CSC 591 GDM Capstone Project Final Report

1.Research question

Our topic is clustering in multiple graph. In our lecture, all the lectures are about how to find the community in a single graph. However, there are many situations that can involve in multiple graph. For instance, a person may have several social network accounts. Nevertheless, the user may well not synchronize all the information among all accounts. Therefore, there will be several different graphs based on this situation. A possible application is recommendation. We can find the potential relationship in different types of graph by clustering and further recommend the user. Another possible application is finding the community on different network. A user may watch several types of video. We can find what other video are similar to them and further broadcast to the user. Overall, we think this problem can be applied in many situations and thus is important.

Summaries of relevant research papers

1. **Community Detection in Multi-Layer Graphs: A Survey**

   Nowadays, people have several relationships among social network. Thus, it is important to analyse multiple interdependent graphs ,where each graph represents an aspect of the relationships. This paper introduced a comprehensive understanding of community detection in multi-layer graphs and compare the algorithms with their properties. The paper provides two methods of community detection in more than two layers. The first one is matrix factorization, which use general clustering method to extract common factors from multiple factors. Another method is pattern mining, which find quasi-cliques that appear on multi-layers with a frequency above a given threshold. Also, the paper concludes seven properties of community detection algorithm. Although there are many algorithms that can detect community in multi-layer graphs, the author points out some challenging in detection.

2. **Clustering on Multi-Layer Graphs via Subspace Analysis on Grassmann Manifolds**

   In this paper, the author provides a framework that can merge the graph and preserved the information in each graph. In the traditional spectral clustering algorithm, it can only calculate the subspace of the graph. The author proposes a novel easy method to combine multiple subspace representations into one representative subspace. First, they calculate the Laplacian matrix of each graph and using the eigenvalue and eigenvector to find the subspace of each graph. This approach is like principal component analysis. Then, they sum up over all the Laplacian matrix and subtract the weighted sum of the subspace matrix. This operation combine all the subspace into one subspace. They calculate the eigenvalue and eigenvector again with the merged Laplacian matrix and find the merged subspace. Finally they use k-means to find the

unified cluster. The paper mentions that there are still study to tune the parameter alpha in their approachment. Though they make several models with different values of alpha, they believe there is no specific alpha is perfect. If they know specific prior knowledge of the data, they can have a better clustering performance since alpha is related to calculate the subspace.

3. **Evaluating accuracy of community detection using the relative normalized mutual information**

   The paper points out that most people use similarity and normalized mutual information to measure the clustering. However, the author indicates that both of the measurement have some problems. For the former, when the graph is large, maximizing overlap is difficult. For the latter, the size of the graph will affect the approxImate probability. The author proposes a new metric for the accuracy of community detection, called the relative normalized mutual information(rNMI). They introduce a null model to reduce the bias. They done the same algorithm but use rNMI and found that rNMI can actually reduce the bias.

Proposed solution

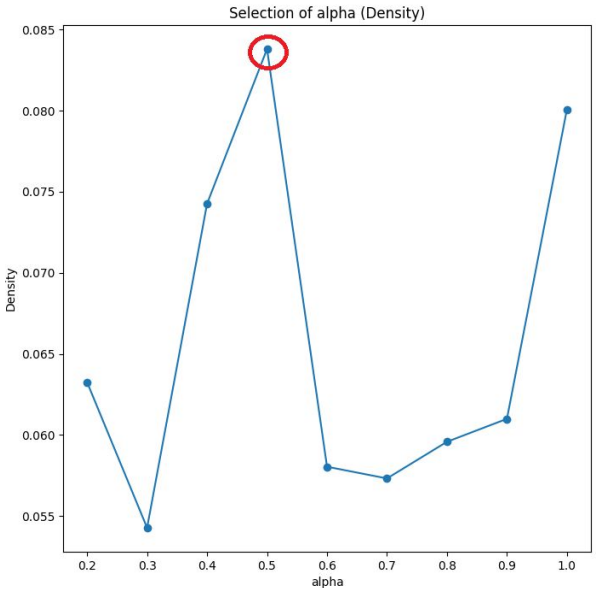
We Implement the algorithm in our second reference paper. We use the dataset called AUCS dataset which clearly represents in figure 4 in the first reference paper. The meaning of the dataset is much like what we proposed. First of all, we seperate all the edges into their corresponding graph. Then, we calculate the normalized Laplacian matrix of each graph. We further use the normalized Laplacian matrix to calculate the eigenvalue and eigenvector in order to extract the subspace of each graph. Then we sum the normalized Laplacian matrix and subtract with weighted sum of all subspace matrix. The purpose of summation is to aggregate the subspace. Then we calculate the subspace of aggregate modified Laplacian matrix by calculating the eigenvalue and eigenvector again. For the last step, we use k-means to find the unified cluster.

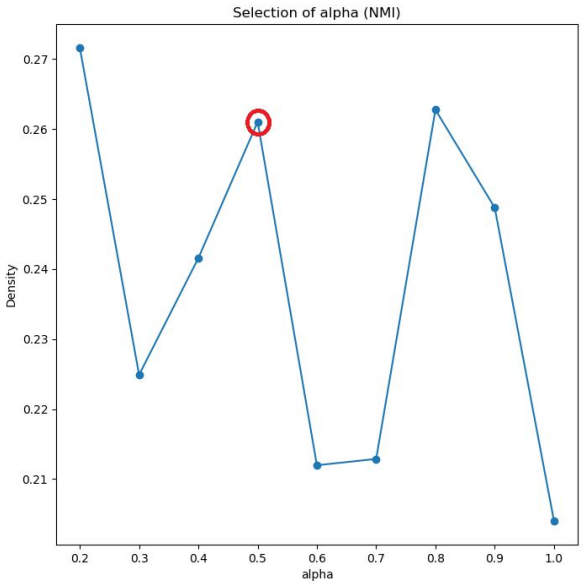
Evaluation of the proposed solution

1. Density: is a great indicator to help us identify the community clustering results. The higher value of density, the more likelihood it presents a good community with strong connections. As the description of subspace analysis on Grassmann Manifolds algorithm, we need to select a regularization parameter alpha to construct a subspace matrix for clustering.
2. Normalized Mutual Information (NMI): is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). It also be used on training for the best alpha selection.

3. Purity: is a simple and transparent evaluation measure.. Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N.

Result:





With the results of selection of regularization parameter alpha, we can observe that when alpha = 0.5. We have the best performance in Density and NMI score. Thus, we will use the parameter to cluster multilayer graph.

Since our ground truth contain 8 different group clusters, we don't need to decide the number of clusters. The following table is the result of clustering community score.

Metrices	Lunch	Facebook	Work	Leisure	Coauthor	Multilayers
Purity	0.4909	0.2909	0.4727	0.4727	0.4	0.3472
NMI	0.40782	0.2371	0.3602	0.3836	0.3053	0.2986

Each column indicates different single layer score and multilayer graph score. We can find that their corresponding subspaces are close to each other on the Grassmann manifold. We can find that the algorithm enforces the solution of the layer 'Facebook' which is relatively lower quality because the information in the representative subspace improve the clustering performance in the result. Therefore, the algorithm is expected to provide complementary but not contradictory information and allow to employ the information of subspace from other layers to improve the performance of clustering in community detection.

Reference:

1. Dong, Xiaowen, et al. "Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds." IEEE Transactions on signal processing 62.4 (2013): 905-918.
2. Kim, Jungeun, and Jae-Gil Lee. "Community detection in multi-layer graphs: A survey." ACM SIGMOD Record 44.3 (2015): 37-48.
3. Zhang, Pan. "Evaluating accuracy of community detection using the relative normalized mutual information." Journal of Statistical Mechanics: Theory and Experiment 2015.11 (2015): P11006.