CS550: Massive Data Mining and Learning
Homework 3

4ex Due 11:59pm Friday, Apr 9, 2021

Only one late period is allowed for this homework (11:59pm Saturday Apr 10)

4

Submission Instructions

4ex

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 Canvas. 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):
Rohan Shah (ras513)

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

I acknowledge and accept the Honor Code.

(Signed) RG

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

Answer to Question 1(a)

The modularity of G when removing the edge (A, G) is 0.48.

Answer to Question 1(b)

The modularity of G when adding the edge (E, H) is 0.41. This means that G is partitioned less well than it was in **1.a**. This measure decreases in value because the edge (A, G) is now present in G, which blurs the division between the two partitions defined in **1.a**. However, this effect is offset by the addition of the edge (E, H), which increases the number of edges within the second group of the partition and contributes to establishing community structure.

Answer to Question 1(c)

The modularity of G when adding the edge (A, F) is 0.32. This means that G is partitioned less well than it was in **1.a**. Again, as in **1.b**, the presence of the edge (A, G) obscures the two partitions. In contrast to **1.b**, however, this effect is augmented by the addition of edge (A, F), which is an additional edge between the two partitions that results in decreased modularity.

Answer to Question 2(a)

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

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

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

Answer to Question 2(b)

The eigen values and corresponding eigen vectors are:

$$0.00 \rightarrow \begin{bmatrix} -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \end{bmatrix} \quad 0.35 \rightarrow \begin{bmatrix} -0.25 \\ -0.38 \\ -0.38 \\ -0.38 \\ 0.38 \\ 0.38 \\ 0.25 \\ 0.38 \end{bmatrix} \quad 1.00 \rightarrow \begin{bmatrix} 0.00 \\ -0.00 \\ -0.00 \\ 0.00 \\ 0.41 \\ 0.41 \\ 0.00 \\ -0.82 \end{bmatrix} \quad 3.00 \rightarrow \begin{bmatrix} 0.00 \\ -0.00 \\ -0.00 \\ 0.00 \\ 0.71 \\ -0.71 \\ 0.00 \\ 0.00 \end{bmatrix}$$

$$4.00 \rightarrow \begin{bmatrix} 0.57 \\ -0.05 \\ -0.48 \\ -0.04 \\ -0.19 \\ -0.19 \\ 0.57 \\ -0.19 \end{bmatrix} \quad 4.00 \rightarrow \begin{bmatrix} 0.22 \\ -0.45 \\ 0.69 \\ -0.46 \\ -0.07 \\ -0.07 \\ 0.22 \\ -0.07 \end{bmatrix} \quad 4.00 \rightarrow \begin{bmatrix} 0.00 \\ -0.71 \\ 0.01 \\ 0.70 \\ -0.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix} \quad 5.65 \rightarrow \begin{bmatrix} 0.66 \\ -0.14 \\ -0.14 \\ -0.14 \\ 0.14 \\ 0.14 \\ -0.66 \\ 0.14 \end{bmatrix}$$

Answer to Question 2(c)

The second smallest eigen value and its corresponding eigen vector is:

$$0.35 \rightarrow \begin{bmatrix} -0.25 \\ -0.38 \\ -0.38 \\ -0.38 \\ 0.38 \\ 0.38 \\ 0.25 \\ 0.38 \end{bmatrix}$$

Using a boundary of 0, the graph would be split into the following two communities: $ABCD$ and $EFGH$.

Answer to Question 3(a)

The nodes of G are connected to one another if they share a common factor. In this context, a common factor is simply a positive integer greater than 1 that divides evenly into two or more greater positive integers. Therefore, if i is any integer greater than 1, all nodes that are divisible by i will be connected to one another. These same nodes compose C_i, and since any given pair of nodes in C_i is connected via an edge, C_i is a clique. In the case where there are less than 2 nodes in C_i, C_i is still a clique since there are no pairs of nodes within C_i that are not connected via an edge.

Answer to Question 3(b)

C_i is a maximal clique if and only if i is prime and $i \leq 1000000$. The following proof assumes that $i \geq 2$.

First, we prove that C_i is a maximal clique if i is prime and $i \leq 1000000$.

1. If i is not prime, then its set of factors F includes factors that are not 1 or i. Let $F' = F - \{1, i\}$. Since all nodes of C_i are divisible by i and i is divisible by all $f \in F'$, all nodes of C_i are also divisible by all $f \in F'$. All $f \in F'$ are guaranteed to be less than i since they are factors of i, so they are not divisible by i and therefore not members of C_i. This means that there are at least $|F'|$ nodes that can be added to C_i while still allowing C_i to remain a clique, since all of these extra nodes $f \in F'$ share at least the common factor of f with all nodes currently in C_i. This defies the definition of a maximal clique and can only be avoided when F' is empty; F' will only be empty if the only factors of i are 1 and i, i.e. only if i is prime. Therefore, C_i is not a maximal clique when i is not prime.

2. There are no nodes in G with value greater than 1000000. So, if $i > 1000000$, C_i will be empty. Any singular node in G can then be added to C_i, and C_i would still be a clique. Therefore, C_i is not a maximal clique when $i > 1000000$.

Next, we prove that i is prime and $i \leq 1000000$ if C_i is a maximal clique. If C_i is not a maximal clique, then there exists a node n in G that is not in C_i but has an edge to every current member of C_i. If $C_i + \{n\}$ is a clique, then n is not a multiple of i but n shares a common factor with i greater than 1. This means that i has factors other than 1 and i, meaning i is not prime. Furthermore, since C_i would be empty if $i > 1000000$, this scenario as a whole is impossible as there are no nodes in C_i that any other node can have an edge to. Therefore, if i is not prime or $i > 1000000$, then C_i is not a maximal clique; or, by contrapositive, if C_i is a maximal clique, then i is prime and $i \leq 1000000$.

Answer to Question 3(c)

As proven in **3.b**, C_i is a maximal clique if i is prime and $i \leq 1000000$. $i = 2$ satisfies this condition, so C_2 is a maximal clique. A pattern becomes distinguishable when considering prime numbers between 2 and 1000000 for i:

- $i = 2 \rightarrow 1$ out of every **2** consecutive integers (nodes) will be a member of C_2
- $i = 3 \rightarrow 1$ out of every **3** consecutive integers (nodes) will be a member of C_3
- $i = 5 \rightarrow 1$ out of every **5** consecutive integers (nodes) will be a member of C_5

This pattern indicates that C_2 will have $1000000/2 = 500000$ member nodes, C_3 will have $1000000/3 = 333333$ member nodes, and so on. As 2 is the smallest prime number between 2 and 1000000, there can be no other value for i that results in a larger $|C_i|$ than $|C_2|$.