Introduction to Graph Databases

Activity – Implementing Graphs using Relational Databases

Google released, in 2002, a subset of the structure of the WWW. In this dataset, web pages are represented by graph nodes such that when a web page A contains a hyperlink to web page B, a directed edge is created from node A to node B.

In this activity, we will focus on the performance of different queries. Therefore, we will use three PostgreSQL tables, which are subsets of different sizes of the web structure released by Google:

- **webgraph1** table: 605 nodes (web pages) and 1521 edges (hyperlinks)
- **webgraph2** table: 1622 nodes (web pages) and 6288 edges (hyperlinks)
- **webgraph3** table: 4122 nodes (web pages) and 14356 edges (hyperlinks)

If you are using your own local PostgreSQL database, you need to create and populate it. Use the three files provided.

```
CREATE TABLE  webgraph1 (fromnode int,  tonode int);

COPY webgraph1   FROM   'X:..\webgraph1.txt'


CREATE TABLE  webgraph2 (fromnode int,  tonode int);

COPY webgraph2   FROM   'X:..\webgraph2.txt'


CREATE TABLE  webgraph3 (fromnode int,  tonode int);

COPY webgraph3   FROM   'X:..\webgraph3.txt'
```

Below, we show the list of different uses cases we want to analyze:

Use Case A: For each pair of connected nodes, find the **1-hop paths**. Include four columns in the resultset: **fromNode, toNode, length, path**, which correspond to the source node, target node, length of the path, and visited nodes, respectively. Exclude repeated nodes in the path. That is, if A -> A, do not consider that A, A, 1, A-A is a valid 1-path from A to A.

Use Case B: For each pair of connected nodes, find the **2-hop paths**. Include four columns in the resultset: **fromNode, toNode, length, path**, which correspond to the source node, target node, length of the path and the visited nodes, respectively. Exclude repeated nodes in the path. That is, if A -> B and B->A, do not consider that A, A, 2, A-B-A is a valid 2- path from A to A.

Use Case C: For each pair of connected nodes, find the **3-hop paths**. Include four columns in the resultset: **fromNode, toNode, length, path**, which correspond to the source node, target node, length of the path and the visited nodes, respectively. Exclude repeated nodes in the path. That is, if A -> B, A->C and B->A, do not consider that A, C, 3, A-B-A-C is a valid 3-path from A to C.

Use Case D: for each pair of connected nodes, find the **N-hop paths (the value of N is not known in advance)**. Include four columns in the result set: **fromNode, toNode, length, path**, which correspond to the source node, target node, length of the path and the visited nodes, respectively. Exclude repeated nodes in the path like in case "C".

Exercise 1

1.1) Show the SQL queries that solve each use case. Use an alias for the table, such that it can be easily rewritten for different table names.

Use Case	SQL (use an alias for the table in the from clause)
A (1-hop)	
B (2-hop)	
C (3-hop)	
D (N-hop)	

1.2) Run each query in Part 1.1., against tables of different sizes. For each run, record the execution time and the size of the result set (number of tuples), and complete the following comparative tables.

Use Case A (1-hop)		
Table	Execution Time (msec)	Resultset Size (#tuples)
1521 tuples		
6288 tuples		
14356 tuples		

Use Case **B** (2-hop)		
Table	Execution Time (msec)	Resultset Size (#tuples)
1521 tuples		
6288 tuples		
14356 tuples		

Use Case C (3-hop)		
Table	Execution Time (msec)	Resultset Size (#tuples)
1521 tuples		
6288 tuples		
14356 tuples		

Use Case D (N-hop)		
Table	Execution Time (msec)	Resultset Size (#tuples)
1521 tuples		
6288 tuples		
14356 tuples		

Exercise 2

2.1) Probably, some queries above will run indefinitely. Analyze the strategy used by PostgreSQL for executing each query. More precisely, find out the query plan chosen for the largest table (webgraph3), for each one of the use cases. Complete the following table.

Use case	Query Plan for webgraph3
A (1-hop)	
B (2-hop)	
C (3-hop)	
D (N-hop)	

2.2) Briefly sketch the idea behind the query plans proposed by PostgreSQL .

Exercise 3

In Exercise 1, probably **N-hop** queries over table **webgraph3** ran indefinitely. However, note that a 3-hops query could also be solved by an SQL recursive query limited to N=3.

3.1) Rewrite the **recursive** SQL query **over the webgraph3 table**, limiting it to retrieve **only 3-Hops.** Verify the result, checking that you obtained the same results as with the 3-hop non-recursive version, that is, your SQL recursive query limited to N=3 is equivalent to a non-recursive triple join query.

Answer:

3.2) Now, we want to study if the execution time could be improved using indexes. We will take into consideration both query variations, i.e. 3-Hops (triple join), and SQL recursive limited to obtain 3-Hops.

3.2.1) Which index could be useful for avoid the TableScan+Sort? Which index would be useful for the Merge Operator? Write the SQL syntax for creating the index/es proposed.

Answer:

3.2.2) Create the index/es proposed. Run both queries. Compare the result against the 3-hop performance in exercise 1. Complete the following table. Did you obtain any improvement?

Use Case C (3-hop)		
Table with index/es	Execution time (sec)	Resultset size (#tuples)
14356 tuples		

Use Case D (N-hop) - SQL recursive limited by N=3		
Table with index/es	Execution time (sec)	Resultset size (#tuples)
14356 tuples		

3.3.3) Analyze the query plan for both queries (with indexes). Complete the following table:

Use case	Query plan with index/es for webgraph3
C (3-hop)	
D (N-hop) limited by N=3	

3.3.4) Did PostgreSQL use the same strategy in both queries? Explain the reason in detail.

Answer

Exercise 4

When looking for N-paths, where N is large (>3) or unknown, we need to use Recursive SQL. Execute the recursive SQL query and find the **longest N value** that allows us to obtain a result set in **less than 5 minutes**, for the **webgraph3** table.

Answer: