Exploration of Various Community Detection Algorithms and Clusterization of Artworks via Community Detection

Lee, Jae Kyeong

December 27, 2018

Abstract

The report is about the exploration of various methods of community detection. For community detection, Exponential Random Graph Model(ERGM), Modularity maximization, Stochastic Block model(SBM) with latent space, Degree-corrected Stochastic Block model(DCSBM) and Stochastic Block model Mixture of a finite mixture(MFM-SBM) will be used. The main focus of the report is capturing the differences between various methodologies of detecting communities in the network and underlying assumptions of these methods. As well as understanding, application on real data was needed. Before applying on real data, applying algorithms to famous "Zachary Karate Club" data will be preceded. After that, to grasp the differences between multiple algorithms of community detection on the perceivable level beyond the analytic level, algorithms were deployed for unstructured data such as images. For image, color distribution similarity and structural similarity were used to construct the network within the artworks, with the color histograms and the vectors obtained through the trained feature map of the convolutional neural network.

1 Introduction

Communities in a network are the groups of nodes, which are highly connected than to the rest of the nodes in the network. Community detection is the key characteristic, which could be used to extract useful information from networks. The more general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same communities, and less likely to be connected if they do not share communities. One problem is community detection, where the goal is to find a community that a certain vertex belongs to.

Many metrics exist to describe the structural features of an observed network such as density, centrality, or assortativity. However, these metrics describe the observed network which is only one instance of a large number of possible alternative networks. This set of alternative networks may have similar or dissimilar structural features. To support statistical inference on the processes influencing the formation of network structure, a statistical model should consider the set of all possible alternative networks weighted on their similarity to an observed network. Exponential-family random graph models (ERGMs) provide a principled and flexible way to model and simulate features common in social networks.

Modularity is one measure of the structure of networks or graphs. It was designed to measure the strength of the division of a network into modules (communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for detecting community structure in networks.

As opposed to working with "edge counting" objectives like modularity, the stochastic block model (SBM) is an example of a probabilistic or generative model. Generative models are a popular way to encode assumptions about the way that latent or unknown parameters interact to create edges. Then, they assign a probability value for each edge in a network.

Some of these algorithms will be deployed. Communities suggested by some of these algorithms of community detection on famous data "Zachary Karate Club" will be explored to see the apparent differences between algorithms. After that, community detection will be used to cluster artworks from a few art movements.

2 Community Detection Algorithms

Modularity Optimization Algorithms

Modularity Maximization is one measure of the structure of networks or graphs. It was designed to measure the strength of the division of a network into modules(community). Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. The value of the modularity lies in the range $[-1, 1]$. It is positive if the number of edges within groups exceeds the number expected based on chance. For a given division of the network's vertices into some modules, modularity reflects the concentration of edges within modules compared with a random distribution of links between all nodes regardless of modules. The general definition for the score of modularity, it is given as below.

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$$

where, A_{ij} is adjacency matrix,m is total number of connections, $k_i k_j$ is number of connection in between i,j,and $\delta(c_i, c_j)$ is membership. For modularity optimization, there are many algorithms such as Greedy Algorithms, Spectral Methods, Extremal Optimization, Simulated Annealing, Sampling Technique, Mathematical Programming. For this project, I majorly focused on generative models such as SBM or ERGM. Thus, even though there are various algorithms for modularity, I only used a greedy algorithm for this project.

The greedy algorithm is an agglomerative hierarchical clustering method. Initially, every node belongs to its community. Then, at each step, the algorithm repeatedly merges pairs of communities and chooses the merger for which the resulting modularity is the largest. The change in Q upon joining two communites c_i, and c_j is

$$\Delta Q_{c_i,c_j} = 2\left(\frac{|E_{c_i,c_j}|}{2|E|} - \frac{|E_{c_i}||E_{c_j}|}{4|E|^2}\right)$$

where, $|E_{c_i,c_j}|$ is the number of edges from community c_i to community c_j and $|E_{ci}| = 2|E^{in}_{c_i}| + |E^{out}_{c_i}|$ is the total degrees of nodes in community c_i. The partition with the largest value of modularity, approximating the modularity maximum best, is the result of the algorithm.

Exponential Random Graph Model(ERGM)

Exponential Random Graph Model(ERGM) is a probablistic model of Y that takes the following mathematical form:

$$P(Y = y) = \frac{exp\{H(y;\theta)\}}{\kappa_H(\theta)}$$

where $H(y;\theta)$ is the graph Hamiltonian, and $\kappa_H(\theta)$ is the normalizing constant corresponding to the probability mass function $P(Y = y)$.

In general, the graph Hamiltonian $H(y;\theta)$ can be any function of y. Often, assumes $H(y;\theta)$ is finite and takes a form as

$$H(y;\theta) = \sum_{k=1}^{p} \theta_k z_k(y)$$

An important property of ERGMs is that it allows to define a probabilisty measure of link $Y_{ij} = y_{ij}$ that is dependent on values of other links via specifying the Hamiltonian $H(y;\theta)$

If Y_{ij} is independent of other link variables Y_{-ij},

$$\begin{aligned} P(Y_{ij} = y_{ij}|Y_{-ij} = y_{-ij}) &= \frac{P(Y_{ij=y_{ij}}, Y_{-ij} = y_{ij})}{P(Y_{-ij} = y_{-ij})} \\ &= \frac{P(Y_{ij} = y_{ij})P(, Y_{-ij} = y_{ij}))}{P(Y_{-ij} = y_{-ij})} \\ &= P(Y_{ij} = y_{ij}) \end{aligned}$$

If Y_{ij} is dependent of other link variables Y_{-ij},

$$\begin{aligned} P(Y_{ij} = y_{ij} | Y_{-ij} = y_{-ij}) &= \frac{P(Y_{ij=y_{ij}}, Y_{-ij} = y_{ij})}{P(Y_{-ij} = y_{-ij})} \\ &\neq P(Y_{ij} = y_{ij}) \end{aligned}$$

Under an ERGM,

$$P(Y_{ij} = y_{ij} | Y_{-ij} = y_{-ij}) = \frac{exp\{H((y_{ij}, y_{-ij}); \theta)\}}{\kappa_H^+(\theta) + \kappa_H^-(\theta)}$$

where,

$$\begin{aligned} \kappa_H^+(\theta) &= exp\{H((y_{ij} = 1, y_{-ij}); \theta)\} \\ \kappa_H^-(\theta) &= exp\{H((y_{ij} = 0, y_{-ij}); \theta)\} \end{aligned}$$

Using this equation, pseudo-likelihood method which operates by maximizing the so-called pseudo-likelihood defined by

$$l(\theta) = \sum_{ij} ln(P(Y_{ij} = y_{ij} | Y_{-ij} = y_{-ij})$$

Although this method is intuitively appealing, the properties of the resulting estimator for exponential graph models are unknown. Thus, MCMC method was proposed for estimating the parameters.

Stochastic Block Model(SBM)

Stochastic Block Model(SBM) is a generative model for random graphs. This model tends to produce graphs containing communities, subsets characterized by being connected with particular edge densities. For example, edges may be more common within communities than between communities.
Suppose, each vertex i has type $z_i \in \{1, ..., k\}$ where k is number of community. M is stochastic block matrix of group-level connection probabilites and Q_{z_i,z_j} is probability that i,j are connected. The likelihood function of probability of A given labeling z and block matrix M when $A_{ij}|M \sim Bernoulli(Q_{z_i,z_j})$

$$\begin{aligned} P(A|z, M) &= \prod_{(i,j)\in E} Q_{z_i,z_j} \prod_{(i,j)\in E} (1 - Q_{z_i,z_j}) \\ &= \prod_{rs} Q_{r,s}^{e_{r,s}} (1 - Q_{r,s})^{n_s n_r - e_{r,s}} \end{aligned}$$

General SBM is,

$$P(A|z, M) = \prod_{ij} f(A_{ij} | M_{R(z_i,z_j)})$$

where, A_{ij} is value of adjacency, R is partition of adjacencies, f is probability function. When f is binomial it means simple graphs, poisson for multi-graphs and Normal for weighted graphs.

If $M \sim_{iid} Multinomial(\rho)$, $A_{ij}|M \sim Bernoulli(Q_{z_i,z_j})$ and block assignment is stochastic but iid, log-likelihood gets very complicated as below,

$$L(Q,\rho) = log \sum_{z \in \{1:k\}^n} \left[\prod_{r,s} Q_{rs}^{e_{rs}(z)} (1-Q_{rs})^{n_r n_s(z) - e_{rs}(z)} \prod_{r} \rho_r^{n_r} \right]$$

Thus, use EM algorithm, Gibbs sampling, etc.

Then, swap any two of the block labels and measure differences in M between estimates in permutation-invariant ways such as minimum over permuting 1 to k.

Degree Corrected Stochastic Block Model(DCSBM)

In the degree-corrected blockmodel, the probability distribution depends not only on the parameters introduced previously but also on a new set of parameters θ_i controlling the expected degrees of vertices i.

$$P(A_{ij}|z, M, \theta_i, \theta_j) = \theta_i \theta_j Q_{rs}$$

θ helps account for broad degree distribution.
Math simplifies if we pretend $A_{ij} \sim Poisson(\theta_i, \theta_j Q_{z_i,z_j})$.

According to the paper, the uncorrected model prefers to split networks into groups of high and low degree..The degree-corrected model correctly ignores divisions based solely on degree and hence is more sensitive to underlying structure.

Mixture of Finite Mixture-Stochastic Block Model(MFM-SBM)

The mixture of Finite Mixture-Stochastic Block Model(MFM-SBM) is SBM implemented with the Bayesian non-parametric techniques to overcome the limitation in the uncertainty of the chosen number of clusters. At the same time, a mixture of a finite mixture (MFM) was suggested to minimize the snags of the Chinese Restaurant Process(CRP) which makes extraneous clusters.

General framework for prior specification is

$$\begin{aligned} &z = (z_1, ..., z_n) \\ &(z,k) \sim \Pi \\ &Q_{rs} \sim U(0,1) \\ &A_{ij}|z, M, k \sim Bernoulli(Q_{z_i,z_j}) \end{aligned}$$

Π is a probability distribution on the space of partitions of $\{1, ...n\}$ With known k,

$$\begin{aligned} z_i|\pi &\sim Multinomial(\pi_1, ..., \pi_k) \\ \pi &\sim Dir(\alpha/k, ..., \alpha/k) \end{aligned}$$

For unknown k, through Chinese Restaurant Process(CRP)

$$P(z_i = c|z_1, ..., z_{i-1}) \propto \begin{cases} |c|, \text{at an exisiting table labeled c} \\ \alpha \end{cases}$$

Marginally, the distribution of z_i is given by the stick-breaking formulation of a Dirichlet process.

$$z_i \sim \sum_{h=1}^{\infty} \pi_h \delta_h, \quad \pi_h = \nu_h \prod_{l<h} (1 - \nu_l), \quad \nu_h \sim Beta(1, \alpha)$$

CRP assigns large probabilities to clusters with relatively smaller size.
A modification of the CRP based on a mixture of finite mixtures (MFM) model proposed to circumvent this issue.

$$k \sim p(\cdot), \quad (\pi_1, ..., \pi_k)|k \sim Dir(, ..., \gamma), \quad z_i|k, \pi \sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1, .., n$$

where $p(\cdot)$ is a proper p.m.f on 1,2.., and δ_h is a point-mass at h.
The joint distribution of $(z_1, .., z_n)$ under these conditions admit a Polya urn scheme akin to CRP.

$$P(z_i = c|z_1, ..., z_{i-1}) \propto \begin{cases} |c| + \gamma, \text{at an existing table labeled c} \\ \frac{V_n(t-1)}{V_n(t)} \gamma \text{if c is a new table} \end{cases}$$

$$where, V_n(t) = \sum_{n=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^(n)} p(k)$$

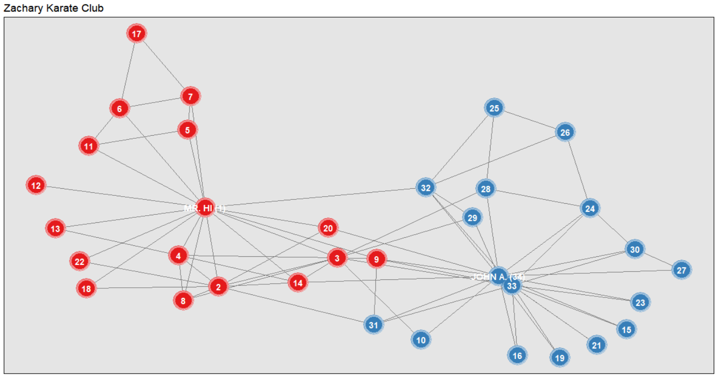
Adapting MFM to the SBM setting, our model and prior can be expressed hierarchiacally as

$$\begin{aligned} & k \sim p(\cdot), \quad \text{where} \quad p(\cdot) \quad \text{is a p.m.f on } 1, 2.., \\ & Q_{rs} = Q_{sr} \sim Beta(a, b), \quad r, s = 1, ..., k \\ & \pi|k \sim Dir(\gamma, ..., \gamma) \\ & pr(z_i = j|\pi, k) = \pi_j, \quad j = 1, .., k, \quad i = 1, .., n \\ & A_{ij}|z, Q, k \sim Bernoulli(\theta_{Q_{z_i, z_j}}) \end{aligned}$$

This model requires four hyperparameters which is λ, γ, α and β. λ is for p.m.f for k which is the number of clusters, γ for the parameter of dirichlet distribution that controls the relative size of clusters. α and β are the parameter for beta distribution that decides on the probability distrivution of two nodes belong to same community.

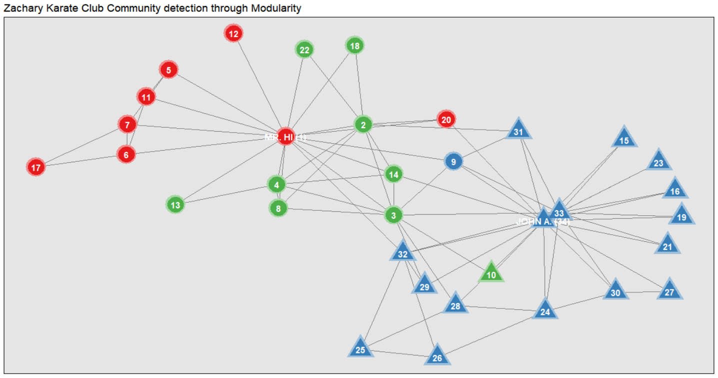
3 Application on Zachary Karate Club

Zachary's karate club is a well-known social network of a university karate club described in the paper "An Information Flow Model for Conflict and Fission in Small Groups" by Wayne W. Zachary. The network became a popular example of community structure in networks after its use by Michelle Girvan and Mark Newman in 2002



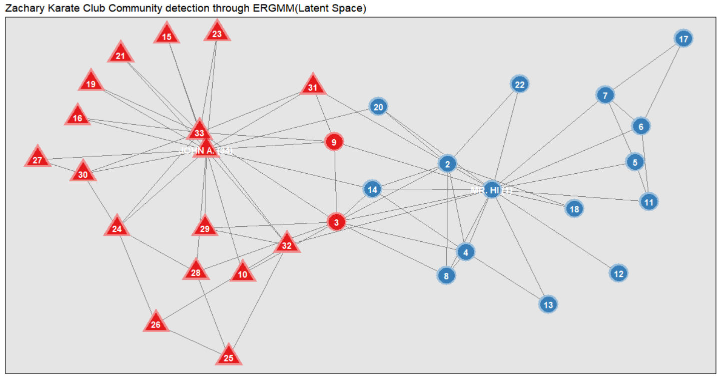
A social network of a karate club was studied by Wayne W. Zachary for a period of three years from 1970 to 1972. The network captures 34 members of a karate club, documenting pairwise links between members who interacted outside the club. During the study a conflict arose between the administrator "John A" and instructor "Mr. Hi" , which led to the split of the club into two. Half of the members formed a new club around Mr. Hi; members from the other part found a new instructor or gave up karate.

Modularity Optimization Algorithms



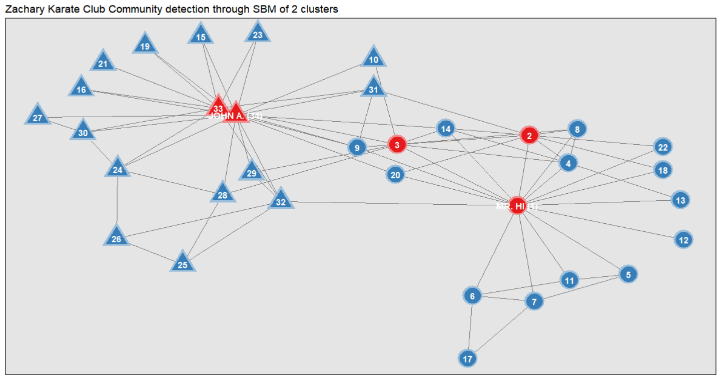
If we consider red and green as one group then, only two of the nodes were misclassified out of 34 nodes.

Exponential Random Graph Model(ERGM)


Zachary Karate Club Community detection through ERGMM(Latent Space)

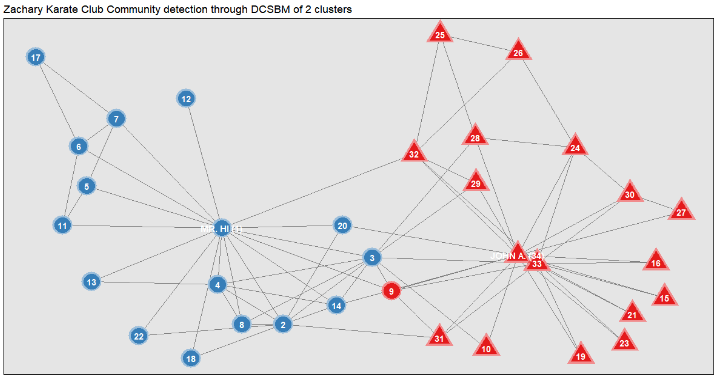
Due to the characteristic of ERGM, it required some specific nodes as reference to detect the community. In Zachary's data, "Mr.Hi" and "John.A" were used as the center character. Furthermore, for this application, ERGM is not solely used. Latent space was implemented in community detection. Only two of the nodes were misclassified out of 34 nodes.

Stochastic Block Model(SBM)


Zachary Karate Club Community detection through SBM of 2 clusters

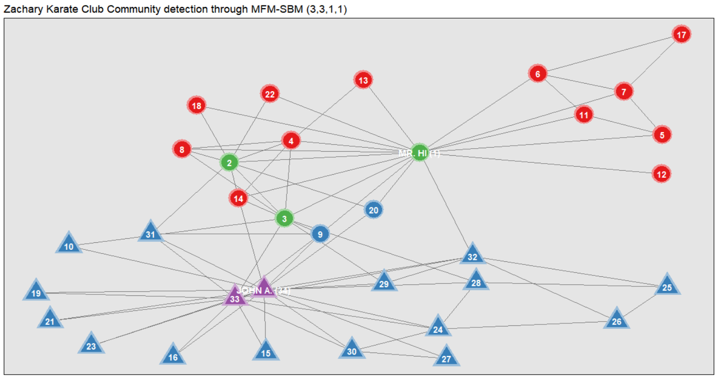
Stochastic Block Model did not great job on this data. According to paper, this model prefers to split networks into groups of high and low degree. As it is on the graph, both central characters, "Mr.Hi" and "John.A" were grouped together, as well as other influential nodes.

Degree Corrected Stochastic Block Model(DCSBM)



Through introducing new parameter on each vertices, DCSBM worked very well on community detection on this data. Only one node was misclassified.

Mixture of Finite Mixture-Stochastic Block Model(MFM-SBM)



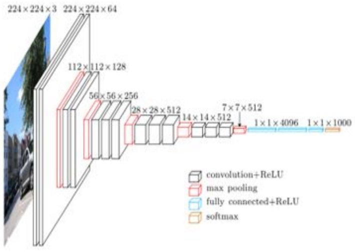
Through MFM-SBM, biggest number of community was suggested. However, how community was formed is interesting. It separated the two group relatively correct and furthermore, it separated once more with influential node and uninflential nodes. This result seems to be a combination of SBM and DCSBM.

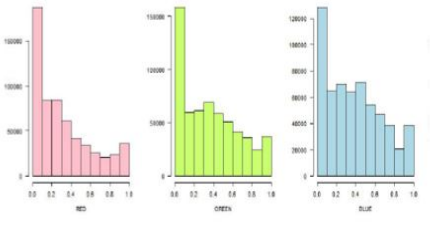
4 Application on Clusterization of Artworks

Using the similarity between unstructured data, I thought community detection algorithms could be applied. Through this trial, I thought that the result may show characteristics of each algorithm on the visually perceptible level better than the Zachary karate club, since best we can see from the result of Zachary karate club is the network graph, whereas people can capture the characteristics of artworks if the picture is shown together. This may end to no avail, but it would worth a try clustering the unstructured data with community detection algorithms.

The first work of clustering the image was to define the word ”similar”. Unlike structured data, there is no firm definition of similarity of images. What I came up with was the color distribution similarity and structure similarity. Since artworks that follow a certain art movement have their tendency of color use. For example, Fauvism tends to use intense colors, whereas impressionism tends to prefer soft colors. Therefore, I thought each color distribution can be a trait that changes the degree of similarity. To express the color distribution into a degree of similarity, I have separated color data in each red, blue, and green dimension of an image. Then I made a color histogram with 50 breaks for each color. Thus, I made a vector of 1 by 150 using proportion according to each break in a histogram. Through this procedure, each artwork obtains a vector that represents its color distribution. Using these vectors, I calculated the cosine similarity to acquire the color distribution similarity. For this process, I should have thought of a more reasonable way to calculate the similarity through other references, but due to time constraint, I could not research enough.

The structural similarity was defined to implicate the overall shape of the artworks. To vectorize the structure of an image, I have thought of the feature map of the convolution neural network. A feature map is a function which maps a data vector to feature space. For this case, I have used the feature map of the convolutional neural network(CNN).CNN is a special type of Feed-Forward Artificial Neural Networks that is generally used for image detection tasks. It accepts a large array of pixels as input to the network. The hidden layers of the network carry out feature extraction from the image. There is a fully connected layer that detects the image and classifies it. The Convolution layer uses a filter





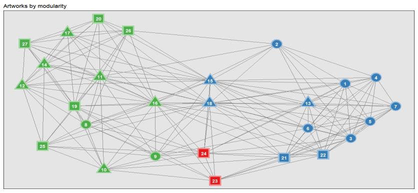
matrix over the array of image pixels and performs convolution operation to obtain a convolved feature map.

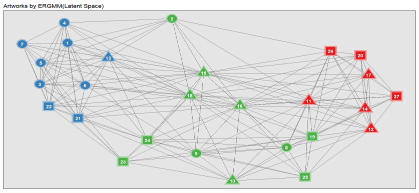
For this specific process, I have used the basic model, VGG16. For this model, I used the pre-trained feature map by 1000 classes from ImageNet. I did not train the model with artworks, because I thought it would be unnecessary in that this is not a process of spotting specific objects from the image but a process to obtain vector to calculate the relative difference between artworks. However, I would be a good try to use a feature map trained with the artworks later. If I put an image to this feature map, it spits out 1 by 1000 vector. With these vectors, I have calculated the cosine similarity.

I used 27 artworks each 9 from 3 different art movements. For 3 of the art movement, Impressionism, Fauvism, Abstractionism was included.

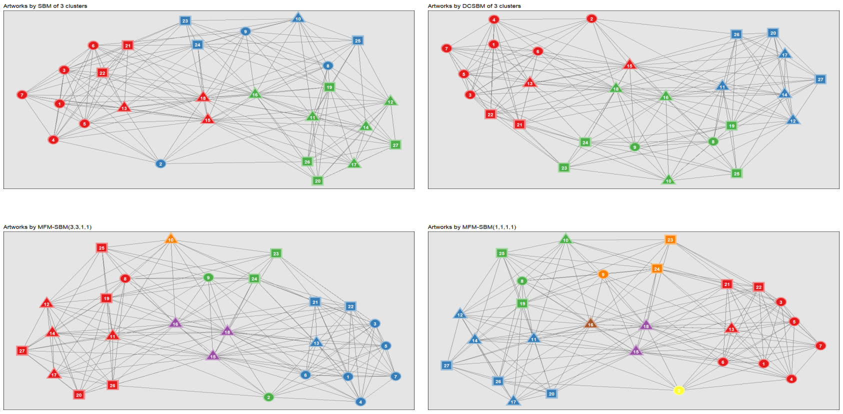
Respectively with color distribution similarity or structural similarity- or with both of them - matrix can be built with the similarity between artworks. However, it has to be reduced since all the nodes are connected. Here I had an issue of how to decide the threshold for cutting out the edges with low similarity. I could not give much thought on this, and just left edges with 10 highest similarities. With this network, I applied community detection algorithms.

Structural Similarity





These are the clustering of the artworks with the structural similarity


Artworks by SBM of 3 clusters
Artworks by DCSBM of 3 clusters
Artworks by MFM-SBM(3,3,1,1)
Artworks by MFM-SBM(1,1,1,1)

through community detection. The shapes of the vertex indicate the art movement and the colors of vertex manifest the community suggested by the algorithm. The algorithm which got the closet result to compare to the art movement is DCSBM. The below figures are the artworks grouped by SBM, DCSBM, and MFM-SBM(3,3,1,1).

Figure 1: Artworks clustered by SBM

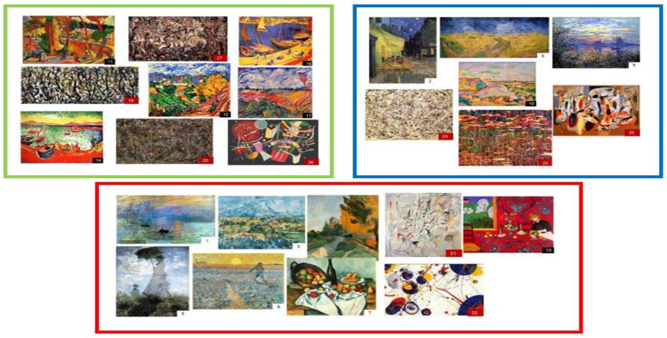


Figure 2: Artworks clustered by DCSBM

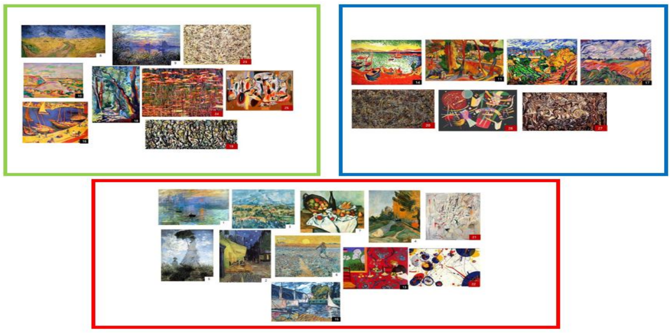
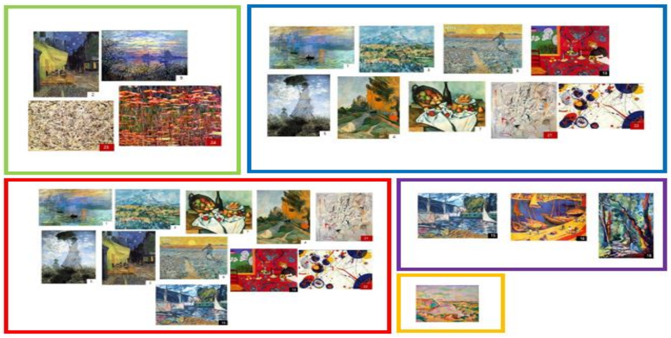


Figure 3: Artworks clustered by MFM-SBM(3,3,1,1)



Clustering with color distribution similarity and both similarity follows a similar pattern of the result of structural similarity. I wanted to post it on the report but due to the deficiency inaptness in latex, I could not manage to upload all the results neatly. So for the pictures, please refer to the presentation material.

Out of all the tries, clustering with both similarity through DCSBM seemed to have the closest result to the ground-truth for an art movement.

5 Limitations and Further Study

There were many limitations and things to be improved. First of all, I Could not fully cover all the algorithms that I planned to study. It is still true that I invested most of my time studying various algorithms, there was always more to study for each algorithm. Since some of the algorithms were distinct to each other, It was hard to fully understand the ideas of each algorithm and variations.

For those algorithms that I had to choose the number of community, I could not try on numbers other than the number which thought to be an answer in problem setting. For the Zachary Karate club, I knew the right number of communities was two and the same for artworks. So I should have tried on different numbers of community on the same data and see how different they are.

The communities within the artworks should be heavily on the definition of similarity of artworks and many other rather arbitrary assumptions I have made. This was the biggest problem of my mini-project and the thing that has to be improved. For both, color distribution similarity and structural similarity, further research on references is needed. Furthermore, since the similarity between the two pictures exists for every pair of pictures, I have picked the top 10 edges with the highest similarity. There must be a better way for this and this choice was rather arbitrary. Also, weights were not taken into consideration, since I was not able to use some of the algorithms properly, so it can be suitable for weighted network.

Used VGG16 which is old and model with the pre-trained model, not training myself with actual artworks. I thought the pre-trained model would be okay but using a model trained on artworks may worth a try. Both for structure similarity and color distribution similarity, only cosine similarity was considered. More ways of calculating similarity can be researched.

Lastly, it was only applied to a small number of artworks. It would be more fun if I applied this process to larger data of artworks.

References

[1] Geng, J., Bhattacharya, A., Pati, D. (2018). Probabilistic community detection with unknown number of communities. *of the American Statistical Association,*,1-32.

[2] Krivitsky, P. N., Butts, C. T. (2013). Modeling valued networks with statnet. The Statnet Development Team, 2013.

[3] Caimo, A., Friel, N. (2012). Bergm: Bayesian exponential random graphs in R. arXiv preprint arXiv:1201.2770.

[4] Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. Electronic journal of statistics, 6, 1100.

[5] Karrer, B., Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. Physical review E, 83(1), 016107.