CS550: Massive Data Mining and Learning
Homework 3

Twisha Gaurang Naik (tn268)

Due 11:59pm Saturday, April 18, 2020
Only one late period is allowed for this homework

Submission Instructions

Assignment Submission Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

Late Day Policy Each student will have a total of *two* free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

Honor Code Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Honor Code.

(Signed) Twisha Gaurang Naik (tn268)

If you are not printing this document out, please type your initials above.

Answer to Question 1(a)

Modularity of the original graph G:

- Adjacency Matrix of Graph G:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Degree distribution:
  k = [4, 3, 3, 3, 2, 2, 4, 1]
- Number of nodes (m) = 11
- S (community label vector) = [1, 1, 1, 1, -1, -1, -1, -1]

Modularity can be computed using the following formula:

$$\boxed{Q = \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})s_i s_j} \tag{1}$$

Putting the values of A, k, m and s in equation (1),

$$\boxed{\text{Modularity of the network Q} = \mathbf{0.39256}}$$

Modularity of the modified graph:
Removing edge (A-G) and partition the graph into two parts we calculate the modularity Q as follows.

- Adjacency Matrix:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & \mathbf{0} & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ \mathbf{0} & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Degree distribution:
  k = [3, 3, 3, 3, 2, 2, 3, 1]

- Number of nodes (m) = 10
- S (community label vector) = [1, 1, 1, 1, -1, -1, -1, -1]

Putting the values of A, k, m and s in equation (1),

$$\boxed{\text{Modularity of the network Q} = \mathbf{0.48}}$$

Answer to Question 1(b)

Keep the original graph and retaining communities as per Q1-(a).
Now, adding edge (E-H), we calculate the modularity (Q) as follows.

- Adjacency Matrix:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 1 & 0 \end{bmatrix}$$

- Degree distribution:
  k = [4, 3, 3, 3, 3, 2, 4, 2]
- Number of nodes (m) = 12
- S (community label vector) = [1, 1, 1, 1, -1, -1, -1, -1]

Putting the values of A, k, m and s in equation (1),

$$\boxed{\text{Modularity of the network Q} = \mathbf{0.413194}}$$

The modularity (Q) of the original graph is **0.393**
The modularity **went up** as compared to Q1-(a).
Explanation: The nodes E and H are within the same community. Thus, adding the edge (E-H) to the original graph increases the intra-community connectivity and results into better overall community structure.
As E and H are in same community, the product $s_i s_j$ will be 1. Hence, it results to one extra positive term in addition for calculation of Q and modularity increases.

Answer to Question 1(c)

Keep the original graph and retaining communities as per Q1-(a).
Now, adding edge (A-F), we calculate the modularity (Q) as follows.

- Adjacency Matrix:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & \mathbf{1} & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ \mathbf{1} & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Degree distribution:
  k = [5, 3, 3, 3, 2, 3, 4, 1]
- Number of nodes (m) = 12
- S (community label vector) = [1, 1, 1, 1, -1, -1, -1, -1]

Putting the values of A, k, m and s in equation (1),

$$\boxed{\text{Modularity of the network Q} = \mathbf{0.31944}}$$

The modularity (Q) of the original graph is **0.393**
The modularity **went down** as compared to Q1-(a).
Explanation: The nodes A and F are in different communities. Thus, adding the edge (A-F) increases the inter-community connectivity. While partitioning the graph into communities, we need to minimize inter-cluster edges. Thus, increasing inter-community connectivity leads to decrease in the modularity of the network.
As the nodes A and F belong to the different communities, the product $s_i s_j$ will be -1. Hence, it results in decrease of the value of Q.

Answer to Question 2(a)

- Adjacency Matrix: (8x8 symmetric matrix)

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Degree Matrix: (8x8 diagonal matrix)

$$D = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- Laplacian matrix: (8x8 symmetric diagonal matrix)

$$L = D - A$$

$$L = \begin{bmatrix} 4 & -1 & -1 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & 0 & 0 & -1 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Answer to Question 2(b)

- 2.1403818880962127e-16, $\begin{bmatrix} -0.35355339 \\ -0.35355339 \\ -0.35355339 \\ -0.35355339 \\ -0.35355339 \\ -0.35355339 \\ -0.35355339 \\ -0.35355339 \end{bmatrix}$

- 0.3542486889354087, $\begin{bmatrix} -0.24701774 \\ -0.38252766 \\ -0.38252766 \\ -0.38252766 \\ 0.38252766 \\ 0.38252766 \\ 0.24701774 \\ 0.38252766 \end{bmatrix}$

- 1.0000000000000049, $\begin{bmatrix} 0.00000000e+00 \\ -3.18493382e-17 \\ 8.59836280e-17 \\ -5.79022479e-17 \\ -4.08248290e-01 \\ -4.08248290e-01 \\ 3.76795815e-18 \\ 8.16496581e-01 \end{bmatrix}$

- 3.0000000000000036, $\begin{bmatrix} 0.00000000e+00 \\ -2.70599246e-17 \\ -1.12776339e-16 \\ 1.50488323e-16 \\ 7.07106781e-01 \\ -7.07106781e-01 \\ -1.06520593e-17 \\ 1.12242936e-16 \end{bmatrix}$

- 3.999999999999996, $\begin{bmatrix} 0.60717154 \\ -0.27939608 \\ -0.1005666 \\ -0.22720886 \\ -0.20239051 \\ -0.20239051 \\ 0.60717154 \\ -0.20239051 \end{bmatrix}$

- 4.000000000000001, $\begin{bmatrix} 0.00000000e+00 \\ 5.62206567e-01 \\ 2.31676233e-01 \\ -7.93882800e-01 \\ 2.16840434e-16 \\ -2.82759927e-16 \\ -1.11022302e-16 \\ -1.71737624e-16 \end{bmatrix}$

- 4.000000000000002, $\begin{bmatrix} -0.07964119 \\ -0.56053094 \\ 0.80283611 \\ -0.16266398 \\ 0.02654706 \\ 0.02654706 \\ -0.07964119 \\ 0.02654706 \end{bmatrix}$

- 5.645751311064582, $\begin{bmatrix} 0.66255735 \\ -0.14261576 \\ -0.14261576 \\ -0.14261576 \\ 0.14261576 \\ 0.14261576 \\ -0.66255735 \\ 0.14261576 \end{bmatrix}$

Answer to Question 2(c)

1. Second smallest eigenvalue $\lambda_2 = 0.3542486889354087 \approx 0.3542$
2. Eigen vector corresponding to the second smallest eigen value = $\begin{bmatrix} -0.24701774 \\ -0.38252766 \\ -0.38252766 \\ -0.38252766 \\ 0.38252766 \\ 0.38252766 \\ 0.24701774 \\ 0.38252766 \end{bmatrix}$
3. Partioning the graph with 0 as the boundary:

   Community 1: (Negative values)

Node	**Eigen vector value**
A	-0.24701774
B	-0.38252766
C	-0.38252766
D	-0.38252766

   Community 2: (Positive values)

Node	**Eigen vector value**
E	0.38252766
F	0.38252766
G	0.24701774
H	0.38252766

Answer to Question 3(a)

Prove: If i is any integer greater than 1, then the set C_i of nodes of G that are divisible by i is a clique.
Proof: We know, all the nodes in C_i are divisible by i. Hence, they have **i** as the common factor apart from 1. Thus, none of the nodes in C_i are relatively prime and each pair have an edge between them. Hence, it forms a clique.

Answer to Question 3(b)

Condition for C_i to be a maximal clique is: **i should be a prime number**.
Proof: A maximal clique is defined as: "A clique for which it is impossible to add a node and still retain the property of being a clique. In other words, a clique C is maximal if every node not in C is missing an edge to at least one member of C."

- i is a non-prime (composite) number
  There will exist a number j such that $1 < j < i$ which is a factor of i. All the nodes in C_i will also have an edge with the node j. As all the nodes in C_i have an edge with a node other than i, it is not a maximal clique.

- i is a prime number
  As i is prime, there is no number smaller than i which divides i. Hence, there cannot exist any number other than i that has an edge with all the nodes in C_i. This makes C_i a maximal clique.

Thus, C_i is a maximal clique for every prime integer $i < 1000000$.

Answer to Question 3(c)

Prove: C_2 is the largest unique maximal clique.
Proof: 2 is an even number and the smallest prime number.

- As proved in the last question, a number has to be prime for the clique C_i to be maximal. 2 being a prime number, the clique C_2 is maximal.

- Now, 2 is a factor of all the even numbers which is half of the total nodes. Hence, it will have the largest number of nodes in its clique. This makes C_2 the largest maximal clique.