Microsoft - DAT278x

From Graph to Knowledge Graph – Algorithms and Applications

LAB 1 – Graph

Construct an Author Collaboration Graph – Instructions

1. Objective

In this lab, we will familiarize you with the contents we have covered in module 2 and 3 about basic graph properties by

- constructing a weighted and undirected author collaboration graph from a subset of the Microsoft Academic Graph (MAG); and
- exploring the created author collaboration graph to examine its properties.

2. Input file

The input file describes Paper-Author relations [**Paper_authors.tsv**], and it is the first file used in the environment setup smoke test.

- It contains a set of academic publications; each publication is associated with its authors.
- The file is formatted as tab-separated value fields 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

3. Output graph – Author Collaboration Graph

What is an author collaboration graph? - It is a weighted and undirected graph.

- **Node**: each node represents an author;
- **Edge**: each edge represents co-author relationship between two authors (nodes); that is, if two authors have published a paper together, there is an edge between these two author nodes;
- **Edge weight**: the weight on edges is defined as the number of publications (collaborations) between two connecting authors (nodes).

With this definition in mind, those single author papers would be ignored during the author collaboration graph creation process.

4. How to Run the Lab

4.1 There is an associated U-SQL script *DAT278x_Lab1_Graph.usql.* Please download and save it.

4.2 Please follow the same steps in the Challenge Lab Setup Instruction, on how to upload the input file (Paper_author.tsv) into the ADL folder (course-kg/input/) – step 4.

The full input file path is "/course-kg/input/Paper_authors.tsv". **[Please note:** this is the **SAME** file that we used for Smoke Test run, you do **NOT** have to do it again if you passed the Smoke Test.**]**

4.3 Please follow the similar steps in step 5 from the Challenge Lab Setup Instruction, open the downloaded *DAT278x_Lab1_Graph.usql.*

4.4 You are expected to fill in the missing code in *DAT278x_Lab1_Graph.usql* script to answer the questions. The missing code is surrounded with some reminder comments like: "//Q1: START CODE HERE" and "//Q1: END CODE HERE" – if we use Question 1 as an example.

```
15
16  //Q1
17  //Count the number of papers in the dataset
18  @paperCount =
19      SELECT "Q1: #papers" AS Question,
20             //Q1: START CODE HERE
21
22             //Q1: END CODE HERE
23      FROM @paperAuthors;
24
```

4.5 For the missing code section, the missing part could be:

4.5.1 either provided directly in the question statement, so that you can copy and paste the code to the script; or

4.5.2 there are multiple choices for the candidate missing code snippets, please pick the right one which can answer the stated question; copy and paste the selected code to the script's expected position.

For example, for Lab1-Q1, after filling in the missing code

COUNT(DISTINCT PaperId) AS Answer

this part of the code will be completed as follows:

```
15
16  //Q1
17  //Count the number of papers in the dataset
18  @paperCount =
19      SELECT "Q1: #papers" AS Question,
20          //Q1: START CODE HERE
21           COUNT(DISTINCT PaperId) AS Answer
22          //Q1: END CODE HERE
23      FROM @paperAuthors;
24
```

4.6 After filling in the missing code lines for all questions, you can follow the Challenge Lab Setup Instruction – step 5.3 and 5.4, to run the script and check the results. The author collaboration graph will be constructed and the answers to the questions will be in the outputting file: *Results_Lab1_Graph.tsv*. Please use the answers in this file to finish all the multiple choices questions as well as numerical input questions in **Challenge Lab 1 on Graph – Questions** section.

5. Question structures and tips of running the script

5.1 Question structures

There are in total 8 questions:

- question 1-3 check the basic properties about the input paper-author relationships;
- question 4-8 construct the author collaboration graph and check its properties.

5.2 Tips of running the script

- for Q3 and Q7, to pick the right code snippet from the multiple choices. The corresponding numeric answers in *Results_Lab1_Graph.tsv* are just for your reference on the resulting values. You do NOT need to report those resulting values anywhere in **Challenge Lab 1 on Graph – Questions** section. However, please be careful about the ALIGNMENT on the Questions and Answers, especially for Q4 and Q8 – make sure you copy/choose the right answers for them.

- What's the best way to run the script and answer questions, all at once or one-by-one?
    - You can either pick all the answers and fill -- "//Q**[N]**: START CODE HERE" and "//Q**[N]**: END CODE HERE" (where **[N]** equals from 1 to 8) -- all 8 blocks at once, and then run the script once. This is the easier, suggested approach to run.
    - Or you can choose to comment out part of the scripts and only run sub-sections as you wish; so that you can finish all 8 questions in multiple runs. E.g. run 1-3 together, and then run 4-8 together.

    If you choose the second approach, please be careful on how to comment out sections; since there are logical/data dependency between code blocks. The second route is only recommended for more experience programmers; but it would help you understand deeper on what each code block is doing and the dependency they build upon each other.

    Do not worry about the computing cost, these jobs are small and cost very little compared with your free account credit quota.

6. More questions to think about

6.1 During author collaboration graph construction process, how do we eliminate those single author papers? Which mechanism guarantees such single-author papers are not included?

6.2 In below code snippet, why do we need to "union all" on these two swapping parts? Did we double count or not?

```
95   //Q6
96   //Get the number of collaborators each author has
97   @authorPairReciprocal =
98       SELECT AuthorA AS AuthorOne,
99              AuthorB AS AuthorTwo
100      FROM @authorPairWeight
101      UNION ALL
102      SELECT AuthorB AS AuthorOne,
103             AuthorA AS AuthorTwo
104      FROM @authorPairWeight;
105
```

6.3 Can you think of a different way to get the number of collaborators each author has, without using the "union all" on the two swapping parts?

We hope you enjoy this lab and learn more about the author collaboration graph construction and properties!