Assignment 2 (Advanced SQL)

Manohar Kaul

March 6, 2022

Given below are a list of queries for the final database you built of scientific publications. For all the queries below, if there is ”missing data”, retrieve the values with NULLs in them! Only if the key to join on is NULL, then it is ok to leave data out.
Hint: Think of this scientific database as a *directed graph* (digraph) of papers citing other papers and papers being cited by other papers.

1) For every paper, list the papers that cite it. All relevant details of the papers that cite should also be provided, e.g., author name(s), title, venue, etc.

2) For every paper, list the papers that it cites. With all relevant details, same as above.

3) For every paper X, list the *second level* of papers that cite paper X. For example, if paper X is cited by paper Y, and paper Y is in turn cited by paper Z, then the pair of papers (X,Z) should be listed! Again provide all necessary details.

4) List the top-20 most cited papers.

5)List all the pairs of authors who have co-authored papers more than once!

6)A *k*-clique in a graph *G* is a *complete subgraph* with *k* vertices within *G*. Detect all “triangles” (3-cliques) between authors in this citation digraph. List a triple of authors (X,Y,Z) along with the counts of how many times this occurs. The order between X,Y, and Z doesn’t matter, which means you must treat pairs (X,Y,Z), (Z,X,Y), (X,Z,Y) as a single pair and count the total occurrences. Also, the relationships between any two authors X and Y can be of “X cited by Y” or “X cites Y” and are treated as being similar.

Why are we doing this? We want to detect groups of authors that keep citing each other’s works to boost their citations! Sometimes also called “clique-farms”.