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

Lee, Jae Kyeong

Department of Statistics
University of Korea

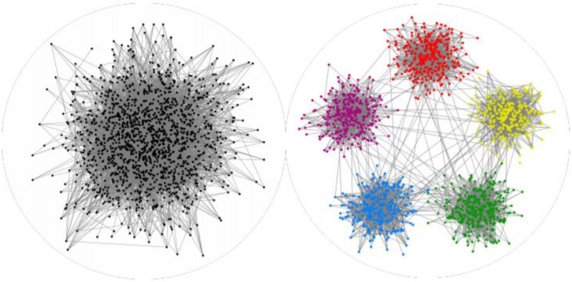
December, 2018

Contents

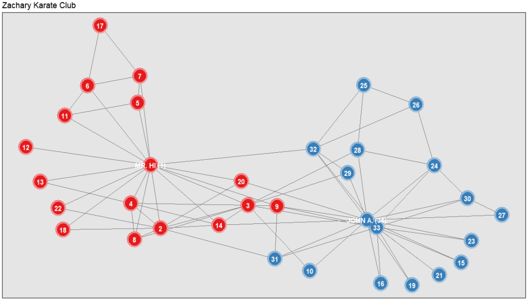
Various Community Detection Algorithms

What is Community Structure?

A network is said to have community structure if the nodes of the network can be easily grouped into sets of nodes such that each set of nodes is densely connected internally.



Zachary's karate club is a well-known social network of a university karate club. The network became a popular example of community structure in networks after its use by Michelle Girvan and Mark Newman in 2002.



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" (pseudonyms), which led to the split of the club into two.

Various Community Detection Algorithms

- Modularity Optimization Algorithms
- Exponential Random Graph Model(ERGM)
- Stochastic Block Model(SBM)
- Degree Corrected Stochastic Block Model(DCSBM)
- Mixture of Finite Mixture - Stochastic Block Model(MFM-SBM)

Modularity Maximization is a one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules(community).
However, it has been shown that modularity suffers a resolution limit and, therefore, it is unable to detect small communities.

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

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.

Modularity Optimization Algorithms

- 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 greedy algorithm for this project.

Modularity measures the difference between the actual fraction of edges within the community and such fraction expected in a randomized graph with the same number of nodes and the same degree sequence.

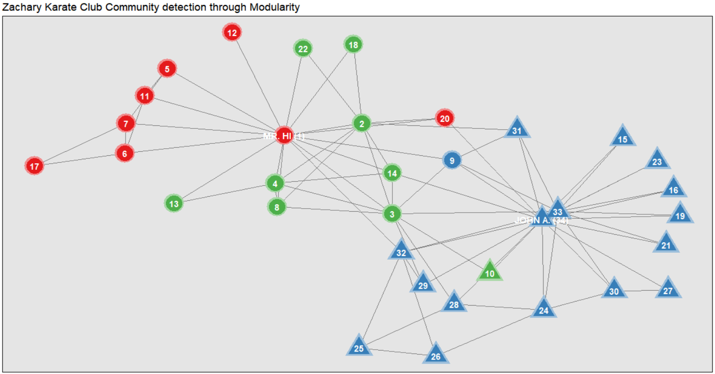
Greedy Algorithms The greedy algorithm is a agglomerative hierarchial clustering method. Initially, every node belings to its own community.
Then, at each step, the algorithm repeatedly merges pairs of communities together and chooses the merger for which the resulting modularity is the largest.

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 $|E_{c\,i}| = 2|E_{c_i}^{in}| + |E_{c_i}^{out}|$ 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.



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) 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,

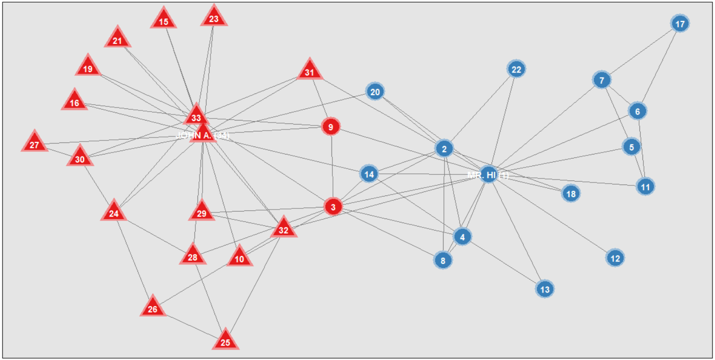
$$
\kappa_H^+(\theta) = exp\{H((y_{ij} = 1, y_{-ij}); \theta)\}
$$
$$
\kappa_H^-(\theta) = exp\{H((y_{ij} = 0, y_{-ij}); \theta)\}
$$

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.

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) is a generative model for random graphs. This model tends to produce graphs containing communities, subsets characterized by being connected with one another 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 $A_{ij}|M \sim Bernoulli(Q_{z_i,z_j})$

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

General SBM is,

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

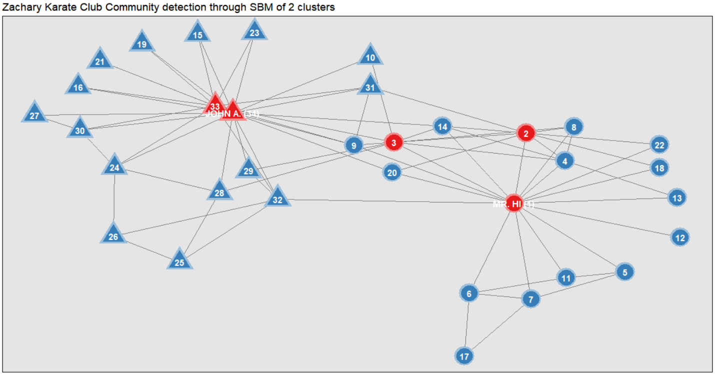
where, A_{ij} is value of adjacency, R is partition of adjacencies, f is probability function. When f is binomial it means simple graphs, poisson for multi-graphs and Normal for weighted graphs.

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.



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.

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.

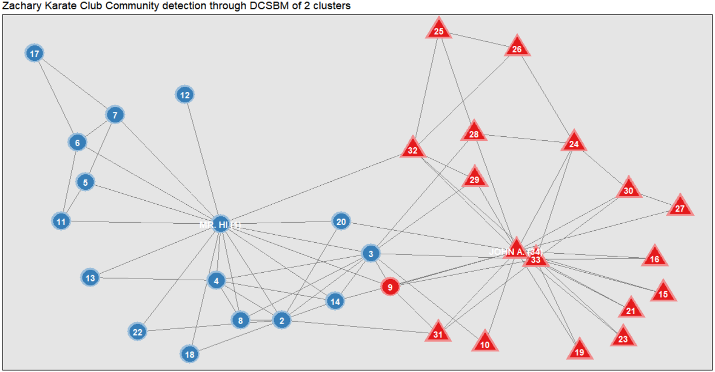
$$P(A_{ij}|z, M, \theta_i, \theta_j) = \theta_i \theta_j Q_{rs}$$

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.

Various Community Detection Algorithms

Degree Corrected Stochastic Block Model(DCSBM)



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

Mixture of Finite Mixture-Stochastic Block Model(MFM-SBM) is SBM implemented with the Bayesian non-parametric techniques to overcome the limitation in uncertainty of the chosen number of clusters.

At the same time, mixture of finite mixture (MFM) was suggested to minimize the snags of Chinese Restraurant 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,

$$z_i|\pi \sim Multinomial(\pi_1, ..., \pi_k)$$
$$\pi \sim Dir(\alpha/k, ..., \alpha/k)$$

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

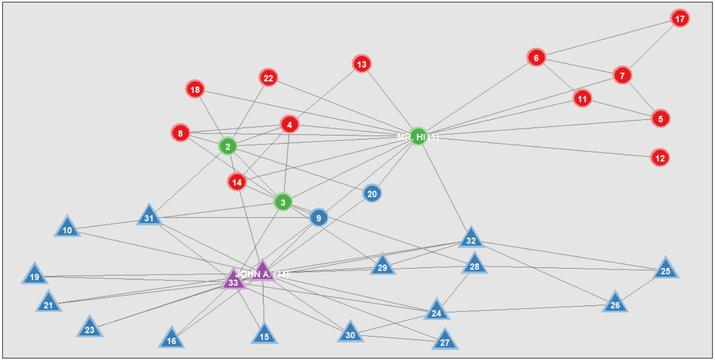
$$
\begin{aligned}
& k \sim p(\cdot), \quad \text{where} \quad p(\cdot) \quad \text{is a p.m.f on } 1, 2.., \\
& Q_{rs} = Q_{sr} \sim Beta(a, b), \quad r, s = 1, ..., k \\
& \pi | k \sim Dir(\gamma, ..., \gamma) \\
& pr(z_i = j | \pi, k) = \pi_j, \quad j = 1, .., k, \quad i = 1, .., n \\
& A_{ij} | z,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

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.

Various Community Detection Algorithms

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

Zachary Karate Club Community detection through MFM-SBM (3,3,1,1)



Application on Clusterization of Artworks

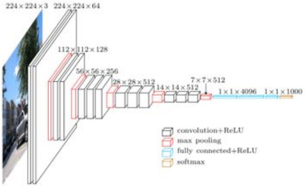
Using the similarity between artworks, let's apply community detection algorithms to the artworks.

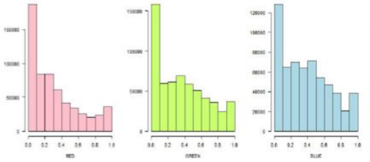
- It would show characteristics of each algorithms on the visually perceptible level.
- It would be fun to just try it.

Application on Clusterization of Artworks

What does "Similar" mean in images.

- Structural Similarity
  - Deployed feature map of Convolutional Neural Network
  - The basic CNN model, VGG16 from ImageNet which was pre-trained with 14million labeled image with 1000 classes was used.
  - Cosine Similarity between the vectors obtained from the feature map.
- Color Distribution Similarity
  - Utilized color histogram of image
    - Composed vector of 1 by 150 composed by 50 partitions of color histogram in RGB respectively.
    - Also, Cosine Similarity between the vectors were calculated



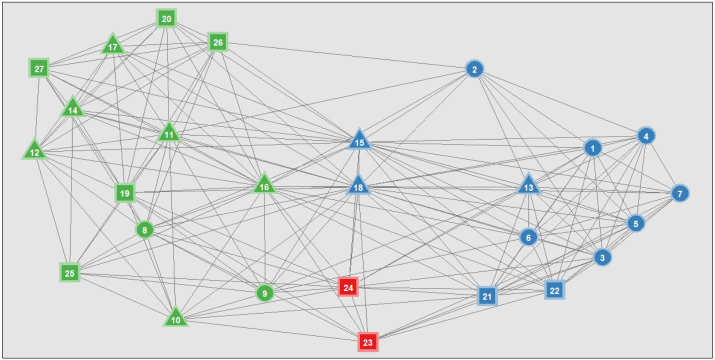


Application on Clusterization of Artworks
Artworks

- 27 Artworks from 3 different art movements
  - Impressionism
  - Fauvism
  - Abstractionism

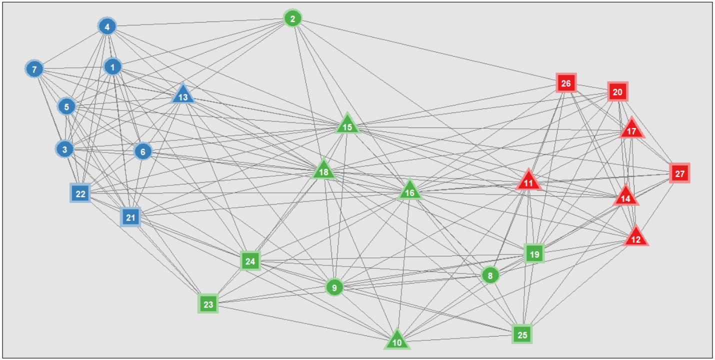
Modularity

Artworks by modularity



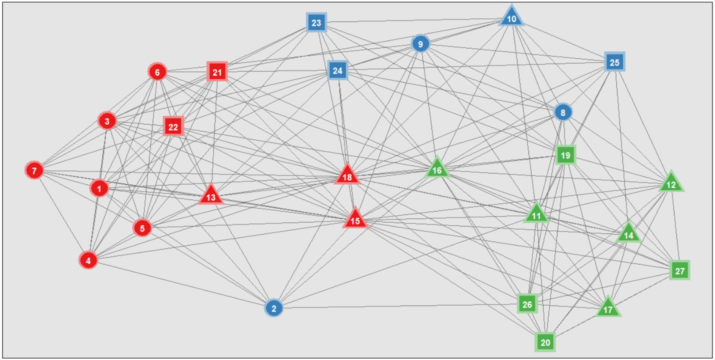
ERGM

Artworks by ERGMM(Latent Space)



SBM

Artworks by SBM of 3 clusters

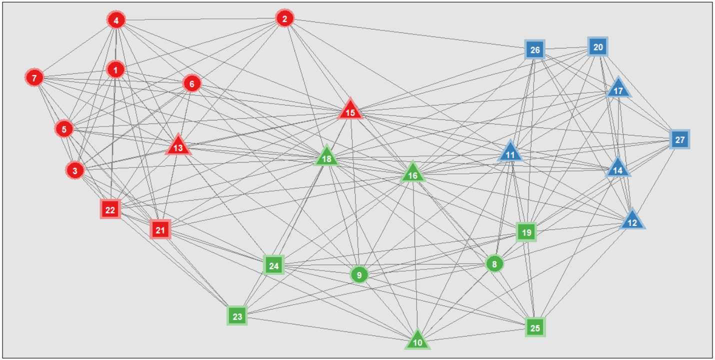


Application on Clusterization of Artworks
Structural Similarity

SBM

DCSBM

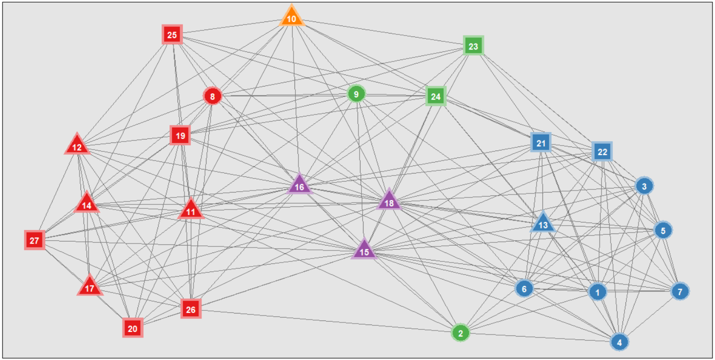
Artworks by DCSBM of 3 clusters



DCSBM

MFM-SBM

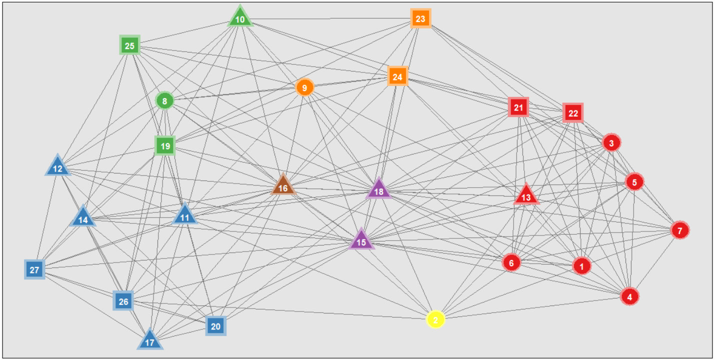
Artworks by MFM-SBM(3,3,1,1)



MFM-SBM

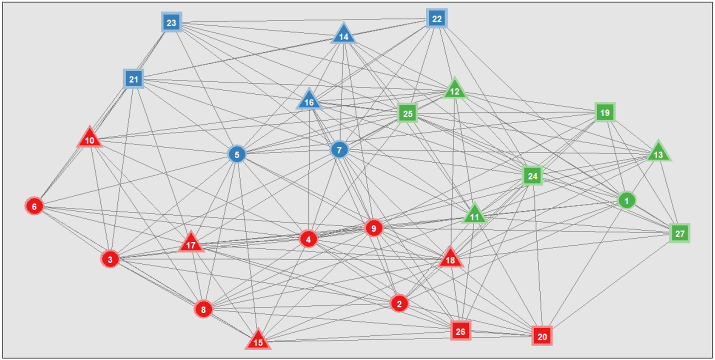
MFM-SBM

Artworks by MFM-SBM(1,1,1,1)



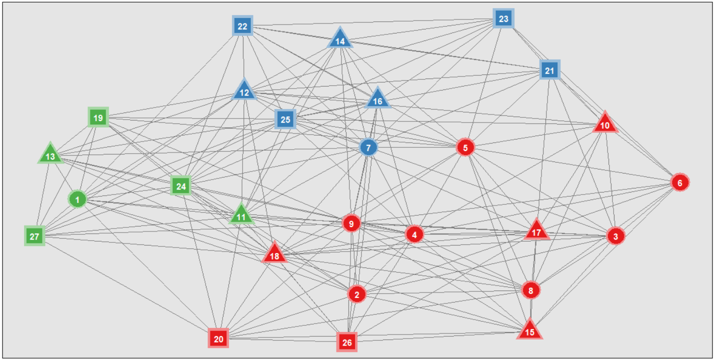
Modularity

Artworks by modularity(color histogram)

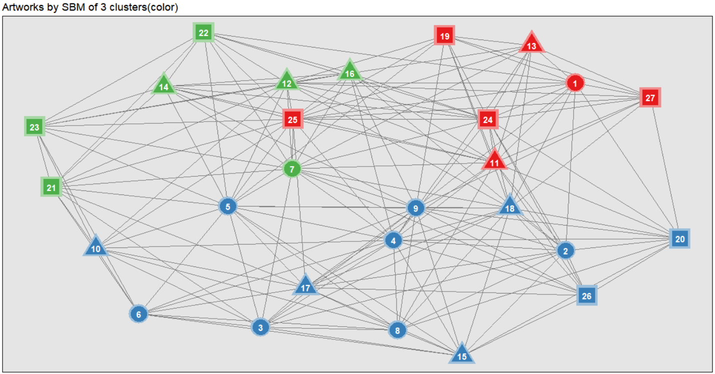


ERGM

Artworks by ERGMM(Latent Space)(color)

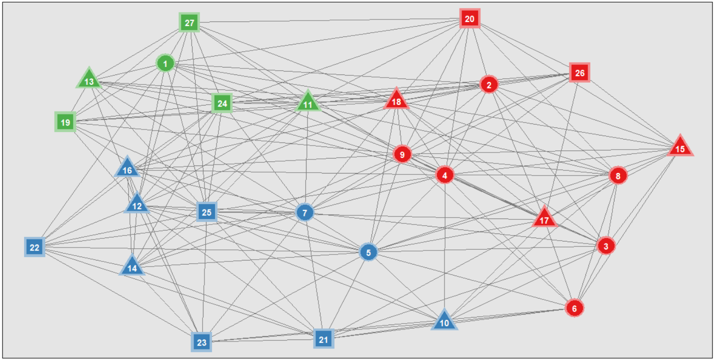


SBM


Artworks by SBM of 3 clusters(color)
22
19
13
16
14
12
1
27
23
25
24
11
7
21
5
9
18
20
4
2
10
17
26
6
3
8
15

DCSBM

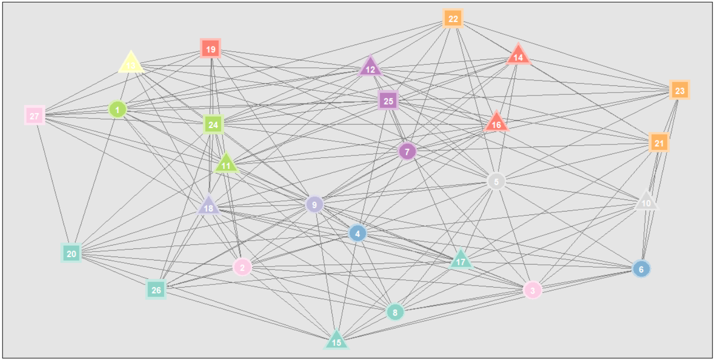
Artworks by DCSBM of 3 clusters(color)



DCSBM

MFM-SBM

Artworks by MFM-SBM(3,3,3,1)(color)



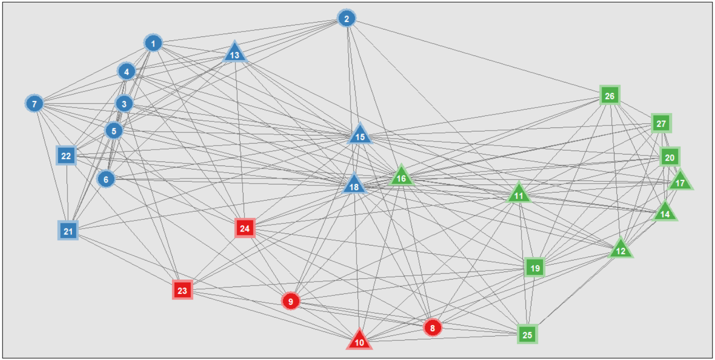
Application on Clusterization of Artworks

Color Distribution Similarity

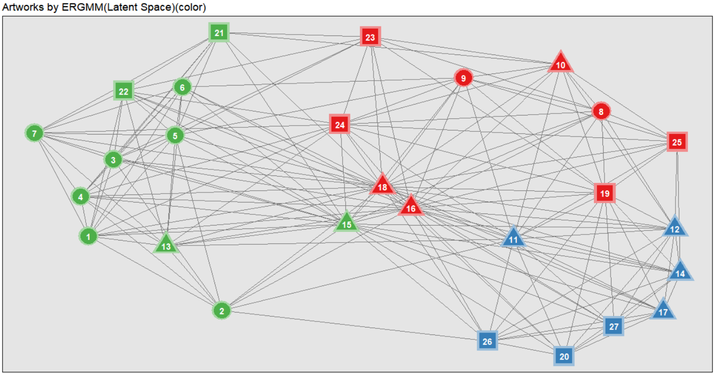
MFM-SBM

Modularity

Artworks by modularity(structure + color)



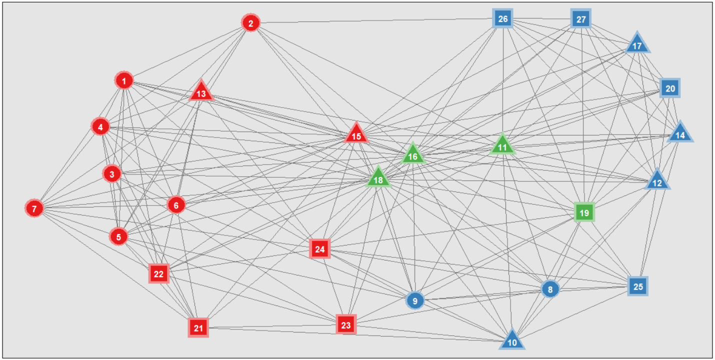
ERGM



Artworks by ERGMM(Latent Space)(color)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

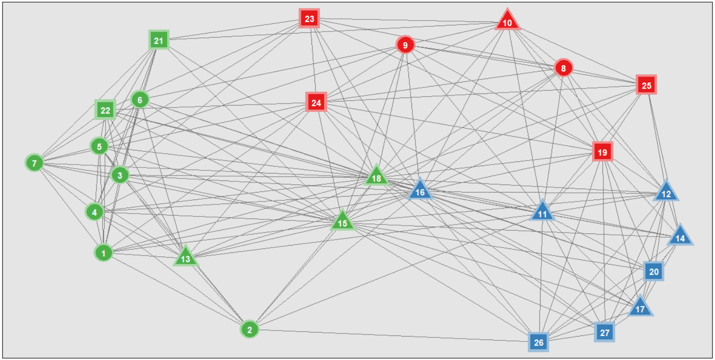
SBM

Artworks by SBM of 3 clusters(structure + color)



DCSBM

Artworks by DCSBM of 3 clusters(color)



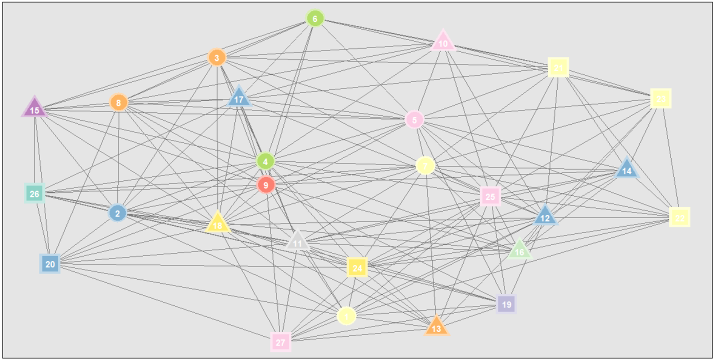
Application on Clusterization of Artworks

Structure and Color Distribution Similarity

DCSBM

MFM-SBM

Artworks by MFM-SBM(1,1,1,1)(structure + color)



- Could not fully cover all the algorithms that I planned to study.
- For those algorithms that I have to choose number of community, I could not try on numbers other than the number which thought to be an answer in problem setting.
- The communities within the artworks should be heavily on the definition of similarity of artworks and many other rather arbitrary assumption I have made.
  - Definition of similarity between picture itself
  - Since the similarity between two pictures exists for every pair of picture, I have picked top 10 edges with highest similarity.
  - 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 pre-trained model, not training myself with actual artworks. I thought 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.
- Only applied on small number of artworks.