

# Maximum number of independence atoms in a relational database

Attila Sali

January 20, 2025

This research problem deals with some combinatorial question of relational databases. For our purposes the database is simply a matrix, whose columns are the *attributes*, the rows are the individual records of data, frequently *tuples*. For example consider Table 1, where the attributes are named.

Course Name	Year	Lecturer	Credits	Semester
Mathematics	2019	Cornelius	5	1
Datamining	2018	Sarah	7	2
Theory of Algorithms	2019	Sarah	7	2

Table 1: A sample database table.

Independence and conditional independence are fundamental concepts in areas as diverse as artificial intelligence, probability theory, social choice theory, and statistics. In this problem we investigate a special case of independence, namely independence atoms. Intuitively, a relation (table)  $r$  satisfies the independence atom  $X \perp Y$  between two disjoint sets  $X$  and  $Y$  of attributes, if for all tuples  $t_1, t_2 \in r$  there is some tuple  $t \in r$  which matches the values of  $t_1$  on all attributes in  $X$  and matches the values of  $t_2$  on all attributes in  $Y$ . In other words, in relations that satisfy  $X \perp Y$ , the occurrence of  $X$ -values is independent of the occurrence of  $Y$ -values.

*Example 1.* Consider a simple database schema that stores information about the enrolment of students into a fixed course. In fact, the schema records for each enrolled student, the year in which they completed a prerequisite course. More formally, we have the schema (set of attributes)  $\text{ENROL} = S(\text{student}), P(\text{prerequisite}), Y(\text{year})$ . Intuitively, every student must have completed every prerequisite in some year. For this reason, for any value in the *Student* column and every value in the *Prerequisite* column there is some year for when this student has completed that prerequisite. That is, the values in the *Student* column are independent of the values in the *Prerequisite* column. A snapshot relation  $r$  over ENROL may be:

<i>Student</i>	<i>Prerequisite</i>	<i>Year</i>
Turing	Math201	1932
Gödel	Math201	1925
Turing	Phys220	1932
Gödel	Phys220	1925

A collection  $\Sigma$  of independence atoms imply independence atom  $\sigma$ , denoted by  $\Sigma \models \sigma$ , if every table that satisfies  $\Sigma$ , also satisfies  $\sigma$ . This implication can be described equivalently by derivation rules, as follows. Let  $X, Y, Z$  be subsets of attributes.

1.  $X \perp \emptyset$  holds always.
2.  $X \perp Y \Rightarrow Y \perp X$  (symmetry).
3.  $X \perp YZ \Rightarrow X \perp Y$  (decomposition).
4.  $X \perp Y \wedge XY \perp Z \Rightarrow X \perp YZ$  (exchange).

So we have that  $\Sigma \models \sigma$  iff  $\sigma$  can be derived starting from independence atoms in  $\Sigma$  using a finite number of applications of the rules above.  $XY$  means the union of attribute sets  $X$  and  $Y$ , similarly  $YZ$  is the union of  $Y$  and  $Z$ . (*In general if two sets of attributes are written next to each other then it denotes the union of the two sets. This is traditional notation in the area.*)

The main concept is non-redundance of set of independence atoms.  $\Sigma$  is *non-redundant* iff no  $\sigma \in \Sigma$  can be derived from  $\Sigma \setminus \{\sigma\}$ , that is no independence atom in  $\Sigma$  can be derived from the other atoms in  $\Sigma$ .

**Research problem** We want to find  $f(n)$  the largest size of a non-redundant collection of independence atoms in a schema of  $n$  attributes.

**What do we know?** One must observe that none of the derivation rules introduces new attributes, so  $\{X_i \perp Y_i : |X_i \cup Y_i| = \lceil \frac{n}{2} \rceil, i = 1, 2, \dots, \binom{n}{\lceil \frac{n}{2} \rceil}, X_i \cup Y_i \neq X_j \cup Y_j \text{ for } i \neq j\}$  with none of  $X_i$  and  $Y_i$  being nonempty, is non-redundant. Thus,  $f(n) \geq \binom{n}{\lceil \frac{n}{2} \rceil}$ .

There is a divide-and-conquer algorithm that decides for a given collection  $\Sigma$  of independence atoms and another independence atom  $\sigma$  whether  $\Sigma$  implies  $\sigma$ . It is implemented in python, if someone wants to try it, write me an email.

We have better constructions than the one indicated above. This shows that  $f(5) \geq 12 > \binom{5}{2}$ .

### Qualifying questions

1. Show that  $f(3) = 6$ .
2. Show that  $f(5) \geq 11$ .
3. Show that  $f(5) \geq 12$ .

4. Show that for pairwise disjoint attribute sets  $U, V, W, W', W'', Y$  the following implication holds.  $UV \perp WW' \wedge UWW'' \perp Y \Rightarrow U \perp YW$ .

Questions, solutions should be directly sent to [saliattila@gmail.com](mailto:saliattila@gmail.com)