Wednesday 14^{th} March, 2018

NP-hardness of k-QID

k-QID is reduced to MINIMUM COVER.

[SP5] MINIMUM COVER

INSTANCE: Collection C of subsets of a finite set S, positive integer $K \leqslant |C|$.
QUESTION: Does C contain a cover for S of size K or less, i.e., a subset $C' \subseteq C$ with $|C'| \leqslant K$ such that every element of S belongs to at least one member of C'?

k-QID

INSTANCE: A finite set of m-tuples of database $\mathcal{D}$ with an m-tuple of column names $\mathcal{C}$ and an m-tuple of finite sets of column domains $\mathcal{F}$, such that $\mathcal{D} \subseteq \bigtimes_{F \in \mathcal{F}} F$ and, for all $F \in \mathcal{F}$, $|\mathcal{D}| \leqslant |F|$; a positive integer k.
QUESTION: Is there a finite set $I \subseteq \text{set}(\mathcal{C})$ with $|I| = k$ such that $|\Pi_I D| = |D|$?

Construction

Given an instance of **MINIMUM COVER**, construct a database $\mathcal{D}$ with $\text{tuple}(C)$ as columns and $\text{tuple}(S \cup \{\perp\})$ as rows. For the content of the database, let

$$\mathcal{D}[i][j] = \begin{cases} i, & \text{if } \text{tuple}(S)[i] \in \text{tuple}(C)[j] \\ \perp, & \text{otherwise} \end{cases}$$

where $\mathcal{D}[i][j]$ means the jth element in the ith tuple in $\mathcal{D}$ and $\perp$ is a fresh symbol not in S.

Proof

Here's the proof that the given instance satisfies **MINIMUM COVER** iff the constructed instance satisfies **k-QID**.

Suppose there is such subset C' of C such that every element in S is in at least one set of C'. Project the columns corresponding to C' in the database. Since every element is in at least one set of C', every row except the $\perp$-row has at least one element unique to that row, so there are no duplicate rows in $\Pi_{C'}\mathcal{D}$; thus, the number of rows in the projected database $\Pi_{C'}\mathcal{D}$ is equal to the original database $\mathcal{D}$.

Suppose there is some subset of columns N such that the number of rows in the projected database $\Pi_N\mathcal{D}$ is equal to the original database $\mathcal{D}$. Per the assumption, it immediately follows that there are no duplicate rows in $\Pi_N\mathcal{D}$; furthermore, per the construction rule of the database, every row must have at least one non-$\perp$ element in some column in N. Thus, choose those subsets C' corresponding to N from C, and every element of S must be in at least one set in C'.

Example

Let $k = 2$, $S = \{a, b, c, d, e\}$, and $C = \{c_1, c_2, c_3, c_4\}$ with $c_1 = \{a, b, c\}, c_2 = \{b, d\}, c_3 = \{c, d\}, c_4 = \{d, e\}\}$. There is a solution $C' = \{c_1, c_4\} = \{\{1, 2, 3\},\ \{4, 5\}\}$.

The constructed database for this instance is:

$$
\begin{array}{llllll}
 & c_1 & c_2 & c_3 & c_4 & \\
\mathcal{D} = \{ & (1, & \perp, & \perp, & \perp), & a \\
 & (2, & 2, & \perp, & \perp), & b \\
 & (3, & \perp, & 3, & \perp), & c \\
 & (\perp, & 4, & 4, & 4), & d \\
 & (\perp, & \perp, & \perp, & 5), & e \\
 & (\perp, & \perp, & \perp, & \perp)\} & \perp
\end{array}
$$

If we project the solution, i.e., $\Pi_{c_1,c_4}\mathcal{D}$, we get:

$$
\begin{array}{lll}
 & c_1 & c_4 & \\
\mathcal{D} = \{ & (1, & \perp), & a \\
 & (2, & \perp), & b \\
 & (3, & \perp), & c \\
 & (\perp, & 4), & d \\
 & (\perp, & 5), & e \\
 & (\perp, & \perp)\} & \perp
\end{array}
$$

where no two rows are identical, meaning every element is in at least one of c_1, c_4.

If we project something incorrect, e.g., $\Pi_{c_1,c_2,c_3}\mathcal{D}$, we get:

$$
\begin{array}{lllll}
 & c_1 & c_2 & c_3 & \\
\mathcal{D} = \{ & (1, & \perp, & \perp), & a \\
 & (2, & 2, & \perp), & b \\
 & (3, & \perp, & 3,), & c \\
 & (\perp, & 4, & 4,), & d \\
 & (\perp, & \perp, & \perp), & e \\
 & (\perp, & \perp, & \perp)\} & \perp
\end{array}
$$

where the e-row and $\perp$-row are identical, meaning e is not in any of c_1, c_2, c_3.