In the name of god

probability and Statistics Project

methods of clustering graphs

Course thacher
Dr. Mohammad Hossein Yasai

Dead line:
2022

1 Introduction

Exams are over and you want to spend your leisure time with a peaceful mind, so you can move on to the next semester with abundant energy. A suitable option for this is watching a movie. You enter Netflix/Disney+/Vimeo through the software or website and on the first page, you see a number of recommendations that are not bad and are in line with your taste. Gradually, you become curious to know on what basis these recommendations are presented to you (and to all users), and here you pause the movie and move on to the sweet project of probability and statistics course...

figure 1: Some of the requested video content services

You are probably familiar with Video On Demand (VOD) services. One of the key elements of these services is providing suitable recommendations to different users. Let's say you have been hired by one of these companies and you want to work on the movie recommendation section for users. To do this, first, you need to identify groups of users who have similar tastes (for example, they all like comedy movies or they all like horror movies, etc.), and then based on that, suggest similar movies to those groups. In this project, we want to first identify groups of users with similar tastes and then recommend movies that suit their preferences. A simple example of this process is shown in Figure 2.

A friendly recommendation at the beginning: Before anything else, read Section 7. After that, at least read the project description carefully and thoroughly at least twice, but don't write the answers to any questions yet. Once you've finished reading the project description, start solving the project and writing the answers to the questions.

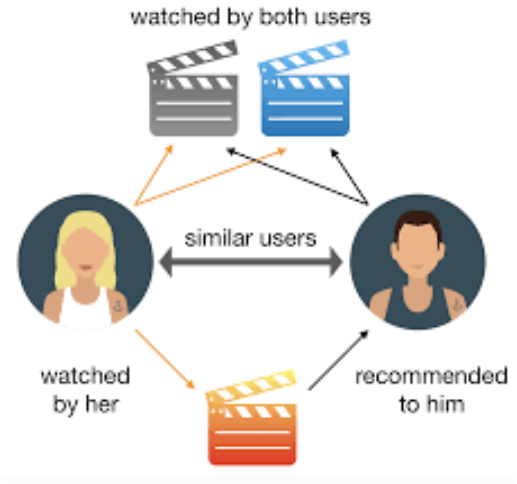


figure 2: The impact of videos viewed by users on system suggestions

2 A

First, we need to find a mathematical model to describe the relationships between individuals. As you have probably guessed by now, the best mathematical model that can describe the relationships between different individuals in this problem is a graph. We can model each individual with a graph vertex and draw an edge between two individuals with similar tastes. For each graph, we can define an adjacency matrix.

definition 1. Adjacency matrix for Graph $\mathcal{G}$ with n Vertex of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ The elements are as the below:

$$A_{i,j} = \begin{cases} 1 & \text{if between vertexes } i, j \text{ there is any edge} \\ 0 & \text{other wise} \end{cases}$$

Our purpose in this project is to find groups of people who share a common taste.

---

theory question 1. If the community can be partitioned into distinct clusters of individuals with the same taste, such that individuals in one cluster have no overlap with individuals in another cluster, then what would be the form of the adjacency matrix for these individuals?

theory question 2. The model you obtained in the previous question does not occur in reality because it is possible for a person in one cluster to have some common preferences (although minimal) with a person in another cluster. Additionally, a person can be a member of multiple clusters simultaneously. For example, a person can be part of a cluster of people who are generally interested in the comedy genre and at the same time be present in a cluster of people who watch drama genre. In this case, what description can be given for the adjacency matrix?

theory question 3. Assume that we have M different genres of movies, denoted as $1, 2, \ldots, M$. If two individuals have an interest in the i-th genre, they are likely to have similar tastes with probability p_i. Now, let's suppose that these two individuals have interests in multiple different genres. If we assume that the interests in different genres are independent of each other, what is the probability that these two individuals have similar tastes?

Why does the likelihood of two individuals being connected increase if they are members of many specific communities? Can you explain this intuitively? In your opinion, what flaw does this model have in representing the relationship between individuals and groups? Please refer to Figure 3 for more information.

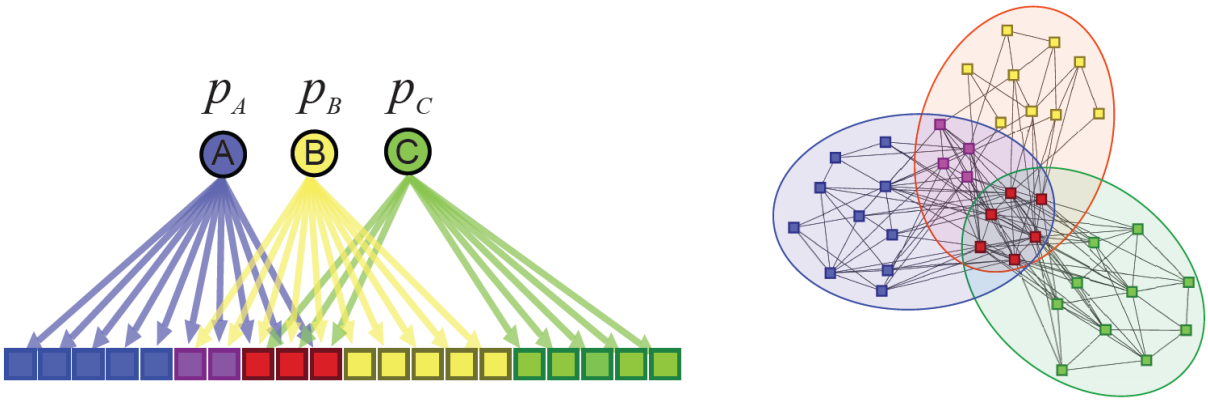


figure 3: An example of different people's interest in different genres.

To solve the issues that previous models had, we need to propose a model for identifying clusters with three features::

1. We consider a parameter to quantify individuals' inclination to belong to a group. This parameter is called "affinity." The affinity of an individual, denoted as u, to cluster c represents their level of willingness to join cluster c. We denote this parameter as $F_{uc} \in [0, \infty)$.
2. The higher the affinity of two individuals to the same cluster, the higher the probability of their homophily. To achieve this, we can define the probability of creating a homophily connection between two individuals, u and v, through cluster c as $P_{uv}(c) = 1 - \exp(-F_{uc}F_{vc})$.
3. Different clusters independently create homophily connections between individuals based on the previous probability.

If we consider the number of individuals as n and the number of clusters as C, by assigning different individuals to different clusters, we can define the affiliation matrix $\mathbf{F} \in \mathbb{R}^{n \times C}$.

Using the matrix $\mathbf{F}$, we can calculate the probability of co-affiliation between two individuals.

theory question 4. According to the above description, calculate the probability of a connection between two individuals, u and v, based on the elements of the matrix $\mathbf{F}$. Show that if the two individuals share common groups, this probability increases.

Note that, if we have the elements of matrix $\mathbf{F}$, we have the tendency of each person towards each movie category, and as a result, we can optimally suggest different movies to different individuals. But in reality, we don't have access to matrix $\mathbf{F}$, so we need to estimate it based on the existence or absence of connections between different individuals. For this purpose, we use the maximum likelihood estimator, and similar to many estimation problems, instead of maximizing the likelihood function, we maximize its logarithm.

theory question 5. We know that by having the adjacency matrix A, we can uniquely determine the graph of co-occurrence relationships among individuals (refer to Definition 1 for the adjacency matrix).
Calculate the log-likelihood function, which is defined as $l(\mathbf{F}) = \log\big(\mathbb{P}[\mathbf{A}|\mathbf{F}]\big)$.

Finding the maximum of the given function in the general case is a challenging task. Therefore, we need to numerically maximize it. To do this, at each step, we slightly move the assignments of a specific person among the different groups in the direction of the gradient. It should be noted that the assignments of person u are located in the u-th row of the matrix $\mathbf{F}$, denoted by $F_{u,:}$, and as a result, the mentioned gradient is given by:

$$\nabla_{F_{u,:}} l(\mathbf{F}) = [\frac{\partial l(\mathbf{F})}{\partial F_{u,1}}, \frac{\partial l(\mathbf{F})}{\partial F_{u,2}}, \ldots, \frac{\partial l(\mathbf{F})}{\partial F_{u,C}}].$$

theory question 6. Please intuitively argue why this method can lead us to the optimal value $\mathbf{F}$, i.e., $\mathbf{F}^* = \arg\max_{\mathbf{F}} l(\mathbf{F})$.

theory question 7. Find $\nabla_{F_{u,:}} l(\mathbf{F})$

By updating the rows of matrix $\mathbf{F}$ along the gradient of the logarithm of the likelihood function with respect to those rows in multiple iterations, we can approach the maximum point.

computer question 1. In the following code sample, the above iterative algorithm for estimating $\mathbf{F}$ is implemented in the train function. The function takes the adjacency matrix and the number of groups as inputs and returns an estimation of the affiliation matrix $\hat{\mathbf{F}}$ as output. Please complete the sections related to the function $l(\mathbf{F})$ and the gradient function.

```
def log_likelihood(F, A):
  #todo
  return log_likelihood

def gradient(F, A, i):
  #todo gradient of log_likelihood respect to person i parameters (F_ic)
  return gradient

def train(A, C, iterations = 200):
  # initialize an F
  N = A.shape[0]
  F = np.random.rand(N,C)

  for n in range(iterations):
    for person in range(N):
      grad = gradient(F, A, person)
      F[person] += 0.005*grad                    # updating F
      F[person] = np.maximum(0.001, F[person]) # F should be nonnegative
    ll = log_likelihood(F, A)
    print('At step %4i logliklihood is %5.4f'%(n,ll))

  return F
```

The number of repetitions can be determined by ensuring that the difference $l(\hat{\mathbf{F}})$ between two consecutive iterations is less than a threshold value like $\epsilon = 0.001$. This way, we can estimate the membership matrix $\mathbf{F}$.

Now, we want to determine which cluster each individual belongs to. For this purpose, we set a threshold value δ and consider individual u as a member of cluster c if $F_{uc} > \delta$.

One approach to determining the value of δ is to set it in a way that the probability of a connection between two like-minded individuals (which is at least equal to $1 - e^{-\delta^2}$) is higher than the probability of a random connection ε (i.e., a connection that is not due to like-mindedness). Additionally, the probability of a random connection between two individuals can be determined as follows: we consider a random subset of individuals and calculate the relative frequency of pairwise connections among them.

---

theory question 8. Let's assume we want to analyze the relationship between two individuals beyond a binary state and delve deeper into it. For instance, we consider the number of shared movies between the two individuals as an integer metric representing the extent of their connection. Instead of using the Bernoulli distribution that we have worked with so far, we consider an appropriate distribution for the relationship between individuals u and v in terms of F_{uc} and F_{vc}. Rewrite the likelihood function and calculate $\nabla_{F_{u,:}} l(\mathbf{F})$ again. .

In the following, we generate a sample of 40 individuals, where the first 25 individuals are in Cluster 1 and the last 25 individuals are in Cluster 2. Then, we pass the matrix $\mathbf{A}$ to the function to determine the clusters. It can be observed that all groups have been correctly identified. (Individuals belonging to the same cluster are more likely to be connected, as indicated by the yellow color in the figure).

```
#testing in two small groups
A=np.random.rand(40,40)
A[0:15,0:25]=A[0:15,0:25]>1- 0.6     # connection prob people with 1 common group
A[0:15,25:40]=A[0:15,25:40]>1-0.1   # connection prob people with no common group
A[15:40,25:40]=A[15:40,25:40]>1-0.7 # connection prob people with 1 common group
A[15:25,15:25]=A[15:25,15:25]>1-0.8 # connection prob people with 2 common group
for i in range(40):
  A[i,i]=0
  for j in range(i):
    A[i,j]=A[j,i]

import matplotlib.pyplot as plt
import networkx as nx
plt.imshow(A)
delta=np.sqrt(-np.log(1-0.1)) # epsilon=0.1
F=train(A, 2, iterations = 120)
print(F>delta)
G=nx.from_numpy_matrix(A)
C=F>delta    # groups members
nx.draw(G,node_color=10*(C[:,0])+20*(C[:,1]))
```

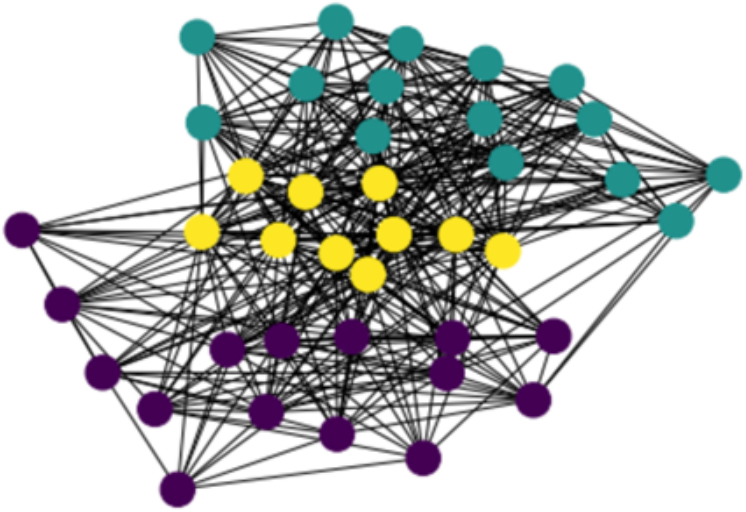
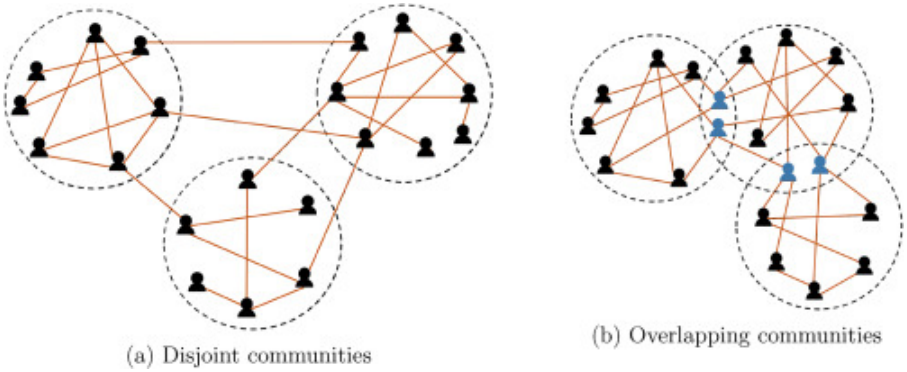


figure 4: Output of the cluster detection algorithm

3 B

In this section, we consider another probabilistic model for the relationship between individuals and their categories.

Suppose there are a limited number of tables in a restaurant, and each table has a specific capacity for accommodating people. When a person enters the restaurant, they can choose one of the tables and sit behind it. It is clear that with the limited number of tables, the choices for subsequent individuals gradually become limited. On the other hand, each person prefers to sit at a table where more of their friends are present. This means that each person has more friends at their table, but it doesn't mean they have no friendship with people at other tables.



(a) Disjoint communities

(b) Overlapping communities

figure 5: probability model of communication between people

Now we try to represent the social connections between individuals in a restaurant as a graph. For this purpose, we consider the following assumptions:

1. Let's assume that the number of people present in the restaurant is $n = 15$, and the number of tables is $k = 3$.
2. We assume that the distribution of people at different tables is homogeneous, meaning that the number of people assigned to each table is equal.
3. We represent the table number behind which each person is sitting as a vector $\mathbf{z}$. In this case, we have $\mathbf{z} \in \mathbb{R}^n$, and for each $i \in \{1, 2, \ldots, n\}$, we have $z_i \in \{1, 2, \ldots, k\}$.
4. Let $\mathbf{z}_0$ be the correct clustering assignment available to us:

$$\mathbf{z}_0 = [3, 1, 2, 1, 3, 1, 2, 2, 2, 3, 3, 2, 1, 1, 3]^\top$$

   The vector $\mathbf{z}_0$ indicates for each person which table they are sitting behind.
5. Now we need to consider the relationships between individuals. According to the initial assumption, among the people sitting at a table, there are more friends than those who are not sitting at the same table. In probabilistic terms, the probability of someone being friends with the person sitting at their table is higher than someone sitting at a different table.
6. Let's assume the probability of friendship between individuals at the same table is $p = 0.6$, and the probability of friendship between individuals who are not at the same table is $q = 0.1$.

7. Form the matrix $\mathbf{Q} \in \mathbb{R}^{k \times k}$ with the following elements:

$$Q_{i,j} = \begin{cases} p & i = j \\ q & i \neq j \end{cases}$$

where i and j are the numbers of different tables.

---

theory question 9. Based on the given information, the element $A_{i,j}$ of the adjacency matrix (you can refer back to the definition 1) represents the friendship relationships in a social network. Describe this random variable and find its probability density/mass function.

theory question 10. Do the matrix you have constructed provide a complete description of the friendships between individuals? For example, you can examine the elements on the main diagonal or consider the independence or dependence of the elements $A_{i,j}$ and $A_{j,i}$ from each other. Describe the matrix $\mathbf{A}$ with the new findings.

---

computer question 2. Create 10 samples from the matrix described in the theoretical question 10.

computer question 3. From the matrix described in the theoretical question 10, create a sample and form a friendship graph based on it. Also, label the desk number of each person on their corresponding node using color or a specific number. Make sure that the resulting graph is connected (has no isolated nodes). If you observe otherwise, reconstruct the graph.

Now we have a set of individuals and their friendship relationships are also defined. The ultimate goal is to determine the grouping of individuals (the table number each person is sitting at) based on the friendship graph. In other words, the final answer to the problem is a vector $\mathbf{z_0}$ which we assume is not available, and we want to estimate $\mathbf{z_0}$ from the elements of the graph $\mathbf{A}$, resulting in an approximation vector $\hat{\mathbf{Z}}_0 \in \mathbb{R}^k$.

Firstly, it is necessary to introduce a criterion for measuring the accuracy of the estimation. We use the Hamming distance, which is defined as follows:

$$d_H : \mathbb{R}^k \times \mathbb{R}^k \to \{1, 2, \ldots, k\}$$

$$d_H(\mathbf{z}_1, \mathbf{z}_2) = \sum_{i=1}^{k} \mathbb{1}\{[\mathbf{z}_1]_i \neq [\mathbf{z}_2]_i\}.$$

In the above definition, the function $\mathbb{1}\{.\}$ is defined for an event as follows:

$$\mathbb{1}\{B\} = \begin{cases} 1 & \text{B is true} \\ 0 & \text{B is false} \end{cases}.$$

---

computer question 4. Write a function that takes two inputs $\mathbf{z}_1$ and $\mathbf{z}_2$ and calculates the Hamming distance between them.

The Hamming distance criterion has one weakness, and that weakness is its dependency on cluster names (or label numbers). The clustering should be independent of cluster names. It doesn't matter what the name of a cluster is; what matters is the structure we have reached. In our example, if we change the names of the labels, it should not create any difference in the clustering. Whether we label the clusters as 1, 2, 3 or 1, 3, 2, the nature of the clustering remains unchanged; only the assigned numbers have changed. In fact, the measure of estimating accuracy should be independent of different permutations of labeling and should not change with its alteration. Now we need to separate the Hamming distance from the cluster names. For example, in a community with $n = 6$ members and $k = 2$ groups, let's assume the correct clustering is as follows:

$$\mathbf{z}_0 = [2, 2, 1, 2, 1, 1]^\top.$$

And let's say we approximate $\mathbf{z}_0$ with the following estimation:

$$\hat{\mathbf{Z}}_0 = [1, 1, 2, 1, 2, 2]^\top.$$

If we use the function mentioned in simulation question 4 to calculate the distance between these two vectors, we would have $d_H(\hat{\mathbf{Z}}_0, \mathbf{z}_0) = 6$. However, if we swap the labels 0 and 1 in the vector $\hat{\mathbf{Z}}_0$ and call the resulting vector $\tilde{\mathbf{Z}}_0$, we would have $d_H(\tilde{\mathbf{Z}}_0, \mathbf{z}_0) = 0$.

For a vector $\mathbf{z}$, we define the set of vectors $\langle \mathbf{z} \rangle$ as the set of all vectors that can be obtained by rearranging the cluster labels in $\mathbf{z}$. For example, in a community with $n = 6$ members and $k = 3$ different clusters, consider the following vector:

$$\mathbf{z} = [3, 1, 2, 2, 3, 1]^\top.$$

We can consider other permutations of the cluster names and form the set $\langle \mathbf{z} \rangle$:

$$\langle \mathbf{z} \rangle = \{[3, 1, 2, 2, 3, 1]^\top, [3, 2, 1, 1, 3, 2]^\top, [1, 3, 2, 2, 1, 3]^\top, \dots\}.$$

Therefore, we define the independent of cluster names distance function as follows:

$$d : \mathbb{R}^k \times \mathbb{R}^k \to \{1, 2, \dots, k\}$$
$$d(\mathbf{z}_1, \mathbf{z}_2) = \min_{\tilde{\mathbf{z}}_2 \in \langle \mathbf{z}_2 \rangle} d_H(\mathbf{z}_1, \tilde{\mathbf{z}}_2).$$

computer question 5. Write a function that calculates the independent distance of cluster names. We call this distance the minimum Hamming distance.

Now we want to estimate the vector $\mathbf{z}_0$ using the maximum likelihood estimation based on an adjacency matrix realization.

theory question 11. Considering the theoretical questions 9 and 10, how many independent elements exist in the matrix $\mathbf{A}$?

theory question 12. Write the probability of the realization of matrix $\mathbf{A}$ given the vector $\mathbf{z}$. The expression you have reached is the likelihood function, which we denote as $L(\mathbf{z}) = \mathbb{P}[\mathbf{A}|\mathbf{z}]$.

theory question 13. Compute the logarithm of the likelihood function, which is defined as $l(\mathbf{z}) = \log\big(L(\mathbf{z})\big)$.

In the following, we attempt to maximize $l(\mathbf{z})$. This is equivalent to minimizing $\tilde{l}(\mathbf{z}) = -l(\mathbf{z})$.

---

computer question 6. Write a function that takes $\mathbf{z}$ and $\mathbf{A}$ as inputs and calculates the value of the expression $\tilde{l}(\mathbf{z}) = -\log(\mathbb{P}[\mathbf{A}|\mathbf{z}])$.

Finding the optimal solution to a problem is not an easy task (read the section title and its footnote again!). Therefore, we need to use numerical methods to minimize it. First, let's assume that we know the number of clusters and the number of members in each cluster. In other words, we know that each element of the vector $\mathbf{z}_0 \in \mathbb{R}^n$ is a number from the set $\{1, 2, 3\}$, and exactly 5 elements of this vector are equal to 1, 5 elements are equal to 2, and 5 elements are equal to 3. Hence, the only thing we don't know is the exact permutation of these elements, which we want to estimate using the help of an adjacency matrix. For this purpose, we use the following algorithm:

1. First, define an initial estimate of $\mathbf{z}_0$ as follows:

$$\hat{\mathbf{z}}_0 = [1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3]$$

2. Calculate the value of $\tilde{l}(\hat{\mathbf{z}}_0)$.
3. For $t = 1, \ldots, T$:

(a) For $i = 1, \ldots, n$: - Replace $[\hat{\mathbf{z}}_{t-1}]_i$ with each element of $\hat{\mathbf{z}}_{t-1}$ one by one. At each step, calculate the value of the function $\tilde{l}$ for the newly obtained vector.

(b) Apply the permutation that results in the greatest reduction in $\tilde{l}(\hat{\mathbf{z}}_{t-1})$ to the vector $\hat{\mathbf{z}}_{t-1}$ and obtain the vector $\hat{\mathbf{z}}_t$.

4. Declare $\hat{\mathbf{z}}_T$ as the output vector.

---

computer question 7. Simulate the above algorithm. At each stage,save the rate of $\tilde{l}(\hat{\mathbf{z}}_t)$ and also $d(\hat{\mathbf{z}}_t, \mathbf{z}_0)$ and finally display on the chart.You are free in the choice of parameter T , but choose the right number.

computer question 8. Run the algorithm for $N = 10$ different starting points $(\hat{\mathbf{z}}_1)$ and compare the results (you can read the title of this section and its footnote again!).

computer question 9. In the previous question, you obtain different outputs when $N = 10$. We represent these outputs as $\{\hat{\mathbf{z}}_T^{(j)}\}_{j=1}^N$. Compare the value of $\tilde{l}(\hat{\mathbf{z}}_T^{(j)})$ for different values of j with $\tilde{l}(\mathbf{z}_0)$. Is there any case where $\tilde{l}(\hat{\mathbf{z}}_T^{(j)}) = \tilde{l}(\mathbf{z}_0)$? In this case, what is the minimum Hamming distance between the estimated vector and $\mathbf{z}_0$?

computer question 10. Is there any j such that $d(\hat{\mathbf{z}}_T^{(j)}, \mathbf{z}_0) = 0$?

computer question 11. Generate two additional samples of matrix A based on $\mathbf{z}_0$. In each of them, try to estimate $\mathbf{z}_0$ using different initial vectors for $N = 10$. Report the best result.

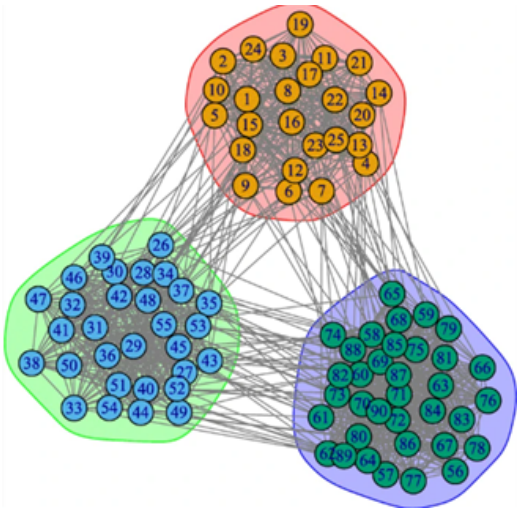


figure 6: An example of clustering people

4 C

figure 7: An example of clustering people

First, let's assume there are only two movie genres, and each person is a fan of exactly one of these two genres. Furthermore, similar to the previous section, we assume that the probability of having a connection (or taste similarity) between fans of the same genre is equal to p, and the probability of having a connection between fans of different genres is equal to q. It is logical to assume that $p > q$, which means the probability of taste similarity between two individuals who are fans of the same genre is higher than the probability of taste similarity between two individuals who are fans of different genres.

With these assumptions, the adjacency matrix (you can refer back to the definition in 1!) is a random matrix, meaning that each of its entries is a random variable.

theory question 14. What random variable is the element $A_{i,j}$ of the matrix $\mathbf{A}$? Calculate its probability density/mass function.

theory question 15. Consider the matrix $\mathbf{W} = \mathbb{E}[\mathbf{A}]$. The elements of this matrix are defined as follows:

$$W_{i,j} = \mathbb{E}[A_{i,j}].$$

Calculate the matrix $\mathbf{W}$.

We can write: $\mathbf{A} = \mathbf{W} + \mathbf{R}$, where $\mathbf{R}$ represents the noise in the observation of $\mathbf{W}$. First, we assume that we have the matrix $\mathbf{W}$ and we attempt to perform clustering using this matrix.

theory question 16. First, let's assume that $n = 4$ and individuals 1 and 2 are fans of genre 1, while individuals 3 and 4 are fans of genre 2. In this case, let's find the eigenvalues and eigenvectors of the matrix $\mathbf{W}$.

theory question 17. In general, show that $\mathbf{W}$ has only two non-zero eigenvalues and calculate the corresponding eigenvectors for these eigenvalues.

theory question 18. Matrix $\mathbf{D_W}$ is considered as a diagonal matrix, where the elements on its diagonal are defined as follows:

$$[D_{\mathbf{W}}]_{i,i} = \sum_{j=1}^{n} W_{i,j}$$

Assuming $\mathbf{W}$ that you have calculated in previous sections, calculate $\mathbf{D_W}$.

theory question 19. The matrix $\mathbf{L_W}$ is defined as $\mathbf{L_W} = \mathbf{D_W} - \mathbf{W}$. Calculate the eigenvalues and eigenvectors of this matrix.

theory question 20. Two eigenvectors corresponding to two smaller eigenvalues of the matrix are considered. In the same simple case where $n = 4$ as mentioned in theoretical question 16, represent the data using these two eigenvectors in a two-dimensional space. In other words, if we denote the two eigenvectors as $\mathbf{u}_1$ and $\mathbf{u}_2 \in \mathbb{R}^n$, consider the point $([u_1]_i, [u_2]_i)$ for the i-th individual. Can individuals be classified using these points?

The algorithm you have observed step-by-step in previous questions is called spectral clustering. As you have seen, if we have the matrix $\mathbf{W}$, we can perform clustering without error. However, in reality, we only have access to the matrix $\mathbf{A}$, which is a noisy version of the matrix $\mathbf{W}$. If we want to replicate the process exactly for the matrix $\mathbf{A}$, we define the matrices $\mathbf{D_A}$ and $\mathbf{L_A}$ and perform the steps mentioned in previous questions, clustering will be performed based on the eigenvectors of the matrix $\mathbf{L_A}$. Fortunately, it can be shown that if the size of the community is sufficiently large, the eigenvectors of the matrix $\mathbf{L_A}$ will be close to the eigenvectors of the matrix $\mathbf{L_W}$.

computer question 12. First, we try to observe the performance of this algorithm through simulation. We consider the number of individuals as $n = 10000$. Additionally, we assume $p = 0.1$ and $q = 0.01$. Generate matrices $\mathbf{A}$ and $\mathbf{W}$ and run the algorithm on them. Compare and report the results of the two algorithms, i.e., the error rate of the algorithm in detecting the clusters in two scenarios. Note that the assignment of individuals to different groups should be done randomly, and it is obvious that a unit assignment should be used to construct matrices $\mathbf{W}$ and $\mathbf{A}$.

theory question 21. It can be shown that the following inequality holds with at least a probability of $1 - 4e^{-n}$:

$$\exists \theta \in \{-1, 1\} : \qquad \sum_{j=1}^{n} |[u_2(\mathbf{L_W})]_j - \theta[u_2(\mathbf{L_A})]_j|^2 \leq \frac{C}{\mu^2}$$

subject to the eigenvalues being arranged in ascending order. In the above expression, $\mathbf{u}_2(\mathbf{L})$ represents the eigenvector corresponding to the second eigenvalue of matrix $\mathbf{L}$, index j denotes the j-th element of the vector, and $\mathbf{L_A} = \mathbf{D_A} - \mathbf{A}$. Additionally, $\mu = \min\{q, p-q\}$ is a constant value.

Using the inequality above, show that the spectral clustering algorithm performs at most a constant number of errors with a probability greater than $1 - \epsilon(n)$ (where $\epsilon(n)$ is a function of n that you need to compute!). Determine the complexity of this algorithm in terms of the minimum required number of data points, n, to achieve a probability of $1 - \epsilon$.

computer question 13. Simulate the question 12 with several different values of n and report and compare the number of errors.

As you can see, the spectral clustering algorithm helps us represent the data in a new coordinate system, where different clusters are well separated. So far, we have only examined the binary case. However, it is possible to generalize this algorithm to any desired number of clusters.

theory question 22. In relation to the spectral clustering algorithm in the case of $k \geq 2$, conduct research and summarize your findings in a report while preserving the LaTeX template.

You should write a general overview of the algorithm and provide a brief explanation of the relevant theory.

We have observed that the spectral clustering algorithm with k clusters can effectively transform the data into a new coordinate system with k dimensions, where the clusters are relatively well-separated. In this case, we cannot visualize the data directly (as their dimensions may exceed 3), so we need to use an algorithm that can handle high-dimensional data and extract clusters. One well-known algorithm for this purpose is the k-means algorithm. The details of this algorithm are provided in the course manual for further reference. To use this algorithm, you can utilize the `sklearn.cluster.KMeans` library. Clicking on the library title will direct you to the documentation page for this library.

computer question 14. The Calfornia Housing dataset, which is provided in the attached files, contains the coordinates of several houses and the average income of residents in each house. We want to cluster the houses based on the average income parameter MedInc. To do this, we define the distance between two houses i and j as $d(i,j) = |\text{MedInc}_i - \text{MedInc}_j|$. Let's consider $k = 3$ as the number of clusters and display the clustering result on a geographical map of the houses. Show houses belonging to one cluster with a different color from the other clusters. Please note that to ensure convergence of the algorithm, you need to increase the number of iterations sufficiently.

So by combining the spectral clustering algorithm and the algorithm k-means clustering can be done for the number of categories you want.

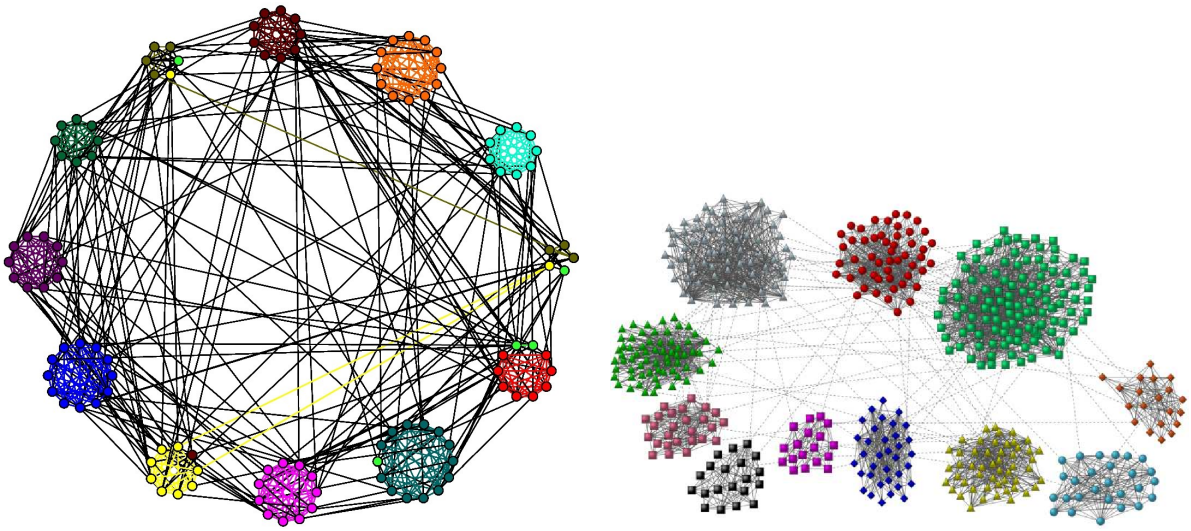


figure 8: Example of clustering people with the help of spectral clustering algorithm and method k–means

computer question 15. Consider the dataset of "Zachary's Karate Club." This dataset represents the interactions between individuals in a karate club. Using spectral clustering algorithm along with k-means, classify this network and display the results for the number of clusters $k = 2, 3, 4$ on three separate graphs.

computer question 16. One of the evaluation criteria for assessing the quality of clustering and community detection is the modularity measure of the identified clusters in a graph. The definition of this parameter for a cluster is as follows:

$$f(S) = \frac{c_S}{2m_S + c_S}$$

where S represents the identified cluster, and m_S and c_S are defined as follows:

$$c_S = |\{(u, v) \in E : u \in S, v \notin S\}|,$$

$$m_S = |\{(u, v) \in E : u \in S, v \in S\}|.$$

In the above definitions, E represents the set of all edges of the graph, and the symbol $|B|$ for a set like B denotes the number of elements in the set B. The average modularity of the identified clusters in the simulation question 15 should be calculated for $k = 2, 3, \ldots, 10$ and plotted on a graph. Based on the graph, how many suitable clusters are there for this network?

5 D

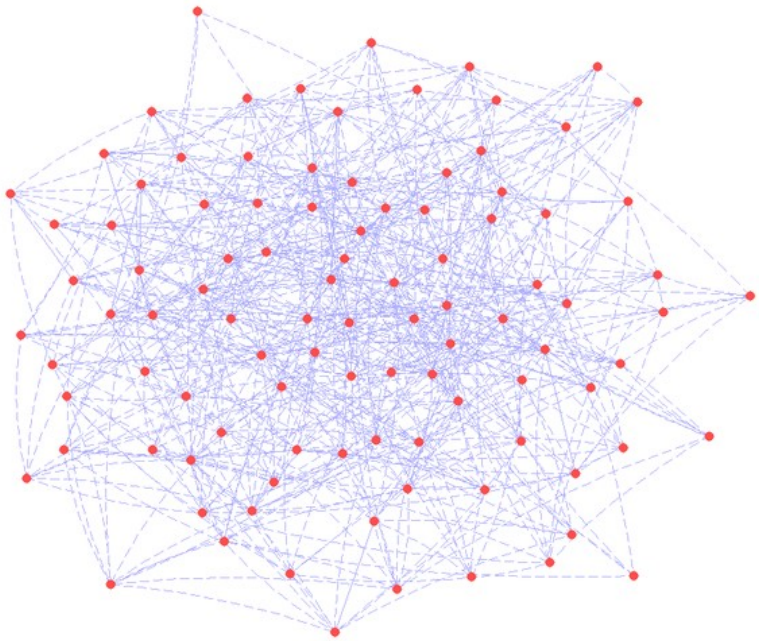


figure 9: An example of relationships within a cluster

In the previous section, we saw that it is possible to model the user graph of Netflix/Disney+/Vimeo with clusters, where the probability of taste similarity between any two individuals within a cluster is equal to p, and the probability of taste similarity between individuals from different clusters is equal to q, where $q < p$.

As we mentioned, within each cluster (e.g., the cluster of users who like comedy movies), the probability of taste similarity between two individuals is p. For this reason, users within each cluster can be modeled using a random graph. Now, in this problem, we focus on the relationships between individuals within one (and only one) of these clusters. A cluster of the community consists of a graph with n vertices, where there is an edge between any two vertices with a probability of p and there is no edge with a probability of $1 - p$.

---

theory question 23. Let's assume there is a friendship relationship among individuals in a cluster under consideration, denoted by m. Based on the random graph model we described, we create friendship relationships between individuals randomly. What is the probability that we have correctly determined all the homophilic relationships? Your answer should depend on n, p, and m.

theory question 24. In this question, let's assume we know the exact value of m, and as a result, we establish the relationship of homophily between these n individuals randomly. Find the probability of correctly determining all the homophily relationships in terms of n and m.

theory question 25. Let's find the probability that we have correctly determined 20% of the intertaste relationships among these n people.

computer question 17. To $p = 0.0034$, $n = 1000$ and $m = 3000$, write a program that repeats the allocation of affinity relationships $N = 10$ times. Each time, save the number of affinity relationships, and at the end, calculate the average of all the obtained values. Is this average approximately (with a maximum error of 5%) equal to m?

theory question 26. For a given n and p as stated in the simulation 17, find this average value theoretically. In general, what relationship should exist between n, p, and m for this average value to be approximately equal to m?

Most likely, you can find people among your acquaintances who have specific tastes. This means that despite their interest in a particular genre (such as comedy), they may still enjoy movies from that genre that many fans of that genre do not like, and vice versa. In this project, we call these individuals "embarrassed"! (You won't be embarrassed, just join the group!)

First, let's define the concept of an embarrassed person.

definition 2 (Embarrassed and Like-minded Person)**.** In a cluster, if each individual on average has a taste similarity of L, then we consider a person embarrassed if they have less than L taste similarity, and we call them like-minded if they have more than L taste similarity.

computer question 18. For $n = 1000$ and $p = 0.00016$, write a program that, with 10 repetitions, calculates the average number of people with the same color. Additionally, to have a better understanding of the distribution of the number of preferences of an individual, plot a graph where the horizontal axis represents the number of preferences and the vertical axis represents the average number of people who have that number of preferences.

theory question 27. For a given n and p stated in question 18, on average, how many acquaintances does each person have?

theory question 28. For given values of n and p as stated in question 18, if we randomly select a person, what is the probability that they are of the same color? Additionally, find the expected value of the number of people of the same color.

Now we want to examine triads of individuals and their co-preference relationships within clusters.

definition 3 (Transitivity Property)**.** In the co-preference relationship between three individuals, A, B, and C, the transitivity property holds if, whenever A has a co-preference with B and B has a co-preference with C, then there is also a co-preference relationship between A and C.

definition 4 (Chain Property)**.** In the co-preference relationship between three individuals, A, B, and C, the chain property holds if A has a co-preference with B and B has a co-preference with C, but there is no co-preference relationship between A and C.

computer question 19. Write a program that, after 5 iterations of randomly assigning affinity relationships between $n = 3000$ individuals with a probability of $p = 0.01$, calculates the average number of affinity relationships with transitivity property and the average number of affinity relationships with chain property.

theory question 29. Calculate the expected number of transitive relationships and the expected number of chain relationships in a cluster.

theory question 30. In a cluster, what fraction of the total number of homophily relationships between three individuals, assuming each of these three individuals is homophilic with at least one of the other two individuals, have the transitive property? Do your simulations agree with the theoretical calculation? Analyze the result.

computer question 20. For a cluster with $n = 1000$ and $p = 0.003$, calculate the average number of homophily relationships among individuals using simulation.

figure 10: An example of relationships within a cluster

theory question 31. Find the expected number of connections between individuals in a cluster with n vertices and probability p of mutual liking. (A closed-form solution is required!)

Until now, we have investigated the properties of the corresponding graph to the relationships of homophily in a cluster of individuals, meaning individuals who are interested in a specific genre of movies. We denote this graph, which has n vertices and each pair of vertices is connected with probability p, as $\mathcal{G}(n,p)$.

According to the "small world" law, any two individuals in the world are likely to know each other with at most 6 intermediaries. For example, if we randomly select two individuals A and B and examine the set of friends of A, friends of friends of A, friends of friends of friends of A, and so on, we can reach individual B within a maximum of 6 steps.

In this section, we want to investigate the existence of this property in a cluster (where the relationships of homophily are randomly arranged, unlike the real world).

First, we will examine the average distance between two individuals in a cluster. It should be noted that the distance between two individuals is the minimum number of homophily relationships connecting them.

---

computer question 21. Write a program to calculate the average distance between two individuals in $\mathcal{G}(n,p)$ for $n = 1000$ and $p = 0.0033$.

Now we turn to examining the maximum distance between two individuals in $\mathcal{G}(n,p)$ (which we refer to as the graph's diameter).

---

computer question 22. Assuming $n = 50$ and $p = 0.34$, generate $\mathcal{G}(n,p)$ randomly 100 times. Each time, find the pair of users with the maximum distance from each other in the graph. Calculate the average maximum distance between two users over these 100 graphs.

computer question 23. With the constant (and equal to the value mentioned in simulation question 22), keep p (the number of vertices) in the range of $[10, 200]$ with a step size of 10 units, and for each n, repeat simulation question 22. Finally, plot the average maximum distance between pairs of individuals (which is averaged over 100 different samples of $\mathcal{G}(n,p)$ with similar specifications) as a function of n. What is the shape of this plot? How does this plot behave as n increases?

---

theory question 32. For two vertices u and v in $\mathcal{G}(n,p)$, where $\mathcal{G}(n,p)$ is a random graph model, we define the random Bernoulli variable $I_{u,v}$ as follows:

$$I_{u,v} = \begin{cases} 0 & \text{if vertices } u \text{ and } v \text{ have a common neighbor} \\ 1 & \text{if vertices } u \text{ and } v \text{ do not have a common neighbor} \end{cases}.$$

Calculate $\mathbb{P}[I_{u,v} = 1]$.

theory question 33. For the graph $\mathcal{G}(n,p)$, the random variable X_n is defined as "the number of pairs of vertices in the graph that have no common neighbor." Find $\mathbb{E}[X_n]$.

theory question 34. By using the Markov's inequality, find the upper bound for $\mathbb{P}[X_n \geq 1]$. Then, investigate the behavior of this bound by taking n towards infinity.

theory question 35. Observe that when n is very large, the diameter $\mathcal{G}(n,p)$ has a high probability of having an upper bound, and also determine the value of this upper bound. Is the graph diameter dependent on large values of n and p? Does the result obtained from theoretical proof match the simulation result?

---

computer question 24. Graph $\mathcal{G}(n,p)$ with $n = 100$ and $p = 0.34$. Generate this graph randomly 100 times and count the number of triangles in each instance. Calculate the average number of triangles in these 100 graphs.

computer question 25. Number of vertices (n) in the range $[10, 100]$ and change it with a step of 10. For each n, p as a function of n is equal to

$$p(n) = \frac{60}{n^2}$$

For each n, calculate the average number of 3-person friendship loops using the simulation-based approach in Question 24 and plot this average as a function of n on a graph.
Does the average tend towards a specific number as n increases? How do you justify this behavior?

computer question 26. Repeat the simulation question 25 with $p = 0.34$. Does the average number of friendship loops converge to a specific value as n increases?

computer question 27. This time, use $p = \frac{1}{n}$. Change n in the range $[50, 1200]$ with a step of 50. Again, plot the average number of friendship loops as a function of n. Also, plot the cumulative mean of this graph. Does it tend towards a specific number?

6 E

In this section, we become familiar with the random walk algorithm for finding groups with similar preferences.

Consider an n-vertex graph G and its adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. The degree of a vertex in the graph $\mathcal{G}$, denoted by $d(i)$, is equal to the number of edges connected to vertex i. Therefore, it is obvious that: $d(i) = \sum_j A_{ij}$.

In the process of random walking, at each step, we randomly and uniformly (read the title of this section again!) move from one vertex to another in its neighborhood.

---

theory question 36. The probability of transitioning from vertex i to vertex j in one step is denoted as $P_{i,j}$.

We can place the calculated $P_{i,j}$ values together and obtain the transition matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$. Additionally, we consider the degree matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$ as a diagonal matrix with diagonal elements $D_{ii} = d(i)$.

---

theory question 37. Write the matrix $\mathbf{P}$ in terms of matrices $\mathbf{A}$ and $\mathbf{D}$.

theory question 38. Considering the matrix multiplication relationship, express the element i, j of the matrix $\mathbf{P}^2$ (denoted as $[\mathbf{P}^2]_{i,j}$) in terms of the matrix powers of $\mathbf{P}$. Can you provide a probabilistic interpretation of the elements of matrix $\mathbf{P}^2$?

theory question 39. Find an expression for the probability of reaching from vertex i to vertex j in t steps, denoted by $P_{i,j}^{(t)}$, while maintaining the LaTeX template.

theory question 40. What is the relationship between $P_{i,j}^{(t)}$ and $P_{j,i}^{(t)}$? Prove that the ratio of these two probabilities depends only on the degrees of vertices i and j.

The main idea of the random walk method is that if a random walker starts moving on a network graph, after a while, with a high probability, the transitions occur only between a subset of nodes that correspond to the desired clusters of similarity. In other words, the random walker gets trapped in a cluster because the number of edges between the members of a cluster is much higher compared to the edges between the members of a cluster and other nodes in the graph.

---

theory question 41. Please describe intuitively what characteristics $P_{ij}^{(t)}$ should have if user i and user j have similar tastes. In that case, what can be said about the probabilities $P_{ik}^{(t)}$ and $P_{jk}^{(t)}$, where user k is an arbitrary person in the network?

theory question 42. To categorize users based on their taste, we need a measure of similarity. We define the "taste difference between two users i and j" as follows:

$r_{ij} = sqrt(sum((P_ik^(t) - P_jk^(t))^2/d(k) for k = 1 to n))$

Now let's consider a category of users with similar tastes and denote it as C. Calculate the probability that a random walker starts from one of the vertices associated with users in this category and reaches a desired user k after t steps. Then, similar to the above equation, define the "taste difference between two categories $C1$ and $C2$".

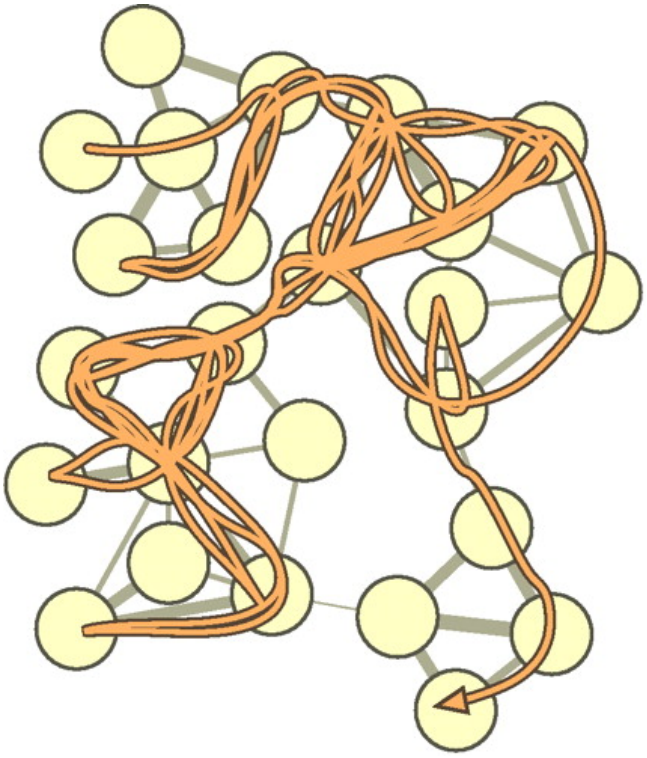


Figure 11: An example of random walk

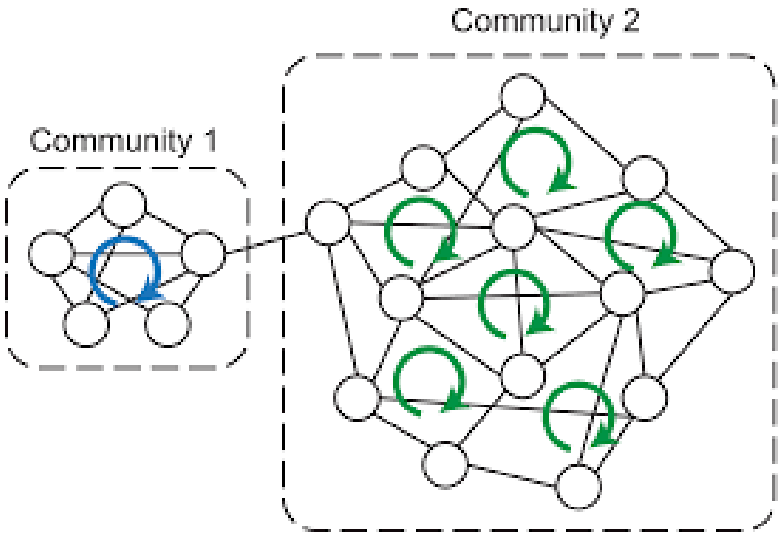


Figure 12: Trapped random walker in clusters

Now we can propose an algorithm for classifying users with similar tastes:

1. Consider each user as a separate category and calculate the taste difference between each pair of users.
2. While the number of categories is greater than 1:
   (a) Select two categories with the smallest taste difference.
   (b) Merge them together ($C_3 = C_1 \cup C_2$).
   (c) Update the taste difference between categories.

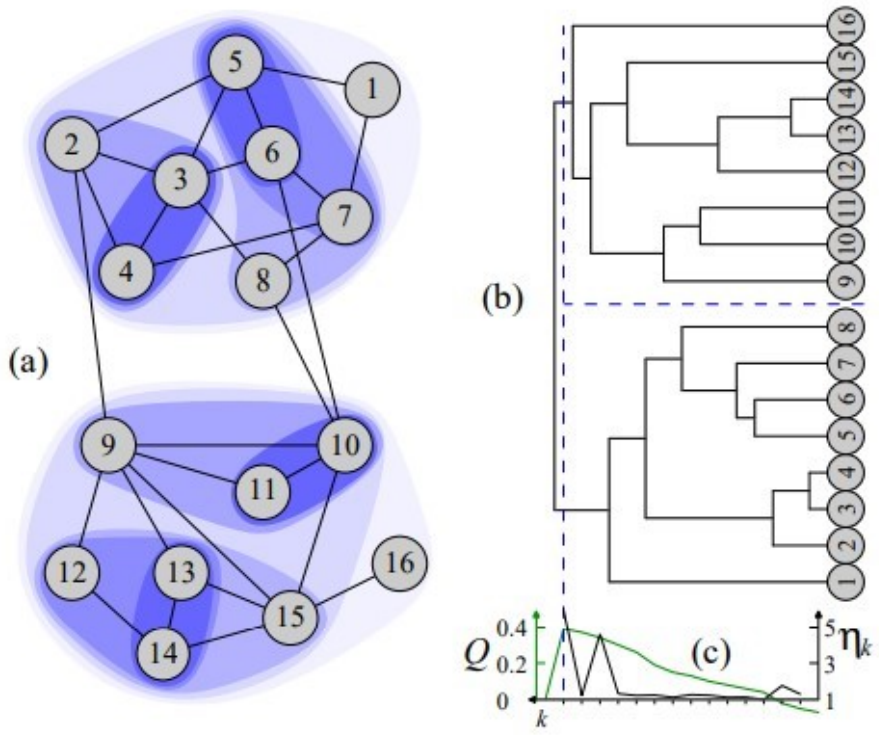


Figure 13: Clustering algorithm for individuals based on random walk

theory question 43. Provide a criterion that indicates at which stage of the algorithm the classification has been performed optimally.

computer question 28. Implement the described algorithm on the dataset *Zachary's Karate Club*. Consider $t = 2$ steps to define the distances.

computer question 29. Consider $t = 5$ steps to define the distances and repeat the simulation question 28.

7 Important Points!

Please pay attention to the following points:

1. This project will contribute to a part of your grade in this course.
2. You can work on the project in groups of 2 or 3. A form will be provided for you to register your groups. Be careful that all group members must be proficient in all sections of the project and ultimately, all members of a group will receive the same grade.
3. The titles of different sections of the project are selected from the works of Persian poets and literary figures. These poems are not related to the concepts you will encounter in each section.
4. All simulations should be done using the Python language. You are only allowed to use the libraries: `networkx`, `numpy`, `scipy`, `random`, and `matplotlib`. Additionally, the use of the library `scikit-learn` is allowed only in the cases explicitly mentioned. If you click on the title of each library, you will be directed to its documentation.
5. In this project, we will use two datasets. The descriptions and instructions for obtaining these datasets are provided in the file "Dataset.txt."
6. In the file "kmeans.pdf," there are some explanations about the k-means algorithm for your optional study. You can also read the file "linear-algebra-prerequisites.pdf" for a better understanding of linear algebra. However, in this project, you will not need advanced linear algebra, and basic operations such as matrix multiplication and calculating eigenvalues of matrices will be sufficient.
7. The project should be submitted as a report along with the written codes. The report should include answers to the questions, images, plots, and necessary conclusions. Note that the main part of the project's weight lies in your report and the results you obtain from the code. Also, the cleanliness of the report is very important. Upload the codes and the report in a compressed file on the course platform.
8. If you have used any external sources (books, articles, websites, etc.) to answer the questions, be sure to provide proper references.
9. Writing the report using LaTeX will earn you bonus points.
10. Simulation questions are marked with the color green and theoretical questions are marked with the color blue.
11. You can write the theoretical parts of the report on paper and include their images in your report, but I kindly recommend not to do so!
12. In case of cheating, both individuals will receive a zero grade.

Good luck!