| Name | Praveenaa Kulandhaivel |
|---|---|
| ASU ID Number | 1225814683 |

# CSE 472: Social Media Mining

## Homework I - Linear Algebra, Graph Essentials, Network Measures

Prof. Huan Liu

Due at 2022, September $8^{th}$, 11:59 PM

This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Gradescope**. This is a fillable PDF and you are able to type into answer boxes provided for each question.

1. **[Linear Algebra]** Consider 2-dimensional data points of [2, 1], [1, 0], [2, 2], [3, 2], [0, 0], [0, 2], [-1, 0].

   (a) All the data points can be gathered together and shown with one matrix. Let's assume $[X]_{2\times 7}$ is that matrix. Fill the following matrix.

$$X = \begin{pmatrix} \boxed{2} & \boxed{1} & \boxed{2} & \boxed{3} & \boxed{0} & \boxed{0} & \boxed{-1} \\ \boxed{1} & \boxed{0} & \boxed{2} & \boxed{2} & \boxed{0} & \boxed{2} & \boxed{0} \end{pmatrix}_{2\times 7}$$

   (b) What is the point showing the center of these points? [*Hint:* Calculate the mean of the values in each dimension].

$$\mu = \begin{pmatrix} \boxed{1} \\ \boxed{1} \end{pmatrix}_{2\times 1}$$

(c) Calculate $Y = (X - \mu)(X - \mu)^T$ in which $X^T$ is the transpose of $X$. To calculate $(X - \mu)$, easily subtract the $\mu$ from all the data points.

$$Y = \begin{pmatrix} \boxed{12} & \boxed{5} \\ \boxed{5} & \boxed{6} \end{pmatrix}_{2\times2}$$

(d) Solve $|Y - \lambda I| = 0$ to extract the values of $\lambda$. $|\cdot|$ is the determinant and $I$ is the identity matrix. $\lambda$ values are called eigenvalues.

$\lambda_1 = \boxed{14.83}$ , $\lambda_2 = \boxed{3.17}$

(e) Calculate the corresponding eigenvector to the **largest** eigenvalue.

$$v = \begin{pmatrix} \boxed{0.87} \\ \boxed{0.492} \end{pmatrix}_{2\times1}$$

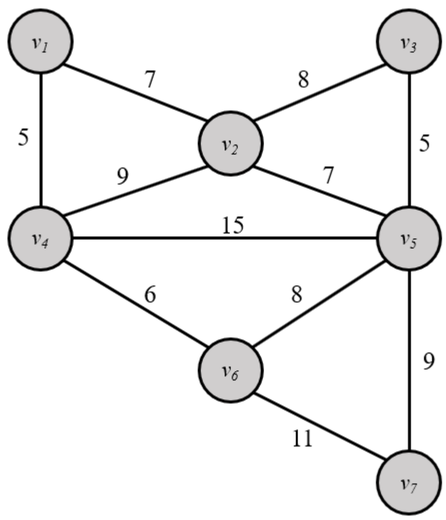
(f) Compute $\hat{X} = v^T X$.

$$\hat{X} = \begin{pmatrix} \boxed{2.232} & \boxed{0.87} & \boxed{2.724} & \boxed{3.594} & \boxed{0} & \boxed{0.984} & \boxed{-0.87} \end{pmatrix}_{1\times7}$$

Congratulations you performed Principle Component Analysis (PCA) procedure, a well known dimensionality reduction method in machine learning. In other words, you projected your 2-dimesional data into 1-dimensional one such that you preserve the variance as much as possible (i.e. the least information has been lost).

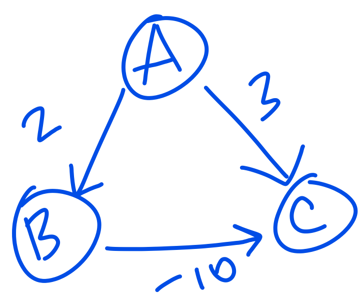
2. **[Graph Algorithms]**

(a) Compute the shortest path between $v_1$ and other nodes using Dijkstra's algorithm for the following graph.




| Node | Distance from $v_1$ |
|---|---|
| $v_2$ | 7 |
| $v_3$ | 15 |
| $v_4$ | 5 |
| $v_5$ | 14 |
| $v_6$ | 11 |
| $v_7$ | 20 |

(b) In the space below, draw a simple example of a directed graph with negative-weight edges for which Dijkstra's algorithm produces incorrect answers.

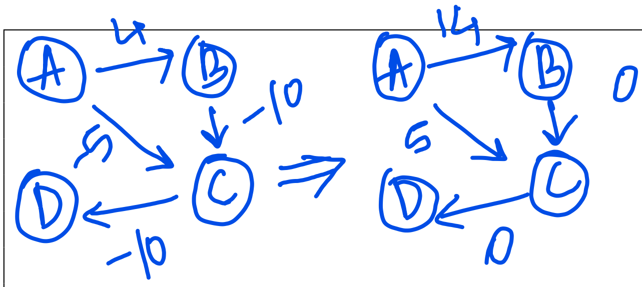



(c) Draw a simple example to argue whether "Algorithm 1" below always produces the shortest paths from one source node to others for graphs that have negative weights but do not have negative cycles.

**Algorithm 1:** Dijkstra Algorithm for graphs with negative weights.

**Input** : Adjacency Matrix $M$, Source node $s$.
**Output:** Shortest Path from $s$ to other nodes.

$C \leftarrow$ Find minimum weight in M
**for** all $i$ and $j$:
    $M[i, j] \leftarrow M[i, j] - C$
**return** Dijkstra(M, s)    *// use the original Dijkstra algorithm to find the shortest paths*



First graph has negative edges and Dijkstra's algorithm will fail.

In second graph, after subtracting the minimum value, all the edges become positive, and Dijkstra's algorithm will work

Therefore, Algorithm 1 will work for graphs with negative weights but no negative cycles

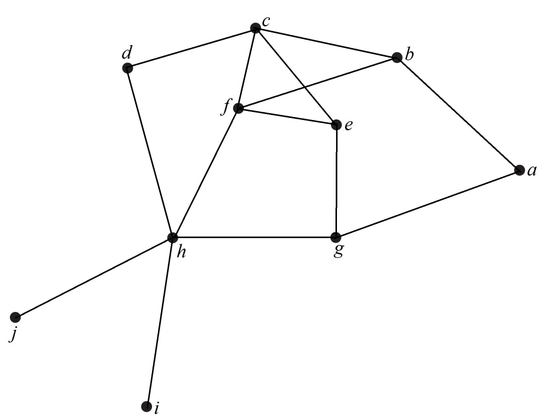
(d) For a real-world social network, is BFS or DFS more desirable? Why? Provide details.

BFS is more desirable in real-world social network, as we interact with our immediate friends first before talking to friends of friends

(e) If we want to calculate Minimum Spanning Trees for graphs, when would we use Kruskal's algorithm over Prim's algorithm?

We prefer Kruskal over Prim's when the graph is sparsely connected.

3. **[Network Measures]** For the given graph compute the following measures.




(a) What is the normalized degree centrality for node $e$?

$$C_d^{norm}(e) = \frac{\boxed{3}}{\boxed{9}} = \boxed{0.333}$$

(b) What is the normalized betweenness centrality for node $h$?

$$C_b^{norm}(h) = \frac{\boxed{35}}{\boxed{72}} = \boxed{0.486}$$

(c) What is the closeness centrality for node $a$ and node $c$?

$$C_c(a) = \frac{1}{\boxed{2.11}} = \boxed{0.474}, \quad C_c(c) = \frac{1}{\boxed{1.77}} = \boxed{0.565}$$

(d) In the case of subgraphs how can you know whether a cycle is socially stable or not?

In social relationships, positive valued edges are considered as friendships and negative value edges are considered to be enmity. A cycle is socially stable in subgraphs when the multiplication of edges gives positive value.

(e) What is the maximum number of edges for a regular graph $G$ with $n$ nodes?

Maximum number of edges of $G$ = (n*(n-1))/

(f) What is the maximum value of local clustering coefficient and when can it be observed for all the nodes in a graph?

Maximum value of local clustering coefficient = 1

When all neighbors are connected in the graph.