# Community Detection In Multi-layer Network

Project report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Masters of Technology

in

Computer Science Engineering

**by**

**Prishni Rateria**

**(16CS60R58)**

**Under the supervision of**

**Professor Bivas Mitra**

**Department of Computer Science and Engineering**

**Indian Institute of Technology Kharagpur**

**Autumn Semester, 2017-18**

**November 2017**

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## KHARAGPUR - 721302, INDIA

## *CERTIFICATE*

This is to certify that the project report entitled "**Community Detection In Multi-layer Network**" submitted by **Prishni Rateria** (Roll No. 16CS60R58) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Masters of Technology in Computer Science Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2017-18.

Date: November 2017

Place: Kharagpur

Professor Bivas Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

Multiple aspects of relationships can be represented by a multi-layer graph comprised of multiple interdependent graphs, where each graph represents an aspect of the relationships. Therefore, great efforts have been made to solve the challenge of community detection in multi-layer graphs. [2]

Communities are defined as groups of nodes that are more densely connected to each other than to the rest of the network. The goal of the community detection algorithms, consequently, is to partition the networks into groups of nodes; large body of work exists on community detection in single and isolated network. Recently, many real networks, including communication, social, infrastructural and biological ones, are often represented as multilayer networks. A multilayer network is comprised of multiple interdependent networks, where each network layer represents one aspect of interaction. Moreover, the functionality of a node in one network layer is dependent on the role of nodes in other layers.

For instance, a location based social networks (say, Yelp) can be represented as a multilayer network (see Fig. 1) where in one layer customers (visitors) are connected via social links and in the other layer location nodes are connected through proximity links. The (coupling) link connecting a customer with a location node represents the visit of a customer to a location.

Community detection in complex multilayer networks is an important research problem. The communities in multilayer networks help to identify functionally cohesive sub-units and reveal complex interactions between multi-type nodes and heterogeneous links.Community detection in multilayer network is challenging as the detected communities have possibility to contain only single or multiple types of nodes.

## 1.2 Current Status and Challenges

**Community Detection in Single Layer Graphs** Many community detection approaches have been proposed for single-layer graphs. Fortunato and Schaeer [2] conducted really extensive survey on this topic. Representative algorithms include graph partitioning algorithms, modularity-based algorithms, spectral algorithms, and structure definition algorithms. The objective of graph partitioning algorithms is to divide the vertices such that cut size is minimal. Cut size is determined by the number of edges lying between partitions. The goal of modularity-based algorithms is to partition the vertices such that modularity is maximal. Modularity is defined by the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. Spectral algorithms partition the graph into communities using the eigenvectors of graph matrices. A graph Laplacian matrix is typically used for the graph matrix. Structure definition algorithms discover communities such that a very strict structural property is satisfied. In other words, they find communities satisfying the meta definitions of a community such as k-clique, r-quasi clique, and s-plex

**Challenges for Multi-Layer Graphs** In contrast to the community detection problem in single graphs, new challenges arise for community detection in multi-layer graphs. Intuitively, each single layer has a piece of meaningful information from its own perspective; however, one can expect improved community detection results through the proper and efficient merging of information in each layer. Thus, an important open question is how to exploit and fuse the multiple aspects of information to generate improved understanding of vertices and their relationships. In addition, since we are confronted with managing multiple layers (often called networks of networks), scalability remains a significant challenge because of the larger resulting search spaces [2].

## 1.3 Objective

Detect both multi-layer and single-layer communities in a multi-layer network, which comprises of

- single type of nodes
- Multiple type of nodes

# Chapter 2

# Dataset

## 2.1 Synthetic Dataset Generation

In this section, we propose a methodology to generate benchmark multilayer networks with ground truth communities. The parameter $\alpha$ regulates the proportion of cross layer vs single layer communities in the benchmark. The network contains M number of different layers where each layer $L_i$ contains $N_i$ nodes with average degree $k_i$. The methodology contains the following three steps:

**Step 1. Single layer communities:** First, we apply the LFR benchmark algorithm to generate communities at each layer $L_i$ with $N_i$ nodes where both degree and community size distributions follow power law distribution with exponents $\gamma_i$ and $\beta_i$ respectively. We fix the mixing parameter as $\mu_i$ to construct $C_i$ single layer communities in layer $L_i$ .

**Step 2. Cross layer communities:** We combine the community $x_i \in C_i$ of layer $L_i$ with community $x_j \in C_j$ of layer $L_j$ to create the cross layer community $x_{ij}$.

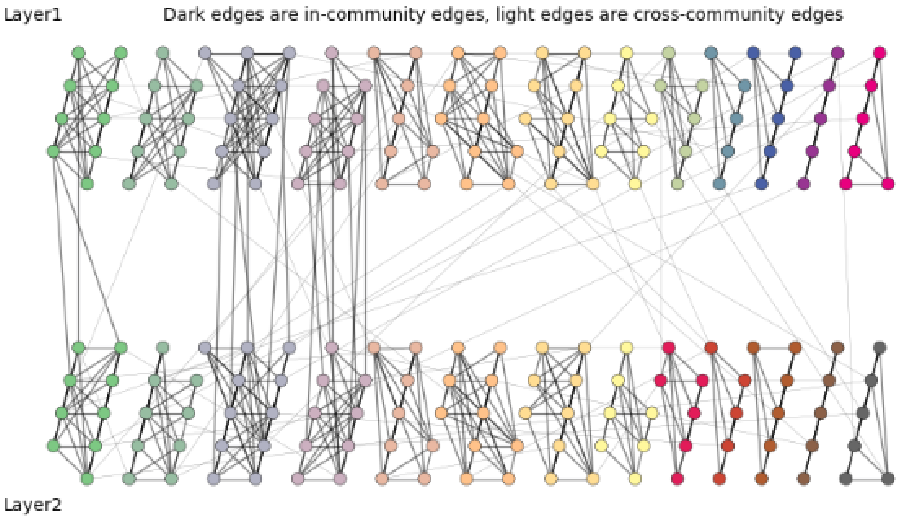

Layer1
Dark edges are in-community edges, light edges are cross-community edges
Layer2

Figure 2.1: An example multi layer graph

Assuming $|C_i|$ and $|C_j|$ as the number of communities in layers $L_i$ and $L_j$ respectively, $|C_c| = \min\{|C_i| \ , \ |C_j|\}$ denotes the maximum possible number of cross layer communities. We construct $(|C_c| \times \alpha)$ cross layer communities by randomly combining single-layer communities from both the layers $L_i$ & $L_j$ respectively; notably each cross layer community $x_{ij}$ may contain one or multiple single layer communities.

**Step 3. Coupling links:** Finally, we create the coupling links between the layers $L_i$ and $L_j$ with density $d_{ij}$ . Fraction p denotes the mixing parameter for the cross layer communities. We first distribute ($N_i$ x $N_j$ x $d_{ij}$ ) coupling links randomly between the layers $L_i$ and $L_j$ where each link has one end in $L_i$ and another end in $L_j$ . Next, we rewire the coupling links such that p fraction of links stay inside the cross layer communities and the remaining $1 - p$ fraction of links connect different cross layer communities.

In summary, we have the following four parameters we can vary to generate different networks:

1. Noise parameter of the LFR algorithm, $\mu$

2. Proportion of multilayer communities, $\alpha$

3. Mixing parameter of multilayer links, $p$

4. Density of cross layer links, $d$

# Chapter 3

# Related Work

The study of community detection in homogeneous single-relational networks or called unipartite networks has a long history. In the past decade, this study has attracted a great deal of interest and various methods were proposed. In particular, a family of methods which are widely used is known as modularity optimization. Modularity was originally proposed by Newman and Girvan [5] [6] for evaluating the "goodness" of a partition of a unipartite network into communities. The definition of modularity involves a comparison of the fraction of intra-community edges in the observed network minus the expected value of that fraction in a randomized network, which is called the null model. More precisely, the mathematical expression of modularity in an undirected single-relational network reads:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \tag{3.1}$$

where,

- $A_{ij} = 1$, if there's an and edge between nodes $i$ and $j$ else 0

- $m$ = Total number of edges
- $k_i$ = Degree of node $i$

There are studies on community detection in heterogeneous single-relational networks. The most simple model of such networks is the bipartite network, where there are two types of nodes and edges that connect nodes of different types. Examples of bipartite networks include author-paper networks, actor-movie networks, and customer-product networks. A direct generalization of bipartite network is the k-partite network, where there are k types of nodes and hyper-edges that connect k nodes of different types. Researchers extended modularity to bipartite networks and k-partite networks.

## 3.1 MetaFac [4]

This algorithm detects communities based on the tensor factorization and requires the number of communities to be specified *apriori*[1]. It detects communities in a way such that each of them contains at least one node from every layer (i.e. only cross layer communities).

## 3.2 CompMod [3]

This algorithm detects communities by maximizing composite modularity, which is a combination of the modularities of different subnetworks. We compute normalized mutual information (NMI) index to compare the detected communities with the ground truth communities. NMI is a measure of similarity of communities, which attains a high value if the ground truth and detected communities exhibit good agreement.

## 3.3 Discovering Community Structure in Multilayer Networks (DSAA)

Proposes a new modularity index ($Q_M$)) for evaluating the communities in multilayer networks and leverages on the single layer community detection algorithms Girvan-Newman [6] and Louvain [5] which detect communities by maximizing modularity. The authors then substitute the Girvan-Newman modularity by proposed modularity index $Q_M$ and develop GN-$Q_M$ and Louvain-$Q_M$. The modularity expression proposed is as follows:

**For single layer community** : In any single layer community C, all the constituent nodes belong to either $L_1$ or $L_2$. Hence, for each single layer community C modularity is computed as:

$$Q_M^C = \forall i,j \in C[\frac{1}{3}\{\frac{1}{2|E_1|}\sum_{i,j\in V_1}(A_{ij} - \frac{h_i * h_j}{2|E_1|})\}] \tag{3.2}$$

**For multi layer community** : Any cross layer community C is composed of three sub modules - two intra layer and one inter layer

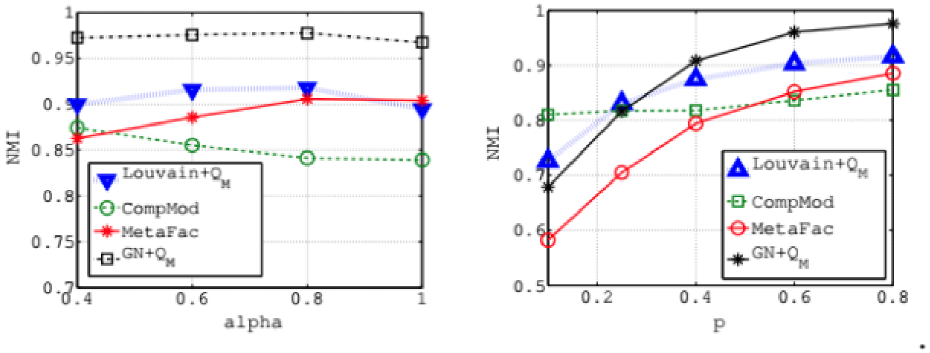
$$Q_M^C = \forall i,j \in C[\frac{1}{3}\frac{1}{2|E_1|}\sum_{i,j\in V_1}(A_{ij} - \frac{h_i * h_j}{2|E_1|})+ \tag{3.3}$$

$$\frac{1}{2|E_1|+2|E_2|+|E_{12}|}\sum_{i\in V_1, j\in V_2}(A_{ij} - \frac{c_i' * c_j'}{2|E_1|+2|E_2|+|E_{12}|})+ \tag{3.4}$$

$$\frac{1}{2|E_2|}\sum_{i,j\in V_2}(A_{ij} - \frac{h_i * h_j}{2|E_2|})] \tag{3.5}$$

where ,

- for any node i, $c_i' = c_i'$ if $c_i > 0$ and $c_i' = h_i$ otherwise.



(a) Varying $\alpha$ values for $p = 0.8$, $\mu = 0.4$ and $d = 0.04$

(b) Varying $p$ values for $\mu = 0.4$, $\alpha = 0.6$ and $d = 0.04$

FIGURE 3.1: NMI of obtained and ground truth communities for various $\alpha$ and $p$ values

- $A_{ij} = 1$, if there's an and edge between nodes $i$ and $j$ else 0
- $h_i$ = intra-layer degree
- $|E_1|$ = Total number of edges in $L_1$
- $|E_2|$ = Total number of edges in $L_2$
- $|E_{12}|$ = Total number of inter-layer edges

While Figure3.1 shows that the method proposed in DSAA performs better than CompMod and MetaFac, it has some limitations.

## 3.4 Limitations of DSAA

Optimizing modularity is proved to be NP-hard [1]. Researchers have developed various heuristic optimization algorithms, of which the the Louvain algorithm [5] is widely used, since it reaches a proper balance between accuracy and speed.

The authors considered Louvain algortihm,i.e. single layer community detection algorithm sufficient to work with multi-layer networks.

### 3.4.1 Louvain Algorithm

The value to be optimized is modularity, defined as a value between -1 and 1 that measures the density of links inside communities compared to links between communities.[2] For a weighted graph, modularity is defined as:

$$mQ = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \tag{3.6}$$

where,

- $A_{ij}$ represents the edge weight between nodes i and j;
- $k_i$ and $k_j$ are the sum of the weights of the edges attached to nodes i and j, respectively;
- $m$ is the sum of all of the edge weights in the graph;
- $c_i$ and $c_j$ are the communities of the nodes; and
- $\delta$ is a simple delta function.

In order to maximize this value efficiently, the Louvain Method has two phases that are repeated iteratively.

First, each node in the network is assigned to its own community. Then for each node i, the maximum gain in modularity is calculated for removing i from its own community and moving it into the community of each neighbor j of i.

In the second phase of the algorithm, it groups all of the nodes in the same community and builds a new network where nodes are the communities from the previous phase. Any links between nodes of the same community are now represented by self loops on the new community node and links from multiple nodes in the same community to a node in a different community are represented by weighted edges between communities. Once the new network is created, the second phase has ended and the first phase can be re-applied to the new network.

For the algorithm to be applicable on multi-layer networks the single modularity index $mQ$ was replaced by proposed multi-layer modularity index $Q_M$.

This algorithm modifies the network structure after every iteration by collapsing all nodes in one community to form one node in the modified network structure. So, after the first iteration the resultant network looses all its layer information and multi-layer structure i.e. after one iteration the network reduces to be a single-layer network.

To overcome this limitation of DSAA algorithm we tried to modify this algorithm which would preserve the layer structure of the network at each step.

# Chapter 4

# Our Contribution

## MixMod Algorithm [7]

With $Q_M$ modularity as the optimal function, the Louvain strategy is adopted to search for an optimal partition. The procedure of MixMod method is illustrated in Figure 4.1. Most of MixMod is similar to the process of Louvain method, except a little change in the second phase.

At first, each node network is assigned to individual community. So, the initial partition consists of communities as many as the number of nodes.

- In the first phase, we calculate the gain of $Q_M$ modularity by removing node i from its module and by placing i in the module of its neighbor j. The node i is then placed in the module for which the gain is maximum and positive. If all gains are less than zero, we keep node i in its original module. Such process is performed repeatedly and sequentially for all nodes until no positive gain is achieved.

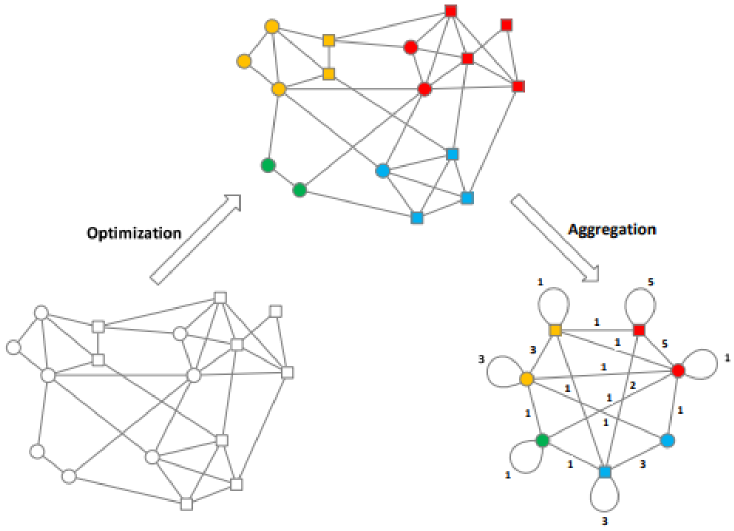


FIGURE 4.1: Main steps of Louvain optimization for MixMod method. This illustration is modified based on the original process of Louvain method

- In the second phase, a new network is constructed according to the partition in the first phase that reaches the local maxima of $Q_M$ modularity. If a module in previous partition includes nodes of both layer 1 and layer 2, two diffrent nodes are added into the new network to represent $L_1$ nodes and $L_2$ nodes in the module respectively. Thus these two nodes belong to a same module (Figure 4.1). The weights of links in the new network are given by the sum of weights of links between two sets of nodes in the original network. It is obvious that this new network is also multi layer. We use this new network as input and then repeat previous two phases until the mixed modularity is maximum. In the end, mixed modules could be identified corresponding to the final partition. [7]

We implemented the new algorithm using

1. The modularity definition for multilayer networks described in Equ.3.2, Equ.3.3 , and

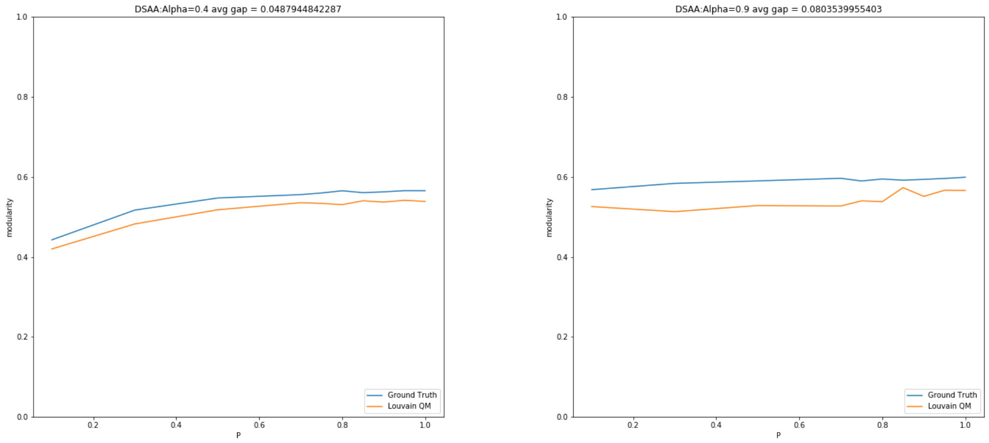


FIGURE 4.2: Modularity plot using Louvain $Q_M$ algorithm for two different values of $\alpha$ (0.4,0.9) and varying P

2. The mixmod algorithm for modularity maximisation in multilayer networks proposed in [7]

and evaluated the performance on our synthetic networks. We also implemented the algorithm proposed in DSAA paper to compare with our results.

## 4.1 Experiments and Evaluation

We evaluated the algoithm on 2-layer synthetic networks with 100 nodes in each layer, generated with maximum degree $k^i_{max} = 10$, average degree $= 6$ and coupling link density, $d = 0.05$.

We vary the synthetic network parameters $p$ and $\alpha$ and use our implemented algorithm to detect communities. We then compute the $Q_M$ value of the detected community structure and compare it with the corresponding value of ground truth community structure.

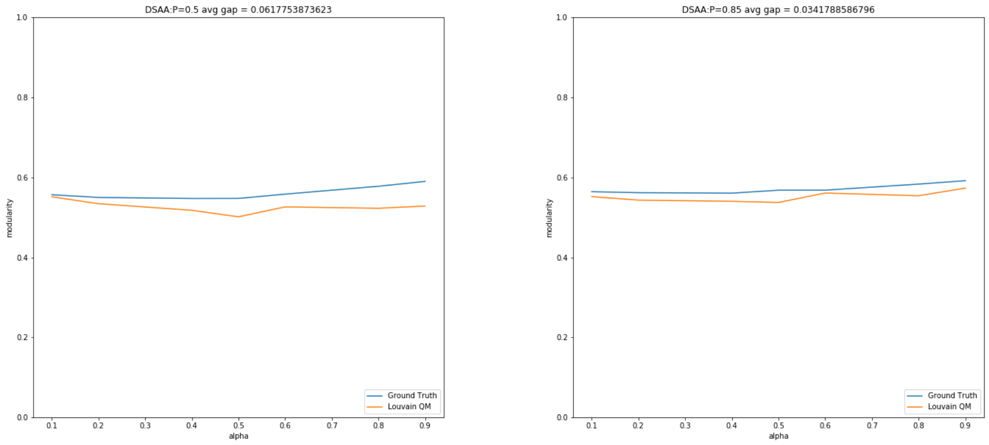


FIGURE 4.3: Modularity plot using Louvain $Q_M$ algorithm for two different values of p (0.85,0.5) and varying $\alpha$

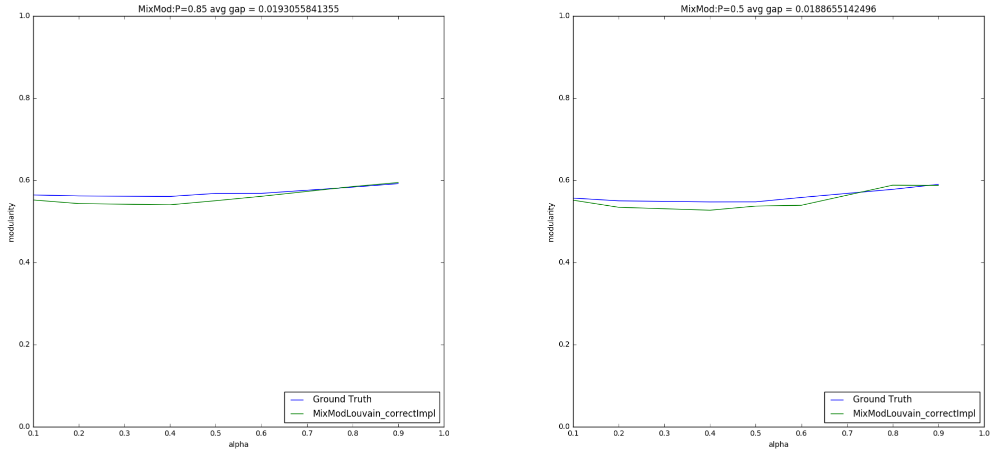


FIGURE 4.4: Modularity plot using Mixmod algorithm for two different values of p (0.85,0.5) and varying $\alpha$

Figure 4.2 and figure4.3 shows the plot for modularity values of communities detected by Louvain $Q_M$ algorithm and the ground truth communities. The plot shows that there is some gap between the detected communities modularity and the groung truth modularity.

After applying the mixmod algorithm with modularity $Q_M$ we plot the above mentioned graphs again (figure 4.4 & figure 4.5) and see a reduction in the gap between

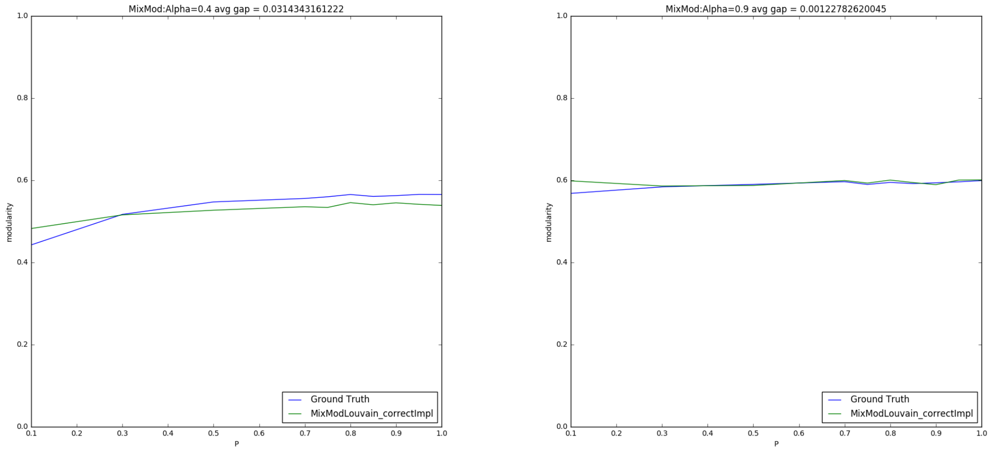


Figure 4.5: Modularity plot using Mixmod algorithm for two different values of $\alpha$ (0.4,0.9) and varying P

the detected communit's modularity and ground truth modularity. Applying mixmod algorithm with the multi layer modualrity index gives a 63.5% improvement in the results.

**Further Analysis** In Figure 4.4 we see that the modularity for the communities detected by our algorithm is lesser than that of ground truth communities for low and moderate values of $\alpha$.

The plot also shows that the algorithm performs well for high values of $\alpha$ which means the algorithm performs well when maximum communities in the ground truth are multi-layer.

Similar behaviour is seen when we plot these graphs for two different values of $\alpha$ (0.4, 0.9) and varying P (Figure 4.5). The algorithm performs poorly for $\alpha = 0.4$ and performs well for $\alpha = 0.9$

From the above observations we conclude that the multi-layer modularity measure $Q_M$ described in Equ.3.2, Equ.3.3 is kind of biased towards detecting multi-layer communities.So, it have a scope of improvement.

## 4.2 Scope of improvement in $Q_M$

The weight of the coupling term 3.3 in the formula is very high. which tries to make every single layer community also as multi-layer in order to maximize the modularity.

$$CouplingtermQ_M = \frac{1}{2|E_1| + 2|E_2| + |E_{12}|} \sum_{i \in V_1, j \in V_2} (A_{ij} - \frac{c'_i * c'_j}{2|E_1| + 2|E_2| + |E_{12}|}) \tag{4.1}$$

where ,

- for any node i, $c'_i = c'_i$ if $c_i > 0$ and $c'_i = h_i$ otherwise.
- $A_{ij} = 1$, if there's an and edge between nodes $i$ and $j$ else 0
- $|E_1| =$ Total number of edges in $L_1$
- $|E_2| =$ Total number of edges in $L_2$
- $|E_{12}| =$ Total number of inter-layer edges

The penalty term in the formula ti too low which overall inflates the value of coupling term.

# Chapter 5

# Conclusion

The algorithm mixmod [7] performs better when used with multi-layer modularity index $Q_M$ than Louvain$Q_M$. And there is some scope of improvement in the modularity index $Q_M$. 3.3 3.2

# Bibliography

[1] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. *On modularity-np-completeness and beyond*. Univ., Fak. für Informatik, Bibliothek, 2006.

[2] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[3] Zhana Kuncheva and Giovanni Montana. Community detection in multiplex networks using locally adaptive random walks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1308–1315. ACM, 2015.

[4] Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 527–536. ACM, 2009.

[5] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[6] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[7] Jianglong Song, Shihuan Tang, Xi Liu, Yibo Gao, Hongjun Yang, and Peng Lu. A modularity-based method reveals mixed modules from chemical-gene heterogeneous network. *PloS one*, 10(4):e0125585, 2015.