In the name of god

probability and Statistics Project

methods of clustering graphs

phase2

Course thacher

Dr. Mohammad Hossein Yasai

Dead line:

2022

1 A

1.1 Watts–Strogatz Random Graph Model

In the fifth section of the first phase of the project, we became familiar with random graphs and observed that according to the small-world phenomenon, any two individuals in the world are connected to each other with approximately one intermediary, with a maximum of 6 degrees of separation.

definition 1 (Average Distance)**.** The average number of edges traversed in the shortest path between any two vertices in a graph is called the average distance of that graph.

definition 2 (Clustering Coefficient)**.** The clustering coefficient of a vertex is defined as the ratio of the number of edges between its neighboring vertices to the total number of possible edges among those neighbors in a complete graph (i.e., a graph where all edges exist).

definition 3 (Average Clustering Coefficient)**.** The average clustering coefficient of the vertices in a graph is called the average clustering coefficient of that graph.

Let's consider the following graph model:

We place n vertices on the circumference of a circle. We connect each vertex to its closest $2m$ neighbors. The resulting graph is called $\mathcal{G}_{\mathrm{WS}}(n, m)$.

For example, for $n = 25$ nodes and parameter $m = 2$, we will obtain such a graph:

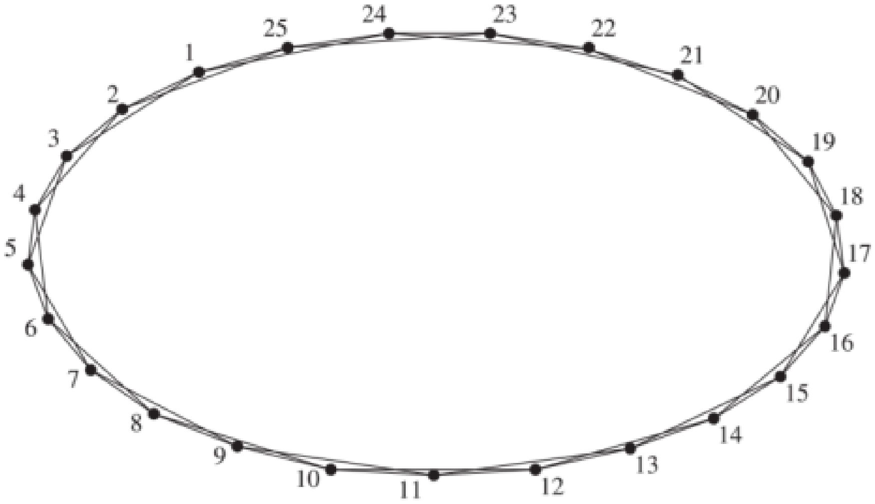

figure 1: an example of model Watts-Strogatz

theory question 1. To find the average distance of the graph $\mathcal{G}_{\mathrm{WS}}(n,m)$,

theory question 2. Prove that in an arbitrary graph, denoted by $\mathcal{G} = (V, E)$, where $V = \{v_i\}_{i=1}^{n}$ is the set of vertices and E is the set of edges of $\mathcal{G}$, the clustering coefficient of vertex v_i, denoted by C_i, is given by the following equation:

$$C_i = \frac{2|\{e_{ij} : v_i, v_j \in N_i, e_{ij} \in E\}|}{k_i(k_i - 1)}.$$

In the above equation, N_i represents the set of neighboring vertices of vertex v_i, e_{ij} is an edge connecting vertices v_i and v_j, and k_i is the number of neighbors of vertex v_i.

theory question 3. Prove that the average clustering coefficient in the graph $\mathcal{G}_{\mathrm{WS}}(n,m)$ is equal to $\frac{1}{2}$.

As seen, the average distance between the nodes in the graph $\mathcal{G}_{\mathrm{WS}}(n,m)$ is high. To reduce the average distance, we need to add some random properties to the initial graph $\mathcal{G}_{\mathrm{WS}}(n,m)$. We do this as follows:

For each vertex $v_i \in V$, we consider the edges that connect v_i to m vertices on its right side. According to the construction method of the graph $\mathcal{G}_{\mathrm{WS}}(n,m)$, the number of these edges is equal to m. We remove each of these edges with probability p and instead, we draw a completely random edge from vertex v_i to one of the vertices that is not connected to v_i through an edge. We call the value p the rewiring probability. As seen in Figure 2, as the value of p increases, this graph becomes closer to a random graph. As the graph approaches a random graph, the average distance decreases, but at the same time, the average clustering coefficient also decreases. However, the good news is that the rate of decrease for these two parameters is different.

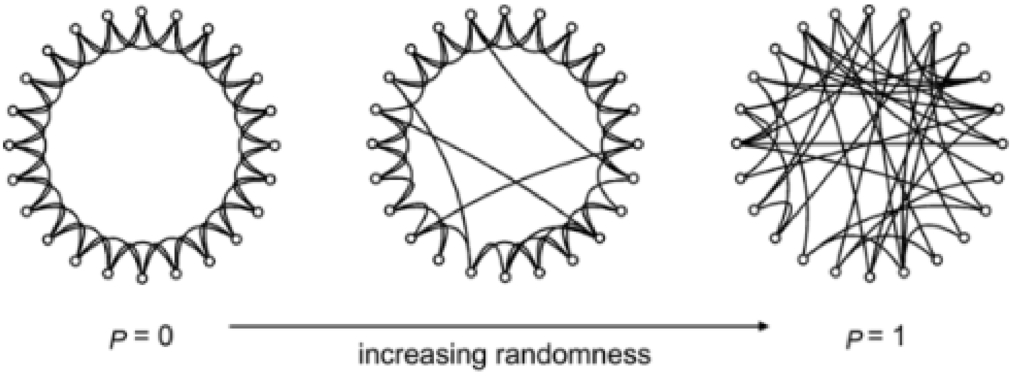

figure 2: The process of adding random properties to $\mathcal{G}_{\mathrm{WS}}(n,m)$

As we can see in figure 3, the rate of decrease in the average clustering coefficient is less than the rate of decrease in the average distance. Therefore, if we adjust the value of p in such a way that the average clustering coefficient does not decrease significantly, but the average distance decreases sufficiently, we can establish the small-world phenomenon for graphs $\mathcal{G}_{\mathrm{WS}}(n,m)$. For example, in figure 3, if $p \approx 0.01$ is chosen, we have achieved our goal to a good extent.

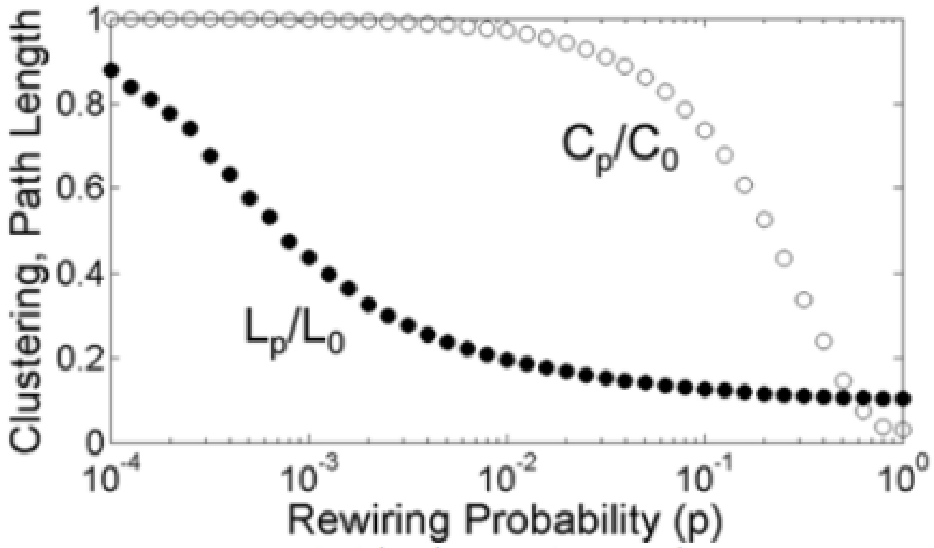

figure 3: An example of the process of reducing the mean distance and the mean clustering coefficient by increasing the random property in the graph $\mathcal{G}_{\mathrm{WS}}(n, m)$

computer question 1. Assume that Netflix/Vimeo/Disney+ have $n = 10000$ users. If we assume $m = 400$, construct the graph $\mathcal{G}_{\mathrm{WS}}(n, m)$.

computer question 2. For different values of p in the interval $[10^{-6}, 1]$, add a random property to the graph in question 1 and plot a graph similar to Figure 3.

computer question 3. Using the graph plotted in question 2, approximately find the optimal value of p (denoted as p^*) and examine the small-world property before and after adding randomness to the graph $\mathcal{G}_{\mathrm{WS}}(n, m)$ with a rewiring probability of $p = p^*$.

1.2 Random model graph Configuration

n this section, we examine another type of random graphs. Let's assume that the sequence $\mathbf{d} = (d_1, d_2, ..., d_n)$ represents the degrees of the vertices of an n-vertex graph. We form the following sequence of vertices:

$$\mathbf{a_d} = (\underbrace{1, 1, \ldots, 1,}_{d_1 \text{ entries}} \quad \underbrace{2, 2, \ldots, 2,}_{d_2 \text{ entries}} \; \ldots, \; \underbrace{n, n, \ldots, n}_{d_n \text{ entries}})$$

We construct the graph $\mathcal{G}_{\mathrm{C}}(n, \mathbf{d})$ as follows:

From the sequence $\mathbf{a_d}$, we randomly and without replacement select two numbers and connect the corresponding vertices. We continue this process until all elements of the sequence $\mathbf{a_d}$ are exhausted.

For example, let's assume $\mathbf{d} = (3, 4, 3)$. In this case, we have:

$$\mathbf{a_d} = (1, 1, 1, 2, 2, 2, 2, 3, 3, 3).$$

Now, let's suppose the randomly chosen numbers are:

$$(1,2),(2,3),(1,3),(2,2),(3,1).$$

In this case, the graph $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$ will look like Figure 4.

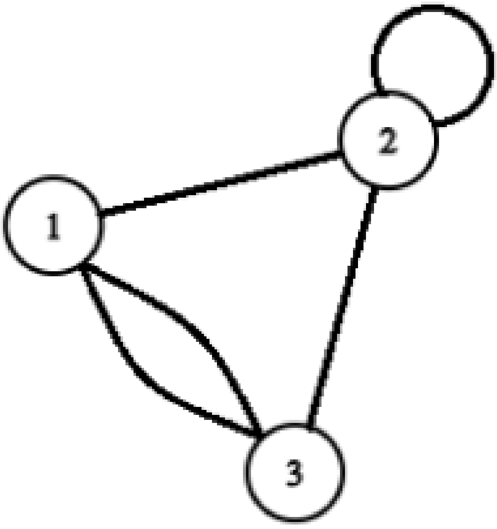

figure 4: example of a graph $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$

definition 4 (Loop)**.** A edge that is connected to itself is called a loop.

definition 5 (Multiple Edge)**.** If there are more than one edge between two vertices, then we say there is a multiple edge between those two vertices.

You have seen the definition of adjacency matrix in the first phase of the project. For graphs that have loops and multiple edges, the adjacency matrix is defined as follows:

definition 6 (Adjacency Matrix for Graphs with Loops and Multiple Edges)**.** The adjacency matrix for a graph $\mathcal{G}$ with n vertices is a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ whose elements are defined as follows:

$$A_{i,j} = \text{the number of edges between vertices } i \text{ and } j.$$

From definition 6, it is clear that $A_{i,i}$ is equal to the number of loops at vertex i.

theory question 4. Prove that:

$$d_i = A_{i,i} + \sum_{j=1}^{n} A_{i,j}.$$

theory question 5. If we denote the number of edges in the graph $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$ as m, prove that $m = \frac{1}{2}\sum_{i=1}^{n} d_i$.

theory question 6. Show that if the sequence $\mathbf{d}$ is given, the probability of observing the graph $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$ with the adjacency matrix $\tilde{\mathbf{A}}$ is equal to:

$$\mathbb{P}[\mathbf{A}=\tilde{\mathbf{A}}] = \frac{1}{\prod_{i=1}^{m}(2i-1)} \frac{\prod_{i=1}^{n}(d_i!)}{\Big(\prod_{i=1}^{n} 2^{\tilde{A}_{i,i}}\Big)\Big(\prod_{1\le i<j\le n}(\tilde{A}_{i,j}!)\Big)}$$

computer question 4. Assuming $n = 20000$, the user is interested in the drama genre and each person exactly shares the same taste with $k = 100$ other users. Let's model the users of this genre with the Configuration model. If we randomly select an edge, what is the probability that it is not a clique or a complete graph?

In fact, one of the weaknesses of the Configuration model is the absence of cliques or complete graphs. Can we overlook this weakness?

1.3 Random model graph Expected Degree

Similar to the previous model, let's assume $\mathbf{d} = (d_1, d_2, ..., d_n)$ is a sequence of degrees of a graph.

If we assume $(\max_{1 \leq i \leq n} d_i)^2 < \sum_{k=1}^{n} d_k$, then we define:

$$p_{i,j} = \frac{d_i d_j}{\sum_{k=1}^{n} d_k}$$

In this model, there is an edge between vertices v_i, v_j with probability $p_{i,j}$, and there is no edge with probability $1 - p_{i,j}$. Furthermore, the existence or non-existence of different edges is independent of each other. Note that in this model, we can also have loops. The graph obtained from this method is called $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$.

theory question 7. Notice that $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ is a random graph, and therefore its properties, such as the degrees of the random variables, vary. If we denote the degree of vertex v_i as D_i, prove that $\mathbb{E}[D_i] = d_i$.

computer question 5. Assuming the comedy genre has n users, consider the sequence of degrees as $\mathbf{d} = (d_1, d_2, ..., d_n)$. Propose two "suitable" values for n and $\mathbf{d}$, and using your proposed values, construct the graph $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ 100 times. Then, check if the average degree sequences of these 100 graphs are equal to $\mathbf{d}$ or not.

theory question 8. Assume Netflix/Disney+/Vimeo has n users who are interested in the comedy genre. Also, assume that each user has the same taste as k other users. In this case, it is obvious that $\mathbf{d} = \underbrace{(k, k, \ldots, k)}_{n \text{ entries}}$ will be the case. Now, show that the probability that the number of similarity relationships of each user in the random graph $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ is equal to k is equal to $\frac{e^{-k} k^k}{k!}$. Then prove that this value is less than $\frac{1}{2}$, indicating a weakness in this model. Note that in the previous model, the degree of vertex v_i was always d_i.

theory question 9. Show that the probability that the degree of vertex v_i becomes zero in the random graph $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ has the following bound: $\mathbb{P}[D_i = 0] \leq e^{-d_i}$.

theory question 10. Prove that if $\sum_{i=1}^{n} e^{-d_i} \leq \epsilon$, then with probability greater than $1 - \epsilon$, the degree of all vertices is greater than or equal to 1.

2 B

Assuming users 1 to 7 are interested in the comedy genre and users 8 to 14 are interested in the drama genre. Within each of these clusters, there are also different subclusters that have more precise and subtle similarities in the interests of their members. For example, among users 1 to 7 who are interested in the comedy genre, users 1 to 3 are interested in comedy TV series, while users 4 to 6 are only interested in comedy movies. On the other hand, as in figure

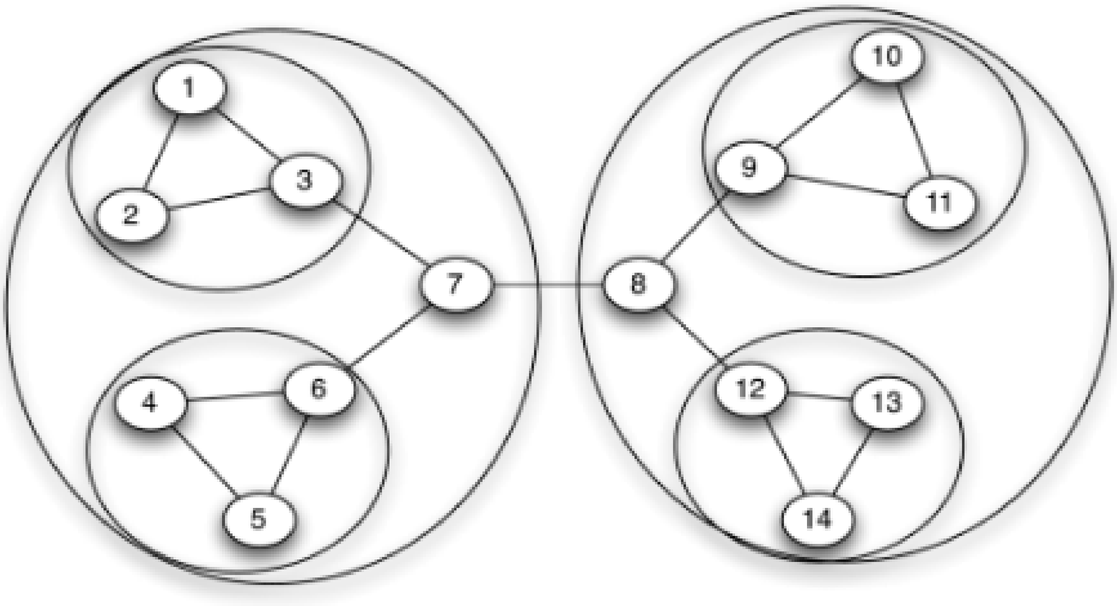

figure 5: An example of clusters and subclusters in a society

5

On the other hand, as seen in Figure 5, user 7 has similar taste to user 8 in the comedy genre and user 7 is likely closer in taste to users 3 and 6, who are in the same comedy enthusiasts cluster, than to user 8, who is in a different cluster. To illustrate this point, we can label the edge representing the taste similarity relationship between user 7 and 8 as ”weak”. Note that the label of the taste similarity edge between user 3 and 7 and between user 6 and 7 can be weak or strong, but we consider the edge between user 7 and 8 (i.e., between two clusters) as weak.

This specific relationship between users 7 and 8 is important for a VOD company because it can recommend movies to user 8 and also suggest those movies to user 7, who has a similar taste but belongs to a different cluster. If user 7 likes those movies, they can also introduce them to other members of the comedy genre cluster who have similar taste to user 7. This approach may enable us to provide new recommendations to users.

theory question 11. According to the given explanations, in real social relationships, do you think it is more likely to obtain a job opportunity or a piece of news from close friends or acquaintances? Why?

Assuming user A is compatible with users B and C. However, since users B and C have not been suggested many common movies, we currently do not know whether B and C are compatible or not, and we need to suggest them several movies to investigate this matter.

theory question 12. Explain why the probability of compatibility between B and C is higher when both are compatible with A compared to the case when they have no common compatibility. This property is called "Triadic Closure." Then conclude that the number of edges in the graph related to the cluster of interest in a genre increases over time.

Now suppose that for any three users A, B, and C, if we know that users B and C are compatible with user A, the probability of creating Triadic Closure and the compatibility between B and C is equal to p.

theory question 13. If we know that users B and C are compatible with users A1, A2, ..., Ak, then find the probability of compatibility between B and C. Justify the probability behavior with increasing k.

3 C

In this section, we want to examine the validity or invalidity of the following hypothesis while preserving the LaTeX syntax.

Hypothesis: Generally, a woman's taste is more similar to another woman's taste than to a man's taste.

We consider the graph of similarity relationships between individuals as $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$, where $\mathcal{G}_1 = (V_1, E_1)$ consists only of female users, and $\mathcal{G}_2 = (V_2, E_2)$ consists only of male users. Therefore, $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$. We also assume that there are n_1 female users, n_2 male users, and a total of $n = n_1 + n_2$ users. We take $\mathcal{G}_1$ as a random graph in which there is an edge between any two vertices with probability p_1, independently of other vertices. We take $\mathcal{G}_2$ as a random graph in which there is an edge between any two vertices with probability p_2, independently of other vertices. Additionally, between any two vertices, such as u, v where $u \in V_1$ and $v \in V_2$, there is an edge with probability p_{12}, independently of other edges.

theory question 14. Write the null hypothesis and the alternative hypothesis based on the above parameters.

Let's assume that the number of same-sex relationships among women is m_1, the number of same-sex relationships among men is m_2, and the number of relationships between a woman and a man is m_{12}.

theory question 15. Design a hypothesis test to test the proposed hypothesis.

computer question 6. Assume that the relationships between individuals are as shown in Figure 6 (pink nodes represent women, and white nodes represent men). Write a program that takes the value of the Type I error (α) as input and prints in the output whether the null hypothesis is rejected or not for this value of α. In what values of α is the null hypothesis rejected?

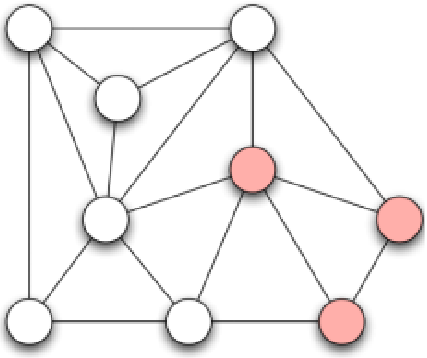

figure 6: example of a relationship between men and women

4 D

Assume Netflix/Disney+/Vimeo has n users. One of the factors that can help us find clusters is the number of common favorite movies between two users. In other words, if two users have a large number of movies that they both like, they are likely to be members of the same cluster. The number of common favorite movies between two users can be represented by the edge weight between their corresponding vertices in the graph. Therefore, the definition of the adjacency matrix changes:

definition 7 (Weighted Adjacency Matrix for Louvain Graph)**.** The adjacency matrix for the weighted graph $\mathcal{G}$ with n vertices is a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ whose elements are defined as follows:

$$A_{i,j} = \text{the weight of the edge between vertices } i \text{ and } j.$$

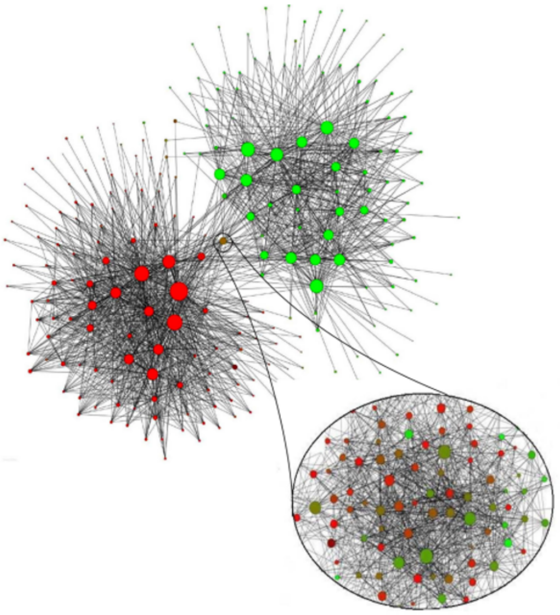

figure 7: example of homogeneous relationships in a graph

In this section, we want to explore another method for identifying clusters. This algorithm is called Louvain [1] . In this method, we first define the modularity for a weighted graph with an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ as follows:

$$Q = \frac{1}{2m}\sum_{i,j=1}^{n}\left[A_{i,j} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{1}$$

In relation to equation (1), $A_{i,j}$ represents the weight of the edge between vertices v_i and v_j, m is the total weight of all edges in the graph, k_i is the sum of weights of edges connected to vertex v_i, and $\delta(c_i, c_j)$ is defined as $\delta(c_i, c_j) = \mathbb{1}\{c_i = c_j\}$. Our goal in this algorithm is to increase the value of Q.

In the algorithm Louvain, we first consider each vertex of the graph as a separate cluster (or equivalently, we consider each person interested in a separate genre, such that no two people

[1] A city in belgium!

have the same interest). Then, in a loop, we try to increase the value of Q in each iteration. We continue this process until we cannot find a significant increase in the value of Q.

The algorithm is as follows:

1. consider each vertex of the graph a separate cluster.
2. for $t = 1, 2, \dots, T$:
   (a) for $i = 1, 2, \dots, |V|$:
      (i) Consider all the neighbors of vertex v_i, which means all the vertices like v_j such that $A_{i,j} > 0$. We put these neighbors in a set called N_i.
      (ii) For each $v_j \in N_i$:
         - Remove the head v_i from its cluster and add it to the cluster where v_j is located. Record the amount of changes Q with this movement. We call this value $\Delta Q_{i,j}$.

      (iii) Define: $\Delta Q_i^{(t)} = \max_{j:v_j \in N_i} \Delta Q_{i,j}$ and $j^* = \arg\max_{j:v_j \in N_i} \Delta Q_{i,j}$.
      (iv) If $\Delta Q_i^{(t)} > 0$, remove vertex v_i from its cluster and add it to the cluster containing v_{j^*}. Otherwise, keep v_i in its own cluster.

   (b) Divide each cluster into a superhead. This superhead has a loop (circle) around it, and the weight of this loop is equal to the sum of the number of edges connecting each vertex to another vertex within the same cluster. Additionally, the weight of the edge between these superheads is equal to the sum of the weights of the edges between the vertices of the two clusters.
   (c) Update the set of graph vertices (V)."

In Figures 8 and 9, we can observe the two main steps of this algorithm.

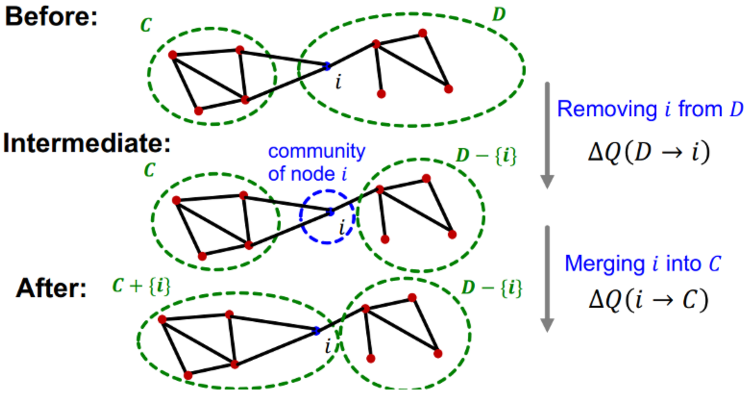

figure 8: example of the displacement of the vertices between clusters

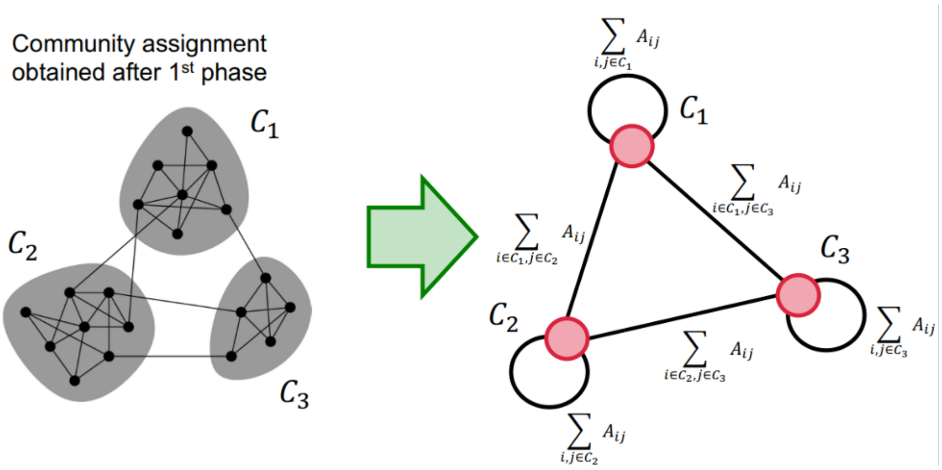
Community assignment
obtained after 1st phase
C_1
C_2
C_3
$\sum_{i,j\in C_1} A_{ij}$
$\sum_{i\in C_1, j\in C_3} A_{ij}$
$\sum_{i\in C_1, j\in C_2} A_{ij}$
$\sum_{i,j\in C_3} A_{ij}$
$\sum_{i\in C_2, j\in C_3} A_{ij}$
$\sum_{i,j\in C_2} A_{ij}$

figure 9: An example of head swapping between clusters

theory question 16. Run the Louvain algorithm manually on the graph shown in Figure 10. Do you obtain the same outputs? (In step 1, examine the nodes in the order of the numbers in the figure.)

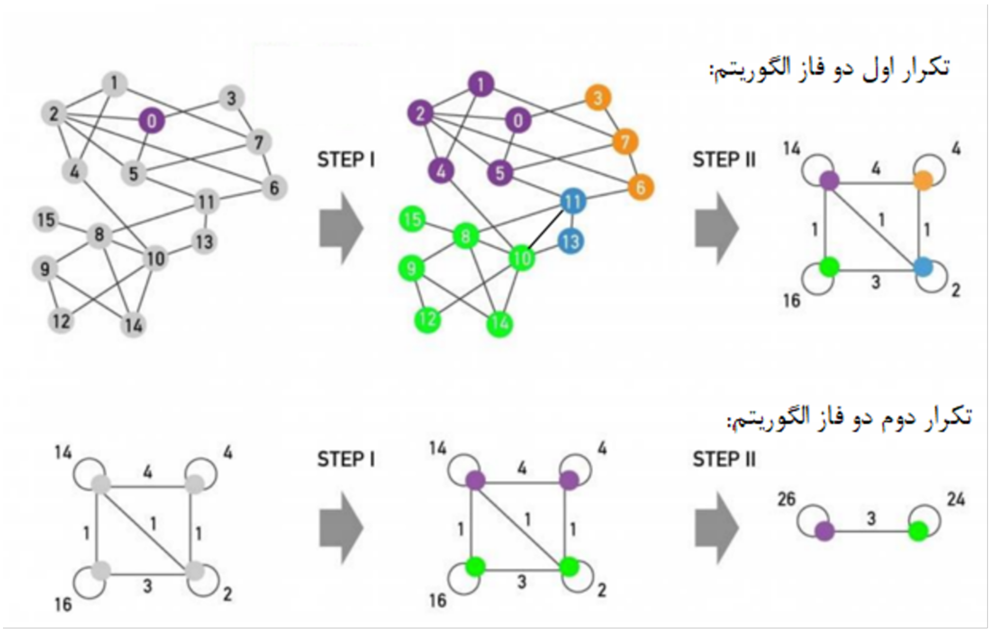

figure 10: example of an algorithm Louvain

theory question 17. By using the equation (1) and the Louvain algorithm, explain why we can better identify clusters by increasing (rather than decreasing) Q.

Assume that vertex v_i wants to be added to the cluster containing vertex v_j, denoted by c. Show that the change in Q due to this move (considering only the addition step) is equal to:

$$\Delta Q_{i,j} = \left[\frac{\Sigma_{\text{in}} + 2k_{i,\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

where Σ_{in} is the sum of the weights of all edges within cluster c, Σ_{tot} is the sum of the weights of all edges connected to the vertices in cluster c, k_i is the sum of the weights of all edges connected to vertex v_i, and $k_{i,\text{in}}$ is the sum of the weights of the edges connecting v_i to any member of cluster c.

computer question 7. In a simple model, if the number of movies that two users have watched in common is above a certain threshold, a weight of 1 is assigned to the edge between them, and otherwise, no edge is drawn between them. Assuming we have examined the similarity relationships among 16 users of a VOD (Video on Demand) service as described, and the resulting graph is shown in Figure 11.

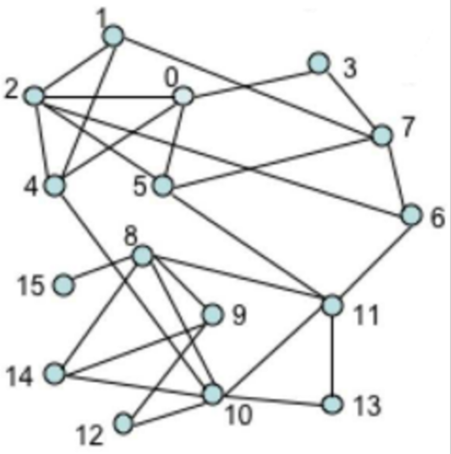

Figure 11: An example of homophily relationships among individuals in a VOD

Write a program that executes the Louvain algorithm on the graph shown in Figure 11 and displays the output for the first to fifth iterations.

Instead of the equation (1), other definitions can be used for Q. For example:

$$Q = \frac{1}{2m} \sum_c (e_c - \gamma \frac{{K_c}^2}{2m}). \qquad (2)$$

In equation (2), e_c represents the number of edges within cluster c, and K_c represents the sum of degrees of nodes within cluster c. Additionally, m represents the sum of weights of all edges in the graph.

computer question 8. Repeat simulation question 7 using the definition of Q given in equation (2). Set $\gamma > 0$ to several different values and report its effect.

computer question 9. Using the Louvain algorithm, obtain the clusters for the dataset ”Zachary’s Karate Club” provided to you among the attachments of the first phase of the project. Calculate the modularity for both the original graph and the clustered graph. You can use existing libraries for this question.

computer question 10. Explain the problems of internally disconnected communities and premature termination in the Louvain algorithm. Why can the modularity measure lead to such problems? Refer to Figure 12 for an example.

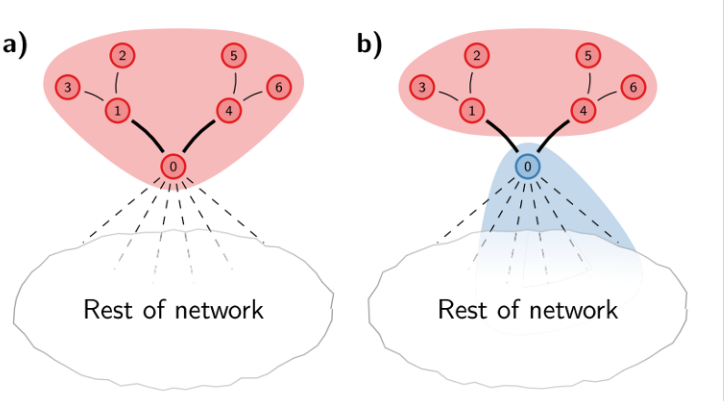

Figure 12: An example of problems with the Louvain algorithm

5 Important Points!

Please pay attention to the following points:

1. This phase of the project is entirely optional.
2. You should complete this phase of the project individually, and you will be graded individually as well.
3. The titles of different sections of the project have been chosen from the works of poets and literary figures of Iran and the world. These poems are not related to the concepts you will encounter in each section.
4. All simulations must 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 only allowed in the cases mentioned. If you click on the title of each library, you will be directed to its documentation.
5. The project should be submitted as a report and the written code. The report should include the answers to the questions, images and graphs, and necessary conclusions. Note that the main part of the simulation burden is on your report and the results you obtain from the code. Also, the cleanliness of the report is very important. Upload the code and report in a compressed file on the course platform.
6. If you have used any sources (books, articles, websites, etc.) to answer the questions, be sure to reference them.
7. Writing the report using LaTeX earns extra points.
8. Simulation questions are marked in green, and theoretical questions are marked in blue.
9. You can write the theoretical sections of the report on paper and include their images in your report, but I kindly advise against doing so!
10. In case of cheating, both individuals will receive a zero grade.

Good luck!