CS 343 Graph Data Science

- Create a projection for Movie database for Person and Movie labels and only ACTED_IN relationship without direction
- Calculate the degree for each node
- Calculate PageRank for each node
- Calculate closeness for each node
- Identify the top 5 nodes for each centrality metric
- Show the network of the node with the highest degree i.e. nodes attached with highest degree node
- Show the network of the node with the highest PageRank i.e. nodes attached with highest degree node
- Show the network of the node with the highest closeness i.e. nodes attached with highest degree node

Examples:

Calculating Degree:

```
CALL gds.degree.stream('interactionGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).username AS username, score AS degree
ORDER BY degree DESC;
```

Calculating PageRank:

```
CALL gds.pageRank.stream('webGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).url AS url, score AS pagerank
ORDER BY pagerank DESC
```

Calculating Closeness:

```
CALL gds.alpha.closeness.stream('interactionGraph')
YIELD nodeId, centrality
RETURN gds.util.asNode(nodeId).username AS username, centrality AS closeness
ORDER BY closeness DESC;
```

Calculating Shortest Path:

```
MATCH (a:Actor)
WHERE a.name IN ['Kevin Bacon', 'Denzel Washington']
WITH collect(id(a)) AS nodeIds
CALL gds.shortestPath.dijkstra.stream('proj', {sourceNode:nodeIds[0], TargetNode:nodeIds[1]})
YIELD sourceNode, targetNode, path
RETURN gds.util.asNode(sourceNode).name AS sourceNodeName,
        gds.util.asNode(targetNode).name AS targetNodeName,
        nodes(path) as path;
```

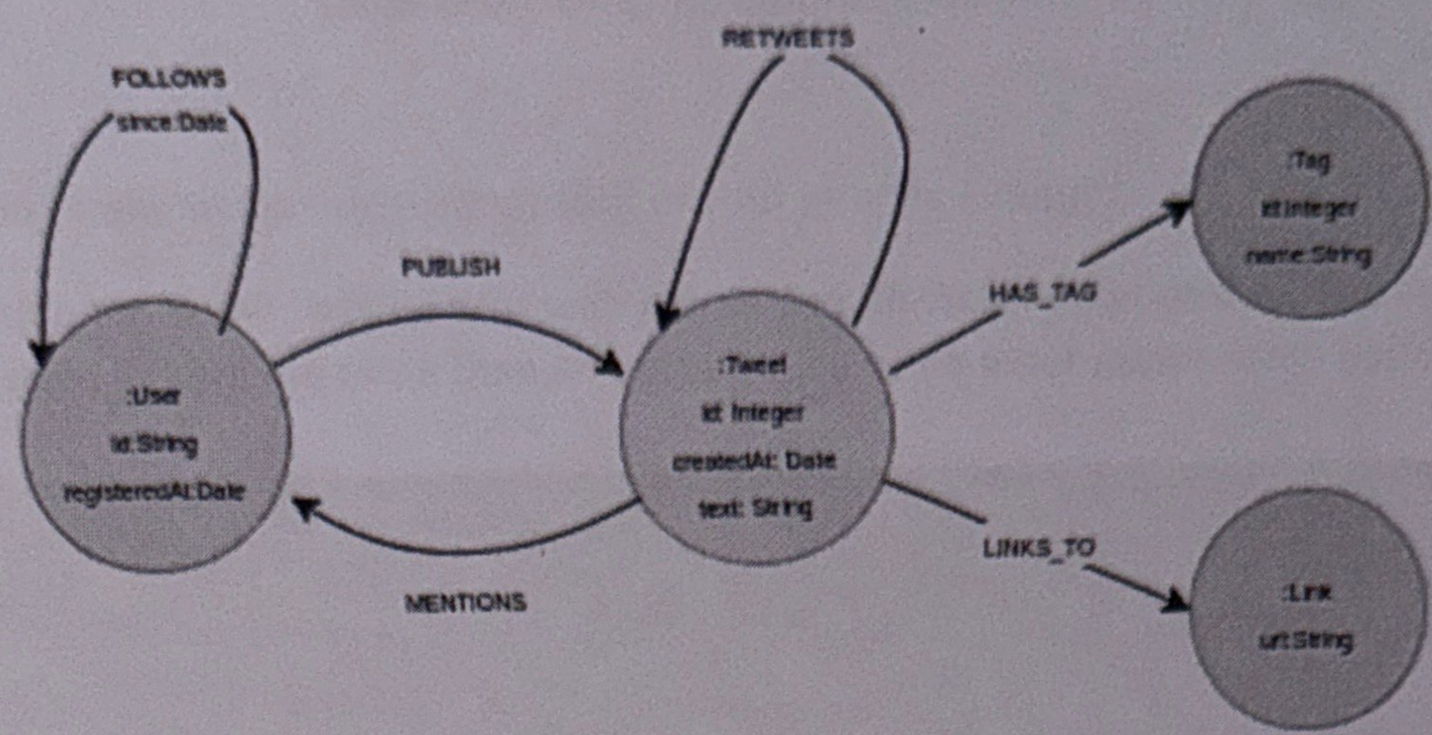
Calculating Breadth-first-search:

```
MATCH (source:Person{name:"Tom Hanks"})
call gds.dfs.stream("test2", {sourceNode: source})
YIELD  sourceNode,  nodeIds,path
RETURN *
```

343 Graph Data Science Quiz 3

Student ID & Name:

1. [50 points] Write the queries for the following considering the following Labelled Graph Property Model:



1.1. Return the number of tweets published by each user.

1.2. Return the number of followers for each user.

1.3. Return the top five users with respect to the number of tweets published by them.

1.4. Return the average number of hashtags used in a tweet.

MATCH (t: Tweet) - [:HAS_TAG]
→ (at : Tags)
WITH t, count(at) as NumOfTags.
RETURN avg(NumOfTags).

1.5. Find the length of the shortest path between User A and User B (where A and B are IDs of the respective users) using the FOLLOWS relationship.

1.6. Find all paths between User A and User B of length up to 3 using the FOLLOWS relationship.

1.7. Generate a projection for analyzing the network of who is following whom.

1.8. Return the top five users followed by other users using gds.degree.stream() – you can refer to the projection generated in the above question.

1.9. Return the top five users following other users using gds.degree.stream().

1.10. Return the top five users having the highest PageRank value using gds.pageRank.stream().

2. Consider a projection where nodes are users and an edge represents a user has mentioned another user in a tweet. Answer the following questions for this network:

2.1. [20] We calculate a weakly connected algorithm which assigns all nodes a component id. On analysis, we found that there are three unique component ids. What have we learned from this information?

2.2. [20] After running a community detection algorithm on this projection, we obtain 10 different communities. The sizes of these communities are 60%, 15%, 5%, 5%, 3%, 2%, 1%, 1%, 1%, and 1%. How would you define a community for this network?

2.3. [10] If we list the top 10 nodes with the highest betweenness values, what characteristics might these nodes possess?

The nodes act like a bridge b/w different communities meaning they
→ have connection spanning multiple clusters
→ play key role in spreading info
→ may include influencers, engaged users, mutual friend.

Breeha Qasim.

Decision Tree Worksheet: Calculating Gini Impurity

Dataset:

ID	Feature A	Feature B	Feature C	Target
1	Yes 1	No	Yes	Yes A
2	No 1	Yes	No	No B
3	Yes	No	No	Yes
4	No	Yes	Yes	No
5	Yes	Yes	No	Yes
6	No	No	Yes	No
7	Yes	No	Yes	Yes
8	No	Yes	No	No
9	Yes	No	No	~~No~~ Yes.
10	No	No	Yes	No

$\left(\frac{2}{2}\right)\left(\frac{2}{2}\right)$

Step 1: Compute Initial Gini Impurity for the Target Variable

- Count how many 'Yes' and 'No' values are in the Target column.
- Compute the initial Gini Impurity using the formula:

$$Gini = 1 - (p_Yes^2 + p_No^2)$$

↳ helps decision tree algo choose best feature to split data.

* to measure how "pure" a node is.
* Low (close to 0): if node is pure
* High (close to 0.5 or 1): if there's a mix of both classes.

Feature A:

Number of rows with Feature A = Yes: 5

- Number of rows Target = Yes: 5
- Number of rows Target = No: 0

] compare 'Yes' of feature A to target.

- Calculate Gini score:

$$Gini_A(Yes) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{5}{5}\right)^2 + \left(\frac{0}{5}\right)^2\right) = 0$$

Number of rows with Feature A =No: 5

- Number of rows Target = Yes: 0
- Number of rows Target = No: 5
- Calculate Gini score:

$$Gini_A(No) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{0}{5}\right)^2 + \left(\frac{5}{5}\right)^2\right) = 0.$$

Weighted Gini Score:

$$Gini\ A = \frac{Number\ of\ Yes}{Total\ Rows} * Gini_A(Yes) + \frac{Number\ of\ No}{Total\ Rows} * Gini_A(No)$$

$$\frac{5}{10} \times 0 + \frac{5}{10} \times 0 = 0$$

$|B| = \sqrt{6^2 + 7^2 + 8^2}$

$\frac{1 \cdot 8 + 2 \times 7 + 3 \times 8 + 4 \cdot 8}{11 \quad 11 \times 11 \quad 11}$

Feature B:

Number of rows with Feature B = Yes: 4

- Number of rows Target = Yes: 1
- Number of rows Target = No: 3
- Calculate Gini score:

$$Gini_B(Yes) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = 3/8.$$

Number of rows with Feature B =No: 6

- Number of rows Target = Yes: 4
- Number of rows Target = No: 2
- Calculate Gini score:

$$Gini_B(No) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right) = 4/9.$$

Weighted Gini Score:

$$Gini\ B = \frac{Number\ of\ Yes}{Total\ Rows} * Gini_B(Yes) + \frac{Number\ of\ No}{Total\ Rows} * Gini_B(No)$$

$$= \frac{4}{10} \times \frac{3}{8} + \frac{6}{10} \times \frac{4}{9} = \frac{5}{12}$$

After Branching at A:

ID	Feature A	Feature B	Feature C	Target
1	Yes	No	Yes	Yes
3	Yes	No	No	Yes
5	Yes	Yes	No	Yes
7	Yes	No	Yes	Yes
9	Yes	No	No	~~No~~ Yes.

Decision Tree

→ a decision tree is a divide and conq. strategy to partition data.
→ ~~form true branch~~ if a decision is pu redo the check.
→ once we've reach a pure node we stop.
→ Now we've to take decision for B & C bec it's impure.

For B: Yes = 1 (1/0), No = 4 (0/0)

For C: Yes = 2 (2/0), No = 3 (3/0)

- Calculate Gini Score:

$$Gini_B(Yes) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 0$$

$$Gini_B(No) = 1 - \sum p^2$$

$$= 1 - \left(\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right) = 0$$

Random Forest

→ Make alot of decision trees
→ computationally expensive.
→ bootstrapping/sampling support.
→ Randomly take data.
→ 2/3 tree used to train
→ the other 1/3 used to evaluate model.

~~Breeha Qasim~~

CS/CE 343/456 Graph Data Science Quiz 4

Student Name & ID: Breeha Qasim (8283). 70

Q. 1. You have built a machine learning model to identify fraudulent transactions (fraud detection) in a financial dataset. The model was trained and then tested on a dataset of 1,000 transactions. Out of these, the model predicted 120 transactions as fraudulent. Among those 120, 70 were actually fraud, and the remaining were legitimate. Out of the 880 transactions predicted as legitimate, 830 were actually legitimate, and the rest were fraud.

1. [15] Construct the Confusion Matrix:

n=1000	Predicted Fraud	Predicted Legitimate
Actual Fraud	70 (TP)	50 (FN)
Actual Legitimate	50 (FP)	830 (TN)
	120	880

15

2. [15] Calculate the following metrics

a. Accuracy: $\frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{70+830}{70+50+50+830} = \frac{9}{10} \times 100 = 90\%$

b. Precision: $\frac{(TP)}{(TP+FP)} = \frac{70}{70+50} = \frac{7}{12}$

c. Recall: $\frac{TP}{(TP+FN)} = \frac{70}{70+50} = \frac{7}{12}$

15

d. F1-Score: $2 * \frac{(Precision * Recall)}{(Precision+Recall)} = 2 \times \frac{(7/12 \times 7/12)}{(7/12 + 7/12)} = \frac{7}{12}$

e. Specificity: $\frac{TN}{(TN+FP)} = \frac{830}{(830+50)} = \frac{83}{88}$

3. [20] Would this model be acceptable in a real-world fraud detection scenario? Why or why not?

Yes, the model will be acceptable in real-world fraud detection scenario because the model is 90% accurate.

10

CS/CE 343/456 Graph Data Science Quiz 4

Q. 2. You are given a dataset of 6 nodes with two numerical attributes (Feature 1, Feature 2) and a class label (either "A" or "B"). The data and pairwise distances between nodes are as follows:

Data

Node	Feature 1	Feature 2	Label
N1	1.2	3.4	A
N2	2.1	2.5	A
N3	3.0	4.0	B
N4	5.1	1.2	B
N5	4.0	3.5	?

Similarity Scores (Euclidean Distance)

N1	N2	N3	N4	N5	N6	
N1	0.00	1.08	1.94	4.29	2.92	1.41
N2	1.08	0.00	1.58	3.26	2.06	0.71
N3	1.94	1.58	0.00	3.13	1.12	2.50
N4	4.29	3.26	3.13	0.00	2.69	3.64
N5	2.92	2.06	1.12	2.69	0.00	2.55
N6	1.41	0.71	2.50	3.64	2.55	0.00

N3
N2

- [30] Perform KNN classification for the test nodes N5 using K=3. Show the working

→ Sort distances

N3, N2, , N4, N1

→ Now we'll pick K=3 nearest neighbours.

$N_3 \rightarrow B$
$N_2 \rightarrow A$
$N_4 \rightarrow B$

→ So we'll classify this, we'll take majority class among K=3 neighbours.

Hence N_5 is classified as 'B'.

30

Q.3. [20] For the given data, calculate Gini Impurity for the target variable:

$$Gini = 1 - (p_{Yes}^2 + p_{No}^2)$$

X1	X2	Class
1	0	A
1	1	B
2	-1	A
3	0	B

$Gini_{x_1}(Yes) =$

$Gini_{x_1}(No) =$

$Gini_{x_2}(Yes) =$

$Gini_{x_2}(No) =$

$Gini_1(X_1) = 1 - ((\frac{2}{2})^2) = 0$

$Gini_2(X_1) = 1 - ((\frac{1}{1})^2) = 0$

$Gini_3(X_1) = 1 - ((\frac{1}{1})^2) = 0$

$Gini_1(X_2) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2)$
$= 1/2$

$Gini_2(X_2) = 1 - ((1/1)^2) = 0$

$Gini_3(X_2)$

2

CS/CE 343/456 Graph Data Science

17

Quiz 3

Student Name & ID: Breeha Qasim. (8283)

Instructions:

- Circle/Mark the correct answer
- Do not use pencil

1. A retail company wants to determine which products significantly influence customer purchases by identifying items frequently bought together. This will help optimize product placement and recommendation systems. Which centrality measure should be used?
   a. Degree Centrality: Measures how many direct connections a product has.
   b. PageRank: Identifies products based on their connections to other frequently purchased items.
   c. Measures how quickly a product can reach all other products in the network.
   d. Focuses on the farthest distance a product has from any other product in the network.

2. A financial institution wants to detect potential fraudsters who act as intermediaries in money laundering schemes. These individuals process transactions between multiple parties without directly appearing as high-volume senders or receivers. Which centrality measure is best suited?
   a. PageRank: Identifies nodes with many important connections.
   b. Betweenness Centrality: Detects nodes that frequently serve as bridges between different parts of the network.
   c. Closeness Centrality: Identifies nodes that can quickly reach all others.
   d. Degree Centrality: Measures how many direct connections a node has.

3. A university wants to streamline its course offerings by identifying key prerequisite courses that serve as essential links between different subjects. These courses help students transition between disciplines effectively. Which centrality measure should be applied?
   a. Betweenness Centrality: Detects courses that frequently connect different subjects.
   b. Eccentricity Centrality: Identifies courses farthest from other courses in the curriculum.
   c. Closeness Centrality: Finds courses that have the shortest average distance to all other courses.
   d. Degree Centrality: Measures the number of direct prerequisite relationships a course has.

4. In a corporate email network, an organization wants to identify employees who act as crucial communicators, bridging multiple departments. These individuals help information flow efficiently across teams. Which centrality measure should be applied?
   a. Betweenness Centrality: Identifies employees who frequently connect different teams.
   b. Eccentricity Centrality: Finds employees who are the farthest from others in communication.
   c. PageRank: Ranks employees based on their importance within the email network.
   d. Degree Centrality: Measures the number of direct email connections an employee has.

5. A city transportation authority wants to determine the most accessible locations in a road network to improve public transportation planning. These locations should have the shortest travel times to other areas in the city. Which centrality measure is appropriate?
   a. Closeness Centrality: Identifies locations that minimize travel distance to all others.
   b. Degree Centrality: Finds locations with the highest number of direct roads.
   c. Betweenness Centrality: Identifies locations that frequently serve as transit hubs.
   d. Eccentricity Centrality: Focuses on locations that are the farthest from any other location.

6. A cybersecurity team is analyzing network logs to detect potential attack points where hackers might control major traffic flows in a system. These points serve as critical gateways through which most data passes. Which centrality measure is most useful?
   a. Betweenness Centrality: Identifies nodes that manage the flow of data.
   b. PageRank: Finds nodes with many important connections.
   c. Degree Centrality: Counts the number of direct connections a node has.
   d. Closeness Centrality: Measures how quickly a node can reach all others.

7. An airline company wants to optimize its network by identifying the most efficient airports for minimizing overall travel time between destinations. These airports should be easily reachable from others to reduce layovers and delays. Which centrality measure should be used?
   a. Closeness Centrality: Identifies airports that minimize travel time to all other airports.
   b. PageRank: Ranks airports based on their importance in the network.
   c. Degree Centrality: Counts the number of direct flight connections an airport has.
   d. Eccentricity Centrality: Measures the farthest distance from an airport to any other airport.

8. On an online job portal, recruiters want to identify well-connected candidates who have a strong professional network, making them valuable hires for referrals and collaborations. Which centrality measure should be considered?
   a. PageRank: Identifies candidates who are linked to influential professionals.
   b. Degree Centrality: Counts the number of direct professional connections a candidate has.
   c. Closeness Centrality: Finds candidates who are centrally positioned in the network.
   d. Betweenness Centrality: Detects candidates who act as bridges between different professional groups.

9. A blockchain-based financial network wants to identify nodes that play a crucial role in transaction verification and securing the system. These nodes facilitate transactions between multiple participants and ensure smooth processing. Which centrality measure is most relevant?
   a. Betweenness Centrality: Identifies nodes that serve as key intermediaries in transaction flows.
   b. PageRank: Ranks nodes based on their importance in the network.
   c. Degree Centrality: Measures how many direct transaction connections a node has.
   d. Eccentricity Centrality: Focuses on nodes that are the farthest from others in the network.

10. A social media platform wants to identify influential users who play a key role in spreading information across the network. These users may not have the most connections, but their endorsement or sharing of content significantly boosts its reach. Which centrality measure should be used?
    a. PageRank: Prioritizes users who are not just well-connected but also followed by other influential users.
    b. Degree Centrality: Counts how many direct friends or followers a user has.
    c. Closeness Centrality: Finds users who are well-connected across the entire network.
    d. Betweenness Centrality: Detects users who act as bridges between different communities.

11. In Neo4j, what is the key difference between the ***write(), mutate(),*** *and* ***stream()*** functions when working with graph projections?
    a. write() modifies the database, mutate() creates in-memory updates, and stream() returns results without modifying data.
    b. write() and mutate() both modify the database, while stream() is only for querying.
    c. mutate() writes data to the database, write() updates in-memory structures, and stream() fetches data in a paginated manner.
    d. All three functions modify the database but differ in how they store results.

CS/CE

CE 343/456 Graph Data Science Quiz 3

12. What is the key difference between a cypher projection and a native projection in Neo4j?
    a. Cypher projections are faster because they work on stored data, while native projections require in-memory transformation
    b. Native projections are optimized for large-scale algorithms, while cypher projections allow flexible graph creation using Cypher queries
    c. Cypher projections permanently modify the database, while native projections do not
    d. Native projections require explicit indexing, whereas cypher projections do not

13. Scenario: A retail store maintains a graph database with the following entities and relationships:
    - Nodes: Orders, Products, Customers, Categories
    - Relationships:
        - Customers ***place*** Orders
        - Orders ***contain*** Products
        - Products ***belong*** to Categories

The retail store wants to analyze which customers frequently purchase the same products. Which Cypher query correctly creates a native projection of customers and products, keeping only the purchased relationship?

A)

```
CALL gds.graph.project(
  'customerProductGraph',
  'Customer',
  'purchased'
)
```

B)

```
CALL gds.graph.project(
  'customerProductGraph',
  ['Customer', 'Product'],
  {
    purchased: {
      type: 'purchased'
    }
  }
)
```

C)

```
CALL gds.graph.project.cypher(
  'customerProductGraph',
  'MATCH (c:Customer)-[:purchased]->(p:Product) RETURN id(c) AS id',
  'MATCH (c:Customer)-[r:purchased]->(p:Product) RETURN id(c) AS source, id(p) AS target'
)
```

D)

```
CALL gds.graph.create(
  'customerProductGraph',
  ['Customer', 'Product'],
  ['purchased']
)
```

14. The retail store wants to analyze product co-purchases by identifying products frequently bought together in the same order. If two products appear in multiple orders together, how should the edges between them be handled?
    a. Do not merge edges; keep a separate relationship for each order.
    b. Merge edges and store the frequency of co-purchases.
    c. Merge edges and sum the total revenue generated from co-purchases.
    d. Merge edges, but store only the most recent order date.

15. A retail store wants to analyze total spending per customer on different products. If a customer purchases the same product multiple times, how should the edges be handled?
    a. Do not merge edges; store each purchase separately.
    b. Merge edges and store the number of times the customer bought the product.
    c. Merge edges and sum the total amount spent on the product.
    d. Merge edges, but store only the last purchase date.

16. Which of the following Cypher queries correctly calculates the in-degree of each node?
    a. MATCH (n) OPTIONAL MATCH (n)-[r]->() RETURN n, COUNT(r)
    b. MATCH (n) OPTIONAL MATCH (n)<-[r]-() RETURN n, COUNT(r)
    c. MATCH (n) OPTIONAL MATCH (n)-[r]-() RETURN n, COUNT(DISTINCT r)
    d. MATCH (n) RETURN n, COUNT(*)

17. Which Cypher query correctly recommends friends of a node "A" based on the Friends-of-Friends method?

    a.
```
MATCH (a {Name:"A"})-[:FOLLOWS]-(x)-[:FOLLOWS]-(b)
WHERE NOT (a)-[:FOLLOWS]-(b)
RETURN b
```

    b.
```
MATCH (a {Name:"A"})-[:FOLLOWS]->(x)-[:FOLLOWS]->(b)
WHERE NOT (a)-[:FOLLOWS]->(b)
RETURN b
```

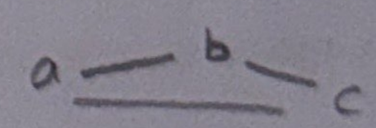
    c.
```
MATCH (a {Name:"A"})-[:FOLLOWS]->(x)<-[:FOLLOWS]-(b)
WHERE NOT (a)-[:FOLLOWS]-(b)
RETURN b
```

    d.
```
MATCH (a {Name:"A"})-[:FOLLOWS]->(x)-[:FOLLOWS]->(b)
RETURN b
```

18. Which Cypher query correctly calculates the number of triangles a node is part of in an undirected graph?

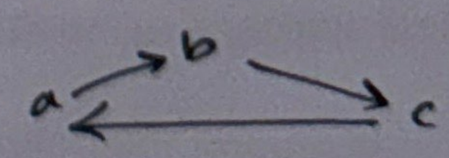
    a.
```
MATCH (a)-[:CONNECTED_TO]-(b)-[:CONNECTED_TO]-(c)-[:CONNECTED_TO]-(a)
RETURN a, COUNT(*) AS triangleCount
```

    b.
```
MATCH (a)-[:CONNECTED_TO]->(b)-[:CONNECTED_TO]->(c)-[:CONNECTED_TO]-
>(a)
RETURN a, COUNT(*) AS triangleCount
```



    c.
```
MATCH (a)-[:CONNECTED_TO]-(b)
MATCH (b)-[:CONNECTED_TO]-(c)
MATCH (c)-[:CONNECTED_TO]-(a)
RETURN a, COUNT(DISTINCT c) AS triangleCount
```

    d.
```
MATCH (a)-[:CONNECTED_TO]->(b)
MATCH (b)-[:CONNECTED_TO]->(c)
MATCH (c)-[:CONNECTED_TO]->(a)
RETURN a, COUNT(DISTINCT c) AS triangleCount
```

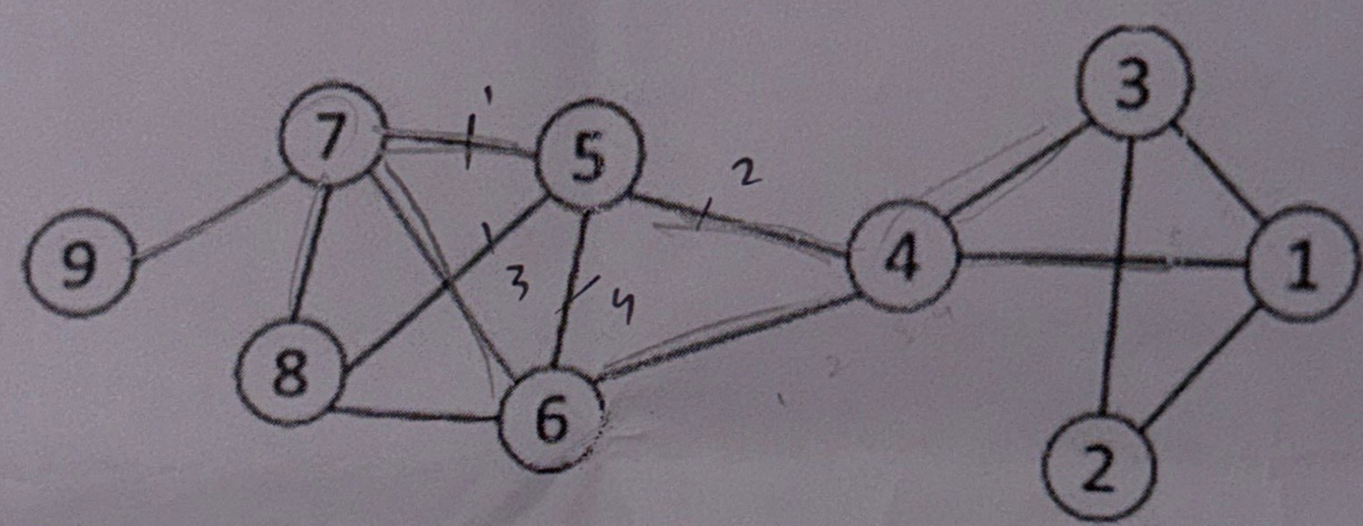


1. [10] What is '
   A. Using
   B. Bui'
   C.
   D
2.

/CE 343/456 Graph Data Science Quiz 3

19. The clustering coefficient of Node '5' in the below given graph is:
    a. 0.33
    b. 0.50
    c. 0.66
    d. 1

20. Which node has expected to have highest betweenness value in the below given graph:
    a. 4
    b. 5
    c. 6
    d. 9



Clustering Coefficient.

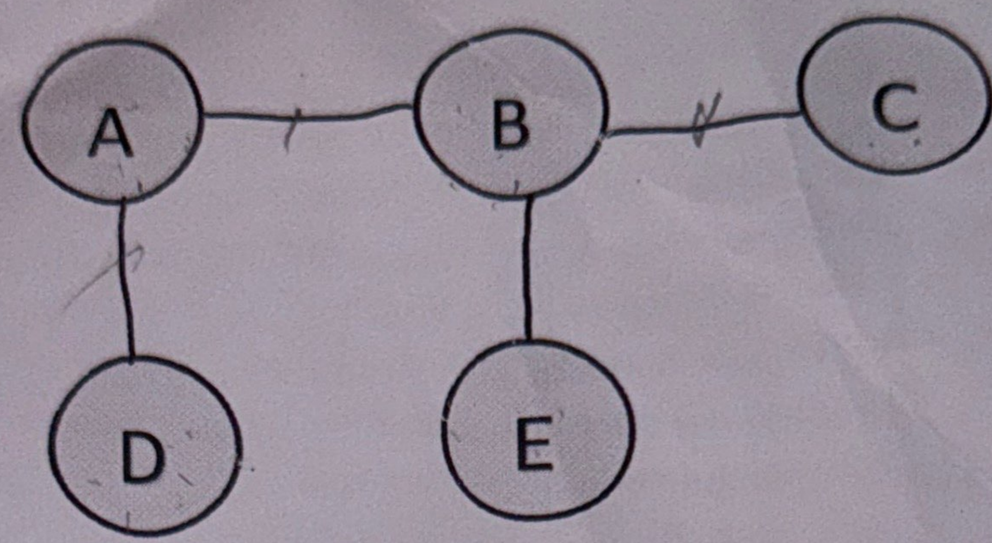
$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

where,
$k_i = 4$
$L_i = 4$

$$= \frac{2 \times 4}{4(4-1)} = 0.66.$$

Graph Data Science

Link Prediction Activity



A D → A

A → A D

A E C

B A

{B, D} ∩ {B}

	Common Neighbors	Adamic Adar	Preferential Attachment	Resource Allocation	Total Neighbors
A,B	0	0	6	0	5
A,C	{B}	1/log(3)	2	1/3	2
A,D	0	0	2	0	3
A,E	{B}	1/log(3)	2	1/3	2
B,C	0	0	3	0	4
B,D	A	1/log(2)	3	1/2	3
B,E	0	0	3	0	4
C,D	0	0	1	0	2
C,E	{B}	1/log(3)	1	1/3	1

Common Neighbors: $CN(x,y) = |N(u) \cap N(y)|$

Adamic Adar: $A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log|N(u)|}$ → degree (log inverse) → to find rare common neighbours connection.

Preferential Attachment: $PA(x,y) = |d(x)| * |d(y)|$

Resource Allocation: $RA(x,y) = \sum_{x \in N(x) \cap N(y)} \frac{1}{N(u)}$ → degree ↑ penalty↓

Total Neighbors: $TN(x,y) = |N(x) \cup N(y)|$

CS/CE 343/465 Graph Data Science (Quiz 1)

Student ID & Name: Breeha Qasim 8283.

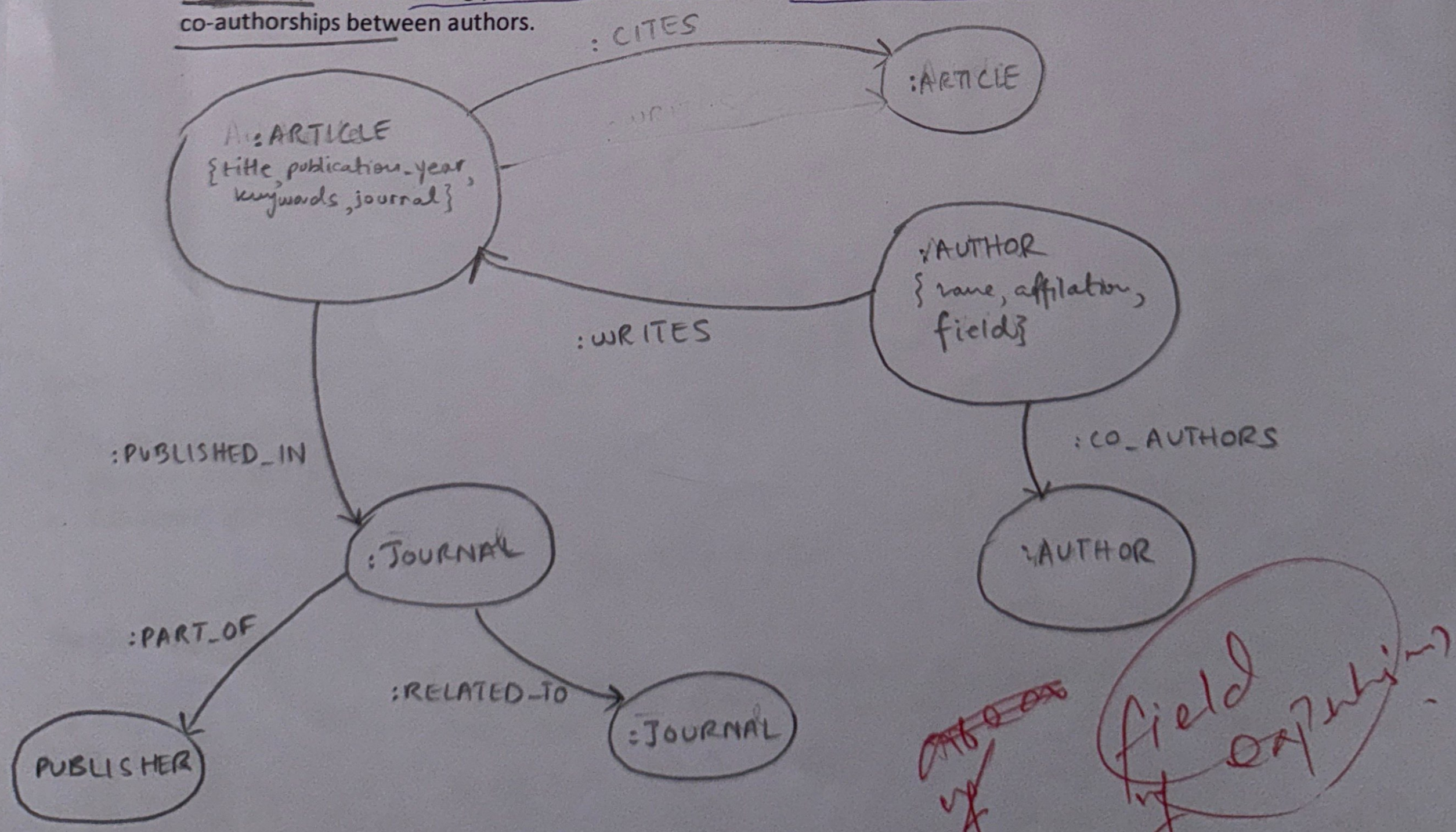
90

1. [10] What is the primary focus of the "Data Representation and Transformation" division in GDS?
   A. Using mathematical and database structures to adapt and transform data into usable formats.
   B. Building predictive models for analysis.
   C. Evaluating workflows for efficiency and effectiveness.
   D. Designing visual dashboards for presenting data.
2. [10] Which of the following best describes the role of the "Computing with Data" division in GDS?
   A. Developing stochastic models for generative analysis.
   B. Designing interactive dashboards for presenting data.
   C. Creating scripts, workflows, and packages to handle data analysis efficiently.
   D. Conducting meta-analyses of data workflows.
3. [10] The division "Science about Data Science" in GDS involves which of the following activities?
   A. Investigating and documenting common workflows in data analysis.
   B. Predicting outcomes using machine learning algorithms for science
   C. Visualizing trends using graphical tools for data science.
   D. Cleaning and preparing raw data for analysis.

4. [30] Create a graph-labeled property model for the following scenario:

node properties -

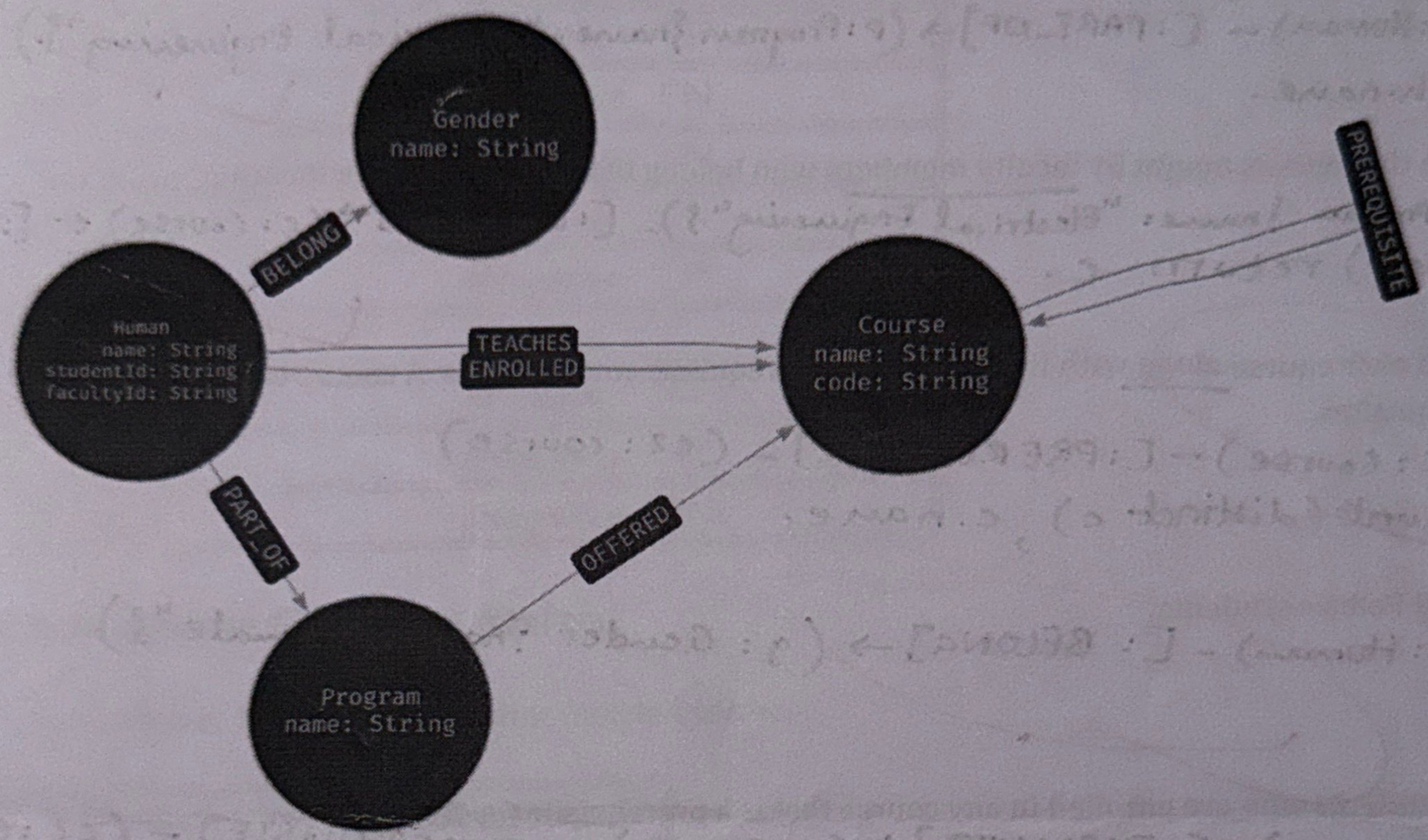
A publishing house wants to model its journal articles, authors, and citations. Each article has attributes like title, publication year, keywords, and the journal in which it was published. Authors are associated with articles, and they have attributes such as name, affiliation, and field of expertise. Articles cite other articles, forming a citation network, and citations should capture the context or sections referenced. Journals themselves may have relationships, such as being part of the same publisher or sharing similar themes. The system should also track co-authorships between authors.



CS/CE 343/456 Graph Data Science **Quiz 2**

Student Name & ID: Breelu Dasin. 75

This graph represents a university academic system, capturing relationships between students, faculty, programs, and courses.



Entities (Nodes)

- Human: Represents students and faculty members.
  - Students enroll in courses.
  - Faculty members teach courses.
- Gender: Represents the gender of each human.
- Program: Represents an academic program (e.g., Computer Science, Electrical Engineering).
  - Students belong to a program.
  - Programs offer courses.
- Course: Represents an academic course (e.g., CS101, CS102).
  - Courses may have prerequisites (i.e., one course must be completed before another).

Relationships (Edges)

- (:Human)-[:BELONGS_TO]->(:Program) → A student belongs to a program.
- (:Program)-[:OFFERS]->(:Course) → A program offers courses.
- (:Human)-[:ENROLLED_IN]->(:Course) → A student is enrolled in a course.
- (:Human)-[:TEACHES]->(:Course) → A faculty member teaches a course.
- (:Course)-[:PREREQUISITE_FOR]->(:Course) → Defines prerequisite relationships between courses.
- (:Human)-[:HAS_GENDER]->(:Gender) → Associates a human with their gender.

CS/CE 343/456 Graph Data Science

Quiz 2

Write Cypher queries for the following:

1. Return the courses offered by the "Computer Science" program.

~~MATCH (c:Courses) - [:OFFERED~~
MATCH (p:Program {name:"Computer Science"}) - [:OFFERED] → (c:Cource)
RETURN c

2. Return the names of students enrolled in the "Electrical Engineering" program.

MATCH (h:Human) - [:PART_OF] → (p:Program {name: "Electrical Engineering"})
RETURN h.name.

3. Return the courses taught by faculty members who belong to the "Electrical Engineering" program.

MATCH (p:Program {name: "Electrical Engineering"}) - [:OFFERING] → (c:Course) ← [:TEACHES]
- (h:Human) return c.

4. Return each course along with the number of prerequisite courses it has. A course can have zero or more prerequisites.

MATCH (c:Course) - [:PREREQUISTE] - (c2: course)
RETURN count(distinct c), c.name.

5. List all Female students.

MATCH (h:Human) - [: BELONG] → (g: Gender {name: "female"})
RETURN h.

6. List students who are enrolled in any course that is a prerequisite for "CS201".

MATCH (h:Human) - [:ENROLLED] → (c:Course) ← [:PREREQUISTE] - (c2:Course
{name: "CS201"}) return h.

7. Add the label **:Teacher** to all nodes representing faculty members (i.e., those teaching at least one course).

MATCH (h:Human) - [:TEACHES] → (c: Couse)
SET h:Teacher.

8. List all courses along with the students enrolled in each course.

MATCH (h:Human) - [:ENROLLED] → (c: Course)
RETURN h.name, c.name.

9. Remove `Computer Science` node from the database.

MATCH (p: Program {name: "Computer Science"}) remove p.name

10. Return the number of courses that must be taken before enrolling in "CS401" (i.e., which courses a student must be taken in order to enroll in CS401).

MATCH (c: Course {name: "CS401"}) - [:PREQUISTE] → (c2: Course)
RETURN count (c2).

Breeha Qasim.

Dataset

A machine learning model was used to classify whether patients have a certain disease (Positive) or not (Negative). Below is the actual data and model's predictions:

Patient	Actual Condition	Model Prediction
1	Positive	Positive TP.
2	Negative	Positive FP.
3	Positive	Negative FN
4	Negative	Negative TN
5	Positive	Positive TP.
6	Negative	Negative TN
7	Negative	Positive FP
8	Positive	Positive TP
9	Positive	Negative FN
10	Negative	Negative TN

Task 1: Construct the Confusion Matrix

Using the dataset above, fill in the confusion matrix below:

	Predicted Positive	Predicted Negative
Actual Positive	TP = 3	FN = 2
Actual Negative	FP = 2	TN = 3

Task 2: Compute Performance Metrics

Use the values from the confusion matrix to calculate the following metrics.

Accuracy: $\frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{3}{5}$

Sensitivity: : $\frac{TP}{(TP+FN)} = \frac{3}{5}$

Specificity: TN / (TN + FP) = $3/5$

Precision: Formula: 2 $* \frac{(Precision * Recall)}{(Precision + Recall)} = 3/5$

- Precision: $\frac{(TP)}{(TP+FP)} = 3/5$
- Recall: $\frac{TP}{(TP+FN)} = 3/5$