# Mathematical and Computational Foundations of Data Science

## Make-Up Midterm Exam

May 2023

Sina Hazratpour
Johns Hopkins University

First Name, Last Name: ______________________________

ID Number: ______________________________

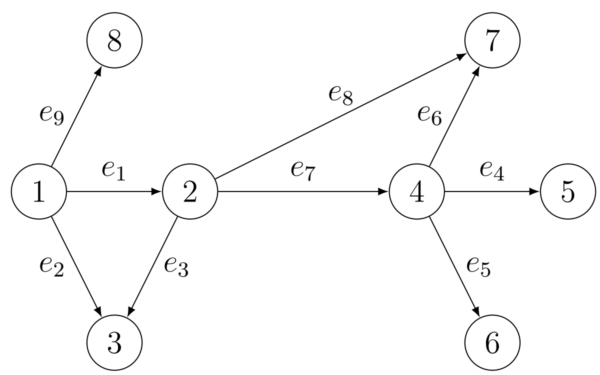
> Answer the questions in the spaces provided. If you run out of room for an answer, continue on the back of the page.

Suppose $G$ is a graph whose vertices are indexed by natural numbers. This gives a natural choice for direction for each edge, indicated by an arrow from one vertex (the source) with lesser index to the other (the target) with higher index. We write $e : E(i, j)$ to say that $e$ is an edge (undirected) between vertex $i$ and $j$. Note that for graphs, given $i$ and $j$, there is at most one such edge $e : E(i, j)$.

The **incidence** matrix of graph $G$, is a rectangular matrix $B_G$ with a row for each vertex in the graph, and a column for each edge, and whose entries are given by

$$b_{ij} = \begin{cases} -1 \text{ if edge } e_j \text{ has source } i, \\ \ \ 1 \text{ if edge } e_j \text{ has target } i, \\ \ \ 0 \text{ if vertex } i \text{ is not on the edge } e_j \end{cases}$$

Here is an example: we have a graph with eight vertices, nine edges, and one connected component (since all vertices are connected to each other):




The incidence matrix of this graph is a $8 \times 9$ matrix $M$ whose rows correspond to the vertices of the graph and whose columns correspond to the edges of the graph. Here is the incidence matrix for the example graph:

$$\begin{pmatrix} -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

1. Show that for any graph $G$ the diagonal entries of $D_G$ are the row sums of $A_G$.

2. Explain why in the incidence matrix each column has exactly one 1 entry and one $-1$ entry and the rest of the entries are zero.

3. In the incidence what does the sum of the entries of each row indicate? What does the sum of the absolute values of the entries of each row indicate? Justify your answers.

4. Consider the matrix $B_G$ as a linear map by letting $\mathbb{R}^V$ and $\mathbb{R}^E$ be the real vector spaces obtained by treating the vertices and edges of $G$ respectively as orthonormal bases. The matrix $B$ then prescribes a linear map $\mathbb{R}^E \to \mathbb{R}^V$ defined by the following action on every basis edge $e$. If $e$ has source vertex $s$ and target vertex $t$, then $B_G(e) = t - s$. Note that $t - s$ is a formal expression and makes sense only in the vector space generated by the vertices of $G$. Show that the number of connected components of $G$ is $|V| - rank(B_G)$. Here $|V|$ is the number of vertices of graph $G$. And, show that the number of undirected cycles in $G$ is $|E| - rank(B)$. Here $|E|$ is the number of edges of graph $G$ (Hint. you can start by observing that $e_1$, $e_2$ and $e_3$ form an undirected cycle. How do you detect this in the incidence matrix?) .

5. Let $A_G$ be the adjacency matrix of a graph $G$ whose vertices are indexed by the natural numbers. The **Laplacian** of $A_G$ is $L_G = D_G - A_G$ where $D_G$ is the degree matrix of $G$, i.e. a diagonal matrix whose diagonal entries are the degrees of the nodes of $G$. Show that $L_G = B_G B_G^T$ where $B_G$ is the incidence matrix of graph $G$.

6. Use the previous part to prove the following explicit formula for the entries of the Laplacian matrix:

$$L_{ij} = \begin{cases} d_i \text{ if } i = j, \\ -1 \text{ if vertices } i \text{ and } j \text{ are adjacent,} \\ \ 0 \text{ otherwise.} \end{cases}$$

   where $d_i$ is the degree of the vertex $i$ (the total number of the incoming and outgoing edges to the vertex $i$).

7. Prove that $L_G$ is positive semi-definite. What can you say about the eigenvalues of $L_G$? Are they all real? Are they all positive?

8. Show that the minimum eigenvalue of the Laplacian $L_G$ is 0. What is the corresponding eigenvector? Justify this using the equation $L_G = B_G B_G^T$.

9. Explain why $L$ has an orthonormal set of eigenvectors.

10. The second eigenvector (corresponding the second smallest eigenvalue) is orthogonal to the first one. What does that tell us about the second eigenvector? In particular, if we interpret the vector $w$ as the vector of weights assigned to the vertices of the graph, what does it mean in terms of weights for $w$ to be the the second eigenvector? What about in the special case where the weights come from the set $\{+1, -1\}$?

11. Show that for any $n \times n$ vector $w$, where $n$ is the number of vertices of the graph, we have

$$w^T L_G w = \sum_{e:E(i,j)} (w_i - w_j)^2$$

12. Suppose $v^{(1)}$ is the first eigenvector of $L_G$ (corresponding to the eigenvalue 0). Consider the space of vectors perpendicular to $v^{(1)}$.

13. We are interested in the *unit* vectors which minimize $w^T L_G w$ and are perpendicular to the first eigenvector $v^{(1)}$ of $L_G$, i.e.

$$\operatorname*{arg\,min}_{x \perp v^{(1)}, \|w\|=1} w^T L_G w = \operatorname*{arg\,min}_{w \perp v^{(1)}, \|w\|=1} \sum_{e:E(i,j)} (w_i - w_j)^2$$

Show that this gives the second eigenvector of $L_G$.

14. Using the second eigenvector $v^{(2)}$ of $L_G$, we find two clusters $P$ and $N$ of the vertices of the graph by defining the following clustering assignment:

$$c \colon V(G) \to \{P, N\}$$

where $c(i) = P$ if $v_i^{(2)} > 0$ and $c(i) = N$ if $v_i^{(2)} \leq 0$.

15. Use the clustering method in part 14 to find two clusters of the example graph from the first page. Is this a good clustering? why/why not?