Reg. No: 2016/21/D/ST6/01485; Principal Investigator: dr Szymon Abram Toruńczyk

Limits of Databases

In this project, we will study an important problem from database theory, called the evaluation problem for conjunctive queries. An example conjunctive query is

$$Q(x, y, z) = E(x, y) \wedge E(y, z) \wedge E(z, x),$$

which selects all triples (x, y, z) of data values from a given database D, such that the pairs (x, y), (y, z) and (z, x) all belong to the table E of the database.

Evaluating such queries quickly is an important task in commercial database systems, which deal with very big databases, storing for example $N = 10^6 = 1000000$ pairs of data values. A naive evaluation algorithm for computing the result of the above query would check all possible combinations (x, y) and (y', z) of pairs of pairs occurring in the table E, and for each such pair, if $y' = y$ and (z, x) belongs to the table E, the algorithm would add the triple (x, y, z) to the set of results. In this case, the algorithm would scan through $N^2 = 10^{12}$ pairs of rows of E. Assuming that the computers processor can process 10^9 operations per second (this is the speed of a 1GHz processor), computing all the results would last for 1000 seconds.

It turns out that much more efficient algorithms can be constructed. The first observation towards such an algorithm is that for any database D whose table E contains N rows, the number of results to the query Q is at most $N^{3/2}$. This is a nontrivial theorem, which can be formulated in two other equivalent ways:

- A graph with N edges contains at most $N^{3/2}$ triangles;
- A solid in three-dimensional space, whose projections onto the planes xy, yz and xz have surface area at most N, has volume at most $N^{3/2}$.

The number $3/2$ in the above theorem is optimal, and depends only on the chosen query Q. This means that there exist databases D having N rows in table E, for which the above query Q results in $N^{3/2}$ triples. It follows that, in the worst case, the running time of an optimal algorithm computing the result of the query Q on the database D with N rows, must be at least proportional to the number $N^{3/2}$, since the algorithm must at least process its results. It turns out that such an optimal algorithm can indeed be constructed. For $N = 10^6$, this algorithm would run 1 second, which is a huge difference comparing to the naive evaluation.

In this project, we will attempt to construct an algorithm which tries to exploit additional statistical information about the database, in order to run even faster. Our approach to the problem will rely on methods from combinatorics and measure theory. Instead of considering very large databases, we will study infinite, limit databases, and we will prove mathematical results about them