Module 2 and 3 Lab: Construct an Author Collaboration Graph

Welcome to the first lab for this course! In this lab, you will construct a weighted and undirected author collaboration graph from a subset of the Microsoft Academic Graph (MAG). You will also explore this author collaboration graph to examine its properties.

The input to the exercise is a set of academic publications, each of which is associated its authors. The input data is formatted as a tab-separated value file with two columns. The first column represents the paper Id and the second column represents one of its associated authors’ Ids. For example, below lists three papers’ author information.

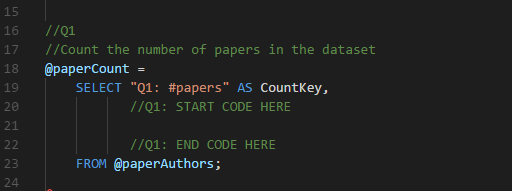
|  |  |
| --- | --- |
| PaperId | AuthorId |
| 120346 | 2604707830 |
| 120346 | 2708960186 |
| 161269 | 2034139186 |
| 161269 | 1984078314 |
| 161269 | 2079572972 |
| 161269 | 2163455850 |
| 672965 | 2403142114 |
| 672965 | 2630491343 |

The goal is to build a weighted and undirected author collaboration graph. Each edge is linked between two authors if they publish a paper together. The weight of each edge is defined as the number of collaborations between two authors. In this process, we will ignore papers with single authors.

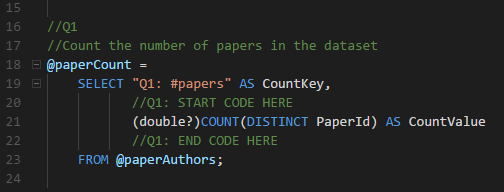
The output could be the weighted edge list of the constructed graph. Note that you can also use different data structures to represent the graph.

In this Lab assignment, please follow the Azure Data Lake (ADL) setup guide to upload the input academic publication data (Paper\_authors.tsv) into the ADL folder (course-kg/input/). The full file path is “/course-kg/input/Paper\_authors.tsv”.

The associated U-SQL script is “LabForModuleTwoThree.usql”. In this Lab, you will fill in the missing code in this script to answer the questions. The missing code is surrounded with “//Q1: START CODE HERE” and “//Q1: END CODE HERE” if taking Lab Question 1 as an example in the following code segment.



After filling in the missing code “(double?)COUNT(DISTINCT PaperId) AS CountValue”, this part of the code will be completed as follows:



To run the script, follow the Azure Data Lake setup guide. After filling in the missing code lines for all questions and then running the script, the author collaboration graph will be constructed and the answers to the questions will be output.

Lab Question 1

Add the following code to answer the question below.

(double?)COUNT(DISTINCT PaperId) AS CountValue

The number of papers in the input data: \_\_\_\_.

Answer 1: 880,571

Lab Question 2

Add the following code to answer the question below.

AVG(PaperCount \* 1.0) AS CountValue

How many papers does each author publish on average (choose the closest one)?

1. 2.37
2. 4.19
3. 11.38
4. 25.60
5. None of the above

Answer 2: A

Lab Question 3:

Following Question 2, pick up the correct code below and add it to get the average number of authors of each paper.

1. MAX(AuthorCount \* 1.0) AS CountValue
2. MIN(AuthorCount \* 1.0) AS CountValue
3. COUNT(AuthorCount \* 1.0) AS CountValue
4. AVG(AuthorCount \* 1.0) AS CountValue

Answer 3: D

Lab Question 4:

Add the following code to answer the question below.

GROUP BY AuthorA, AuthorB;

The number of undirected links in the author collaboration graph is \_\_\_\_\_.

Answer 4: 2,760,375

Lab Question 5:

Add the following code to answer the question below.

MAX(EdgeWeight) AS CountValue

The largest edge weight in this constructed graph is \_\_\_\_.

Answer 5: 149

Lab Question 6:

Add and run the following code to answer the question below.

(double?)COUNT(DISTINCT AuthorOne) AS CountValue

The number of nodes in this constructed graph is \_\_\_\_.

Answer 6: 1,113,841

Lab Question 7:

Following Questions 1 and 2, pick up the correct code below and add it to get the average number of authors of each paper.

1. MAX(NumNeighbors\* 1.0) AS CountValue
2. MIN(NumNeighbors\* 1.0) AS CountValue
3. COUNT(NumNeighbors\* 1.0) AS CountValue
4. AVG(NumNeighbors\* 1.0) AS CountValue

Answer 7: D

Lab Question 8:

Add the following code to answer the question below.

WHERE AuthorOne == 2114716943;

How many collaborators does the author with AuthorId = 2114716943 have?

1. 10
2. 50
3. 100
4. 200
5. None of above

Answer 8: D