branches of mathematics can come together in computer science.



```java
Matrix Ak=(new Matrix(doubles(A))).svd().getU();
Matrix As=Ak.getMatrix(0,Ak.getRowDimension()-1,0,k-1);
System.out.println("Svd complete");
//Project the columns of B indexed by Y2 on Ak to obtain points P
Matrix P=new Matrix(Zlen,Y2len);
for(int i=0;i<Y2len;i++) {
    int[] Btemp=getColumn(B, Y2[i]);
    Matrix project=new Matrix(proj(doubles(Btemp),As));
    P.setMatrix(0, Zlen-1, i, i, project);
}
System.out.println("Projection Complete");
//Initialize b and Lb
double b=0.165;
double L=Math.sqrt((2*n*b)/k)*(p-q);
```

Figure 5: Clustering Algorithm
PC: Chris Gaw through Java

Skills Learned

Throughout SHINE, I learned several skills that will help me in my future research. I learned several new programming languages, including Octave and MATLAB, and improved my abilities in coding Java and Python. I also learned more about Machine Learning and how to successfully implement Machine Learning Algorithms such as linearized regression and graph clustering algorithms. Most importantly, though, I gained insight into the intuition behind how faster algorithms are created and improved.

Next Steps

Joining SHINE and learning directly from Professor Zhang has amplified my interest in theoretical computer science. I will continue to work with Professor Zhang on applying various clustering algorithm on RNA-Sequencing data. To further my knowledge on this subject, I plan on taking subjects relevant to machine learning and major in Computer Science in college.

Acknowledgements

I would like to sincerely thank:

- Prof. Jiapeng Zhang
- Dr. Katie Mills
- Monica Lopez
- SHINE Team
- Cassandra Jeon

for their mentorship and guidance.