



The Iby and Aladar Fleischman
Faculty of Engineering
Tel Aviv University

הפקולטה להנדסה
ע"ש איבי ואלדר פליישימן
אוניברסיטת תל אביב



COMMUNITY DETECTION WITH APPLICATIONS TO MULTIREFERENCE ALIGNMENT

Project Number: 20.1.1.2135

Final report

Written by

Alexey Kiryushkin 324439876

Tomer Matityahu 312116668

Under the instruction of

Mr. Noam Janco Tel-Aviv University

Dr. Tamir Bendory Tel-Aviv University

Project was done from home

TABLE OF CONTENTS

Abstract	2
List of Figures	3
List of Tables	4
List of Abbreviations	5
1 Introduction	6
2 Background	7
2.1 Heterogeneous 1D Multireference Alignment model	7
2.2 Community detection	8
2.3 Performance evaluation	10
2.4 K-means algorithm	11
3 Results	12
3.1 Process overview	12
3.2 Community Detection best performers selection	12
4 Discussion	14
4.1 A subheading	14
5 Conclusion	15
5.1 A subheading	15
6 References	16
7 Appendix	18

ABSTRACT

Single-particle reconstruction in Cryogenic Electron Microscopy (cryo-EM)[2] is a tool for constructing a 3D model of a biological macromolecule using 2D projections of the macromolecules taken by an electron microscope. An unsupervised classification of the 2D images is required in order to separate macromolecular projections of different conformations. Due to high noise levels and data heterogeneity, sophisticated clustering methods are needed.

In our project we employ Community Detection (CD) algorithms to cluster data generated from the Multireference Alignment (MRA) statistical model, the model abstracts away much of the intricacy of cryo-EM while retaining some of its essential features. **Fig. 1** shows a diagram of the process.

In addition to the implementation of CD algorithms on the MRA data, we use the K-means clustering algorithm to perform the same task, and compare the performance of both methods. In the last section of this document we discuss the differences between the methods and the advantages and disadvantages of the method we propose.

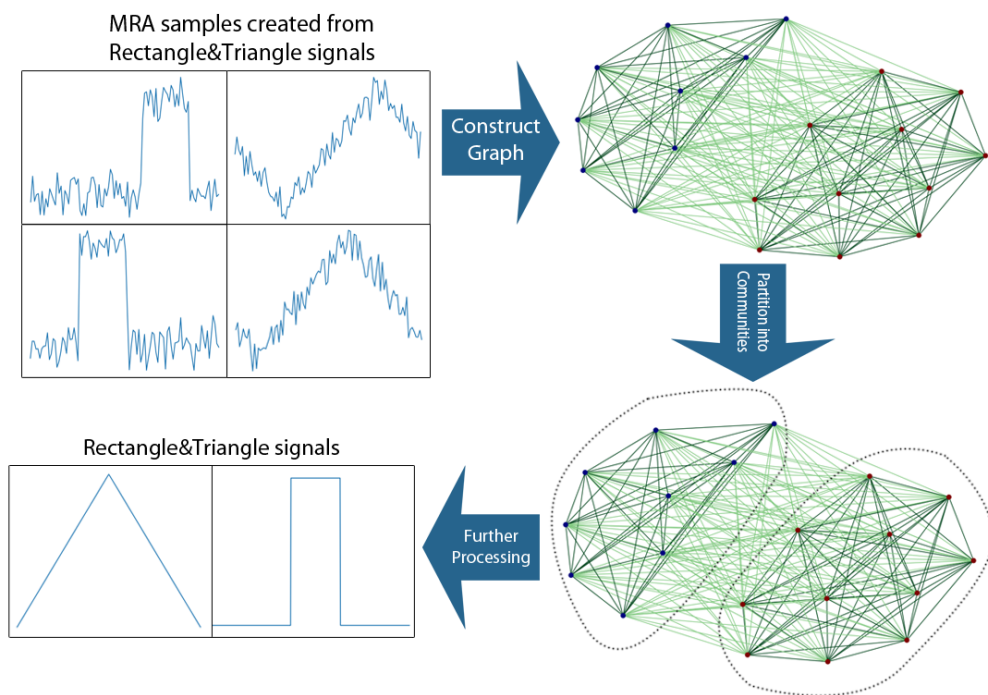


Figure 1: Project process. Further processing stage presents the idea behind clustering the data and is outside of the scope of the project.

List of Figures

1	Project process. Further processing stage presents the idea behind clustering the data and is outside of the scope of the project.	2
2	Example of MRA observations at different SNR levels. Each column shows a shifted distinct signal at different noise levels. We can see that for low SNR the task of signal estimation is quite challenging.	7
3	Graph partition by Community Detection. In this example for the sake of simplicity an unweighed graph is used. Note that in our project we used weighted graphs.	8
4	Full project process.	12
5	Community Detection algorithms performance evaluation. In both graphs, for each SNR value ten different random graphs were evaluated and the results were averaged.	13

List of Tables

1	CD algorithms time complexities. $ E $ represents the total number of edges in the graph and $ V $ the total number of nodes (vertices).	10
---	---	----

List of Abbreviations

C

CD Community Detection. 3, 4, 7, 9, 11

cryo-EM Cryogenic Electron Microscopy. 3, 7, 8

F

FFT Fast Fourier Transform. 8

M

MI Mutual Information. 11

MRA Multireference Alignment. 1, 3, 4, 7–9

N

NMI Normalized Mutual Information. 11

S

SNR Signal To Noise Ratio. 4, 7, 8

1 Introduction

Single-particle reconstruction in cryo-EM is a powerful image-processing tool used to determine the 3D structure of biological macromolecular complexes. 2D images (micrographs) of a macromolecule are taken by an electron microscope. Essentially, the set of all micrographs for a given macromolecule spans a 3D model of the macromolecule. Thus, single-particle reconstruction is using the micrographs to build a 3D model of the macromolecule.

Due to high sensitivity of the biological macromolecules to radiation damage, electron microscope provides limited electron doses when producing micrographs. This and other factors like low contrast of micrographs and digitalization of the images result in cryo-EM data having very low Signal To Noise Ratio (SNR)[2].

The cryo-EM technology has the potential to offer the ability to analyze different functional and conformational states of macromolecules, an important ability for the field of molecular biology. Practically, it entails the classification of heterogeneous cryo-EM data.

Many different approaches for cryo-EM data classification have been developed. Typically, likelihood optimization algorithms and Bayesian inference frameworks are used to deal with data heterogeneity[21, 20, 19, 22, 5].

In our project we propose a different approach for cryo-EM data classification using Community Detection (CD) algorithms that are common in the field of complex networks. According to our approach, data will be classified following the steps:

- Converting data into a weighted graph, where each node corresponds to a sample and the edges between nodes represent the degree of similarity between samples.
- Applying Community Detection algorithms to partition the graph into distinct communities
- Each community represent a single conformation of a sampled macromolecule.

For the sake of an abstraction of the cryo-EM data we will use the Heterogeneous Multireference Alignment (MRA) statistical model. Throughout our project we use the simplified 1D version of the model.

2 Background

2.1 Heterogeneous 1D Multireference Alignment model

In our project we will use the Heterogeneous 1D Multireference Alignment statistical model[4] (MRA shortly) for the sake of cryo-EM data abstraction. Bellow is the definition of the model.

Let $x_1, \dots, x_K \in \mathbb{R}^L$ be K unknown normalized signals (distinct even up to shift) and let R_s be the cyclic shift operator: $(R_s x)[n] = x[\langle n - s \rangle_L]$. We are given N observations:

$$y_j = R_{s_j} x_{k_j} + \varepsilon_j, \quad j = 1, \dots, N \quad (1)$$

where $s_j \sim U[0, L - 1]$, $k_j \sim U[0, K - 1]$ and $\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I)$ is i.i.d white Gaussian noise. Our goal is to estimate the signals x_1, \dots, x_K from the observations.

Simply speaking, MRA observation is a randomly chosen signal x_{k_j} , shifted randomly by s_j and distorted using white noise. **Fig. 2** shows an example of two MRA observations at different noise levels.

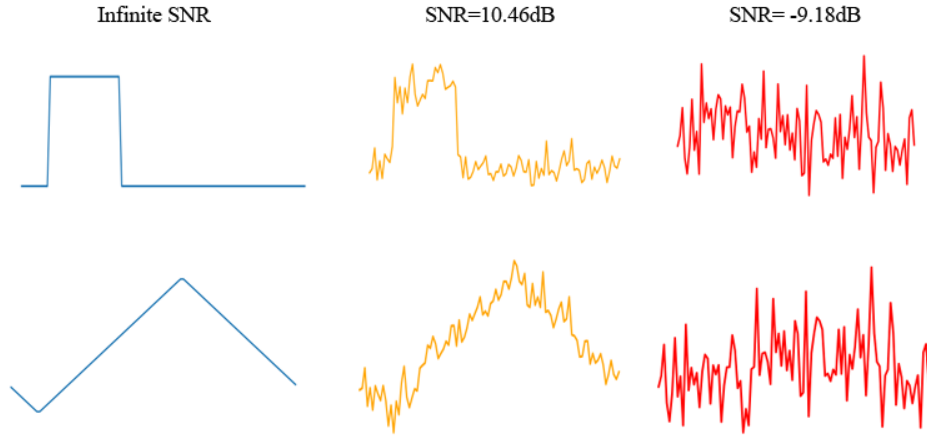


Figure 2: Example of MRA observations at different SNR levels. Each column shows a shifted distinct signal at different noise levels. We can see that for low SNR the task of signal estimation is quite challenging.

In order to build a weighted graph from MRA observations, a similarity measure between observations must be defined. We use the cross-correlation, as its invariance to shift holds great value:

$$(x \star y)_n \triangleq \sum_{l=-\infty}^{\infty} x_l^* (y)_{n+l} \quad (2)$$

The convolution theorem states that the convolution of two signals equals to the inverse Fourier transform of the product of the Fourier transforms of each signal. Thus we can write the equation above in terms of Fourier transforms and later exploit FFT in our simulations.

$$(x \star y)_n = \mathcal{F}^{-1}\{X^* \cdot Y\}_n \quad (3)$$

It should be noted that in order to obtain normalized cross-correlation, MRA samples must first be normalized.

2.2 Community detection

As part of our suggested solution to the heterogeneity problem, we convert the MRA data into a weighted graph. Each node in the graph represents a different MRA observation, as was defined in **Eq. 1**. Each edge connecting a pair of nodes has a weight that represents a similarity measure between the nodes, as was defined in **Eq. 3**.

Our objective is the partition of all MRA observations into K non overlapping classes, where each class represents a distinct signal before it was distorted by **Eq. 1**. Our classification process is carried out using Community Detection (CD) algorithms applied on the produced weighted graph.

Community, in a broad sense, is a set of nodes, which are similar to each other and dissimilar from the rest of the network. Community Detection (CD) algorithms aim to find these communities in a graph. **Fig. 3** shows the result of a CD algorithm.

Finding communities in a graph carries a great value to revealing hidden patterns in data and has many use-cases, such as social behaviour prediction [25] or medical diagnosis[10]. Evidently, Community Detection algorithms are well studied and widely used. In our project we study a group of state-of-the-art CD algorithms and apply them on MRA data.

It is important to note that CD problem is known to be NP-hard[8], meaning that the most efficient CD algorithm that partitions a graph correctly has an exponential time complexity (that is, unless $P = NP$), making it infeasible to run on large graphs. For this reason, it is common to use *heuristic* algorithms that provide approximate solution to the CD problem, with the advantage of polynomial time complexity.

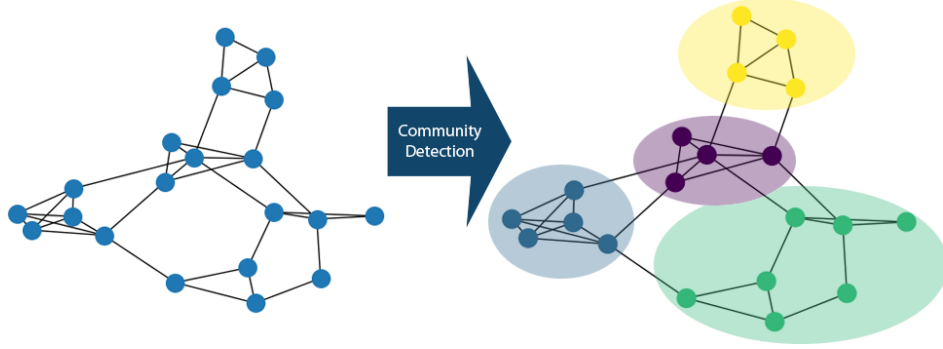


Figure 3: Graph partition by Community Detection. In this example for the sake of simplicity an unweighed graph is used. Note that in our project we used weighted graphs.

Below is a summary of the algorithms we studied.

- *Edge Betweenness*[9] uses the "edge betweenness" as its optimization parameter, which is defined as the number of shortest paths that go through an edge in a graph. Following this definition, edges with higher "betweenness" can be interpreted as connections between different communities. The algorithm computes the "edge betweenness" score for each edge in the graph and iteratively removes the edges with the highest score, while computing the score for each iteration. The end product is a disconnected graph separated into separate communities.
- *Fast Greedy*[13] is based on the idea of *modularity*. Let e_{ij} be the fraction of edges in a graph that connect nodes in group i to those in group j , and let $a_i = \sum_j e_{ij}$. Then modularity is defined as:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (4)$$

Simply speaking, modularity is the fraction of edges that fall within communities minus the expected value of the same quantity if edges fall at random. Thus, low modularity values indicate a random structure of the graph, while higher values point to deviation from randomness and more defined communities within a graph. The algorithm operates by initializing a community for each node in a graph. Then it repeatedly joins nodes in pairs such that the modularity of the graph is maximized.

- *Label Propagation*[16] exploits the basic notion of a community as a set of similar neighbouring nodes. Each node starts with a unique label, and after each iteration, a node updates its label to the most common label of

its neighbours. The algorithm stops when each node has a label that the maximum number of its neighbours have.

- *Leading Eigenvector*[14] is based on the *spectral clustering*. The basic idea behind spectral clustering is to convert the graph into a *similarity matrix* and find the eigenvectors that correspond to the (second) smallest eigenvalue of the matrix. These eigenvectors define the *minimum cut size* of the graph (cut size is defined as the number of edges running between two groups of nodes). The second smallest eigenvalue is chosen because the smallest correspond to minimum cut size of 0, where the whole graph is a single community. Let \mathbf{A} be the *adjacency matrix*:

$$A_{ij} = \begin{cases} 1 & \text{if nodes } (i, j) \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

and \mathbf{P} , such that P_{ij} is the expected number of edges between nodes i, j given a random graph under the constraint that the expected *degree* (number of edges connected to a node) of each node equal to the degree of each corresponding node in the input graph. *Modularity matrix* is then defined as:

$$\mathbf{B} = \mathbf{A} - \mathbf{P} \quad (6)$$

Leading eigenvector algorithm uses the modularity matrix as the similarity matrix and proceeds to perform a spectral clustering of the input graph.

- *Walktrap*[15] is based on the idea that short random walks on a graph tend to stay in the same densely connected area. Each node starts in its own community and a random walk process is run on each node. Based on the random walk, distances between nodes are calculated, and nodes with the smallest distances are lumped together. The process then repeats until all nodes are in the same community. At each step modularity (Eq. (4)) of the partition is computed, and in the end the partition with the maximum modularity is chosen.
- *Infomap*[18] uses *Huffman code* as a way to compress the information about a path of a random walk in a graph. Each node in a graph is given a unique Id. A random walk that explores the entire graph is initialized, note that random walk tends to stay longer in a densely connected areas. This allows to combine nodes Ids into Huffman codes, that will ultimately label the communities in a graph. The algorithm then optimizes the number of Huffman codes such that the encoded information about the path of the random walk in a graph is maximally compressed.
- *Louvain*[3] maximizes modularity (Eq. (4)) in a hierarchical fashion. In the first phase of the algorithm, each node starts in its own community. Each node is then placed in a community with its neighbour such that a maximum modularity score is obtained for the graph. In the second phase a new graph is created, whose nodes are now the communities found during the first phase, such that the weight of the edges between these nodes equal to the edge density between the communities from the first phase. After the second phase is complete, the new graph is passed through the first phase again. The process is repeated until no modularity gain is obtained.
- *Leiden*[23] is an improvement of the Louvain algorithm. Louvain algorithm, in some cases, has proven to produce bad partitions in the form of disconnected communities (a community that contains disconnected nodes). Leiden algorithm introduces a refinement stage in the first phase (see Louvain algorithm). At the second phase, the refined partition is aggregated and initially partitioned based on the unrefined partition. By creating the aggregated graph based on the refined partition, Leiden algorithm has more room for identifying high-quality partitions.
Simply speaking, when a community is aggregated to a node at the second phase in Louvain algorithm, no further changes inside the community are possible. Leiden, on the other hand, gives more flexibility by refining the graph in the first phase.
Leiden also uses a faster method for implementing the nodes transitions between communities in the first phase.
- *SpinGlass*[17] adopts ideas from the field of statistical mechanics for the clustering of the graph, and is based on the Potts model [24]. The algorithm regards to the graph as a system with at most of k *spin states* (which are equivalent to the number of communities we wish to find), and tries to minimize the collective *energy* of the system by arranging its *particles* (which are equivalent to the nodes of the graph) into said spin states. The energy of the system is quantified by the *Hamiltonian* function, which is defined for a group of spin states σ_i as:

$$\begin{aligned}
& - \sum_{i \neq j} a_{ij} \cdot \underbrace{A_{ij} \delta \sigma_i, \sigma_j}_{\text{InternalEdges}} + \sum_{i \neq j} b_{ij} \cdot \underbrace{(1 - A_{ij}) \delta \sigma_i, \sigma_j}_{\text{Non-existingInternalEdges}} \\
& + \sum_{i \neq j} c_{ij} \cdot \underbrace{A_{ij} (1 - \delta \sigma_i, \sigma_j)}_{\text{ExternalEdges}} \sum_{i \neq j} d_{ij} \cdot \underbrace{(1 - A_{ij}) (1 - \delta \sigma_i, \sigma_j)}_{\text{Non-existingExternalEdges}}
\end{aligned}$$

where \mathbf{A} is the adjacency matrix (**Eq. 5**) of the graph, a_{ij}, b_{ij}, c_{ij} and d_{ij} are hyper parameters which affect the contribution of each factor to the hamiltonian and can be changed and $\delta(a, b)$ symbolizes Kronecker delta. In order to minimize the hamiltonian, the algorithm also uses simulated annealing[12].

Table 1 summarizes time complexities of each algorithm.

Algorithm	Complexity	Notes
Edge Betweenness	$O(E V (E + V))$	
Fast Greedy	$O(V (E + V))$	
Label Propagation	$O(E)$	
Leading Eigenvector	$O(V (E + V))$	
Walktrap	$O(D E V ^2)$	D the size of the dendrogram
Infomap	$O(t^2 E)$	t maximal number of iterations
Louvain	$O(E)$	
Leiden	$O(E)$	Considered to be superior to Louvain, i.e its constant factor is smaller
SpinGlass	$O(V ^{3.2})$	For sparse graphs

Table 1: CD algorithms time complexities. $|E|$ represents the total number of edges in the graph and $|V|$ the total number of nodes (vertices).

2.3 Performance evaluation

In order to compare performance between the different CD algorithms, several criteria of performance evaluation need to be defined:

- **The (asymptotic) running time complexity of each algorithm**, or for the sake of simplicity, 'how many steps the algorithm should do on given inputs until it converges, depending on the input's size'. For this criterion, we used the traditional big-O notation.
- **A scoring function to measure the accuracy of the partition an algorithm yielded**, by measuring the degree of similarity between the partition produced by an algorithm and the correct partition of the graph, i.e. the *ground truth*.

As was mentioned previously, CD algorithms are heuristic for solving the NP-hard graph partition problem, therefore cannot always cluster the graph perfectly, assuming that $P \neq NP$. Thus a scoring function can be introduced to provide us a way to quantify the quality of each partition. In our project we chose the Normalized Mutual Information (NMI) as the accuracy score function, as it is considered a good method to measure CD algorithms partition accuracy[6].

We start by introducing the Mutual Information (MI) scoring function, which incorporates ideas from the information theory. Let $G = (V, E)$ be an unweighed graph and let P be a partition of the graph to M communities C_1, C_2, \dots, C_M . *Shannon entropy* is then defined as

$$H(P) = - \sum_i \frac{|C_i|}{V} \log\left(\frac{|C_i|}{V}\right) \quad (7)$$

where $|C_i|$ is the number of nodes in C_i . It is also convenient to define the *conditional Shannon entropy*

$$H(P_1|P_2) = - \sum_{i,j} \frac{|C_i \cap C_j|}{V} \log\left(\frac{|C_i \cap C_j|}{V}\right) \quad (8)$$

where P_1, P_2 different partitions of G , and $|C_i \cap C_j|$ number of nodes that belong to both C_i, C_j . MI is then defined as:

$$MI(P_1, P_2) = H(P_1) - H(P_1|P_2) \quad (9)$$

One can note that the range of the MI quality function is $[0, \infty)$ making it impractical for comparing many cases. Thus it is useful to define the NMI function:

$$NMI(P_1, P_2) = \frac{2MI(P_1, P_2)}{H(P_1) + H(P_2)} \quad (10)$$

Note that the range of the NMI quality function is $[0, 1]$, where $NMI=1$ if the partitions are identical and $NMI=0$ if the partitions are completely different.

A very useful property of the NMI is its independency from the number of communities in each partition, providing us a convenient tool to compare two partitions with different number of communities.

A few important remarks about performance measurement should be noted:

1. **There is no absolute performance metric**, meaning an algorithm's performance may receive a 'good' score by some of the criteria but a 'bad' score by other criteria.
2. **An algorithm performance may depend on the input itself**, that is, an algorithm may perform well on some inputs, but badly on others. In order to reduce this dependability, we should analyze an algorithm's performance by considering the *expectation* of its performance according to each criterion on several random inputs.
3. **Comparison between algorithms should be done on the same inputs exactly**. In other words, it is not enough to run each algorithm on several random inputs separately, and then to compare the results.

2.4 K-means algorithm

In order to examine if CD is beneficial compared to other clustering methods, we compare the CD algorithms performance against the traditional K-means algorithm. K-means is a well known clustering algorithm used widely for pattern recognition applications, below is the description of the K-means algorithm[11].

Let $X = x_1, \dots, x_n \in \mathbb{R}^L$ be the set of L -dimensional points to be clustered into K clusters, $C = c_1, \dots, c_K$. Each cluster is represented by a centroid $\mu_k, k = 1, \dots, K$, which is the mean of all points in the cluster. K-means finds a partition of X into K clusters such that the squared error of the distances between the centroids of the clusters and the points in the corresponding clusters is minimized. Therefore, the goal of K-means is to minimize the function in **Eq. 11**.

$$J(C) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (11)$$

Minimizing the given function is an NP-hard problem [7], thus can only converge to a local minimum.

The algorithm's operation can be summarized in the following steps:

1. Select an initial partition of the data set into K partitions (K is a parameter of the algorithm). In our project we used the Kmeans++ implementation[1]:
 - (a) Choose a random point from X to be the first centroid.
 - (b) Choose the next centroid μ_k , selecting $\mu_k = x' \in X$ with probability $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$, where D is distance from the centroid.
 - (c) Repeat steps (b) until K centroids were chosen.
2. Assign each point in the data set to its closest centroid, i.e assign x_i to cluster c_k such that $\|x_i - \mu_k\|^2$ is minimal for $k = 1, \dots, K$.
3. Recalculate the centroids of the new clusters by computing the mean of each new cluster.
4. Repeat steps (2) and (3) until cluster membership stabilizes.

For a fixed maximal number of iterations t the time complexity of K-means is $O(ndKt)$ where n is the number of data points, d the dimension of each point and K number of specified clusters.

3 Results

To validate the solution we propose, we performed a number of simulations with the aim of testing CD algorithms on the MRA problem. We used three types of signals from which MRA samples were generated:

- Rectangular and triangle pulses, as illustrated in **Fig. 2**, $K = 2$.
- Normally distributed i.i.d signals, i.e $x_j \sim \mathcal{N}(0, \sigma^2 I)$
- Correlated normally distributed signals, ...

All signals have a length of $L = 50$ and are normalized. Simulations have been run on an Ubuntu 18.04 machine, Intel i5 2.5GHz CPU, 4GB RAM using Python. All CD algorithms were implemented using CDlib and NetworkX libraries.

3.1 Process overview

Fig. 4 shows the full process required to partition a graph generated from MRA data using Community Detection. Thus a key requirement is an efficient implementation of each step in the process to minimize the overhead caused by the graph creation.

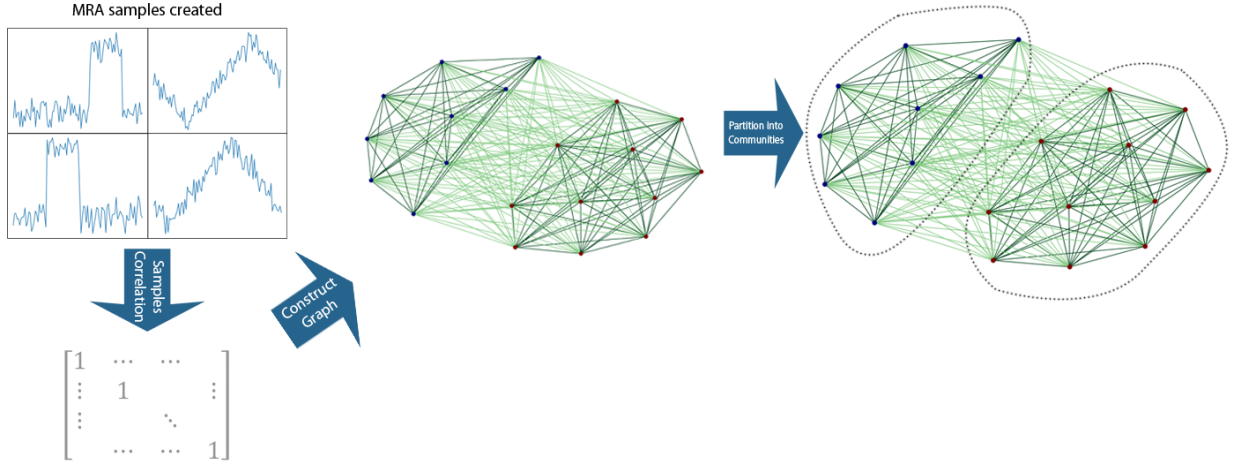
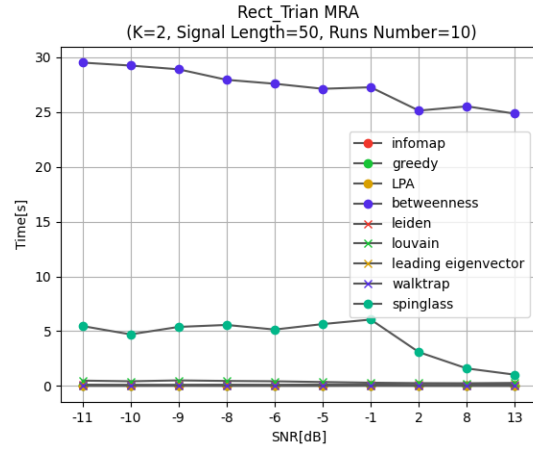
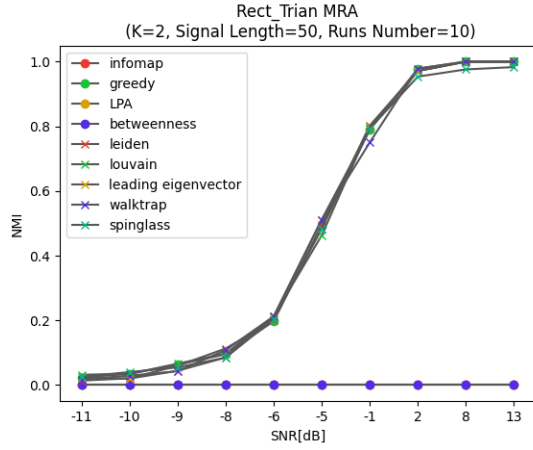


Figure 4: Full project process.

In our project we used the Numpy python package to implement vectorization methods for cross-correlation calculation and graph generation to ensure graph creation will not be the bottleneck of the solution.

3.2 Community Detection best performers selection

Different CD algorithms have different methods to find communities within a graph, some methods are more suitable to the MRA problem than others. We perform a basic simulation on 100 MRA samples generated from rectangle and triangle signals to measure the accuracy of the partitions made against a true partition of the graph. Execution time was also measured to gain an idea about the different algorithms efficiency. **Fig. 5** shows the results of the simulations. Upon observing the clustering quality simulation we conclude that there is a set of algorithms that perform similarly well, while other algorithms fail to partition the data even for high SNR. These algorithms (Fast Greedy, Leiden, Louvain, Leading Eigenvector, Walktrap) are selected as the best performers to be evaluated against KMeans. Spinglass algorithm also partitioned the data quite accurately, but the execution time simulation shows that the algorithm is slow, thus cannot be used on large datasets.



(a) **Clustering quality simulation.** NMI score of the partitions created by each CD algorithm as a function of the samples SNR.

(b) **Execution time simulation.** Average execution time of each CD algorithm as a function of the samples SNR.

Figure 5: Community Detection algorithms performance evaluation. In both graphs, for each SNR value ten different random graphs were evaluated and the results were averaged.

4 Discussion

Some detailed discussion.

4.1 A subheading

Blah blah blah

5 Conclusion

This section would contain the conclusions drawn from the entire body of work.

5.1 A subheading

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

6 References

- [1] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.
- [2] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Processing Magazine*, 37(2):58–76, 2020.
- [3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008.
- [4] Nicolas Boumal, Tamir Bendory, Roy R. Lederman, and Amit Singer. Heterogeneous multireference alignment: a single pass approach, 2018.
- [5] Saikat Chowdhury, Stephanie A Ketcham, Trina A Schroer, and Gabriel C Lander. Structural organization of the dynein–dynactin complex bound to microtubules. *Nature structural & molecular biology*, 22(4):345–347, 2015.
- [6] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008–P09008, Sep 2005.
- [7] Petros Drineas, Alan Frieze, Ravindran Kannan, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 291–299, 10 2000.
- [8] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb 2010.
- [9] Michelle Girvan and Mark Newman. Girvan, m. and newman, m. e. j. community structure in social and biological networks. *proc. natl acad. sci. usa* 99, 7821-7826. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–6, 07 2002.
- [10] Roger Guimerà and Luís Amaral. Functional cartography of complex metabolic networks. *Nature*, 23:22–231, 01 2005.
- [11] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., USA, 1988.
- [12] Scott Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science (New York, N.Y.)*, 220:671–80, 06 1983.
- [13] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), Jun 2004.
- [14] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), Sep 2006.
- [15] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10:191–218, 01 2006.
- [16] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), Sep 2007.
- [17] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), Jul 2006.
- [18] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, Nov 2009.
- [19] Sjörs HW Scheres. Beam-induced motion correction for sub-megadalton cryo-em particles. *elife*, 3:e03665, 2014.
- [20] Sjörs HW Scheres, Mikel Valle, and José-María Carazo. Fast maximum-likelihood refinement of electron microscopy images. *Bioinformatics*, 21(suppl_2):ii243–ii244, 2005.
- [21] Fred J Sigworth. A maximum-likelihood approach to single-particle image refinement. *Journal of structural biology*, 122(3):328–339, 1998.

- [22] Chun Feng Song, Kostas Papachristos, Shaun Rawson, Markus Huss, Helmut Wiczorek, Emanuele Paci, John Trinick, Michael A Harrison, and Stephen P Muench. Flexibility within the rotor and stators of the vacuolar h^+ -atpase. *PLoS One*, 8(12):e82207, 2013.
- [23] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), Mar 2019.
- [24] F. Y. Wu. The potts model. *Rev. Mod. Phys.*, 54:235–268, Jan 1982.
- [25] Wayne Zachary. An information flow model for conflict and fission in small groups¹. *Journal of anthropological research*, 33, 11 1976.

7 Appendix

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.