Corrigendum to “Counting Database Repairs that Satisfy Conjunctive Queries with Self-Joins”

Jef Wijsen
University of Mons, Belgium

Abstract
The helping Lemma 7 in [Maslowski and Wijsen, ICDT, 2014] is false. The lemma is used in (and only in) the proof of Theorem 3 of that same paper. In this corrigendum, we provide a new proof for the latter theorem.

1 The Flaw

The helping Lemma 7 in [MW14] is false. A counterexample is given next.

Example 1. For $S = \{R, S\}$ and $q = \{R(x, y), S(y)\}$, we have $\text{enc}_S(q) = \{N(R, x, y), N(S, y, 0)\}$. From [MW14, Lemma 8], it follows that $\sharp\text{CERTAINTY}(\text{enc}_S(q))$ is $\sharp\text{P}$-hard. From [MW13, Theorem 4], it follows that $\sharp\text{CERTAINTY}(q)$ is in $\text{FP}$. Consequently, assuming $\sharp\text{P} \neq \text{FP}$, there exists no polynomial-time many-one reduction from $\sharp\text{CERTAINTY}(\text{enc}_S(q))$ to $\sharp\text{CERTAINTY}(q)$. Lemma 7 in [MW14] is thus false.

The first part in the proof of Lemma 7 in [MW14] is correct; it shows a polynomial-time many-one reduction from $\sharp\text{CERTAINTY}(q)$ to $\sharp\text{CERTAINTY}(\text{enc}_S(q))$. However, the second part in that proof is flawed when it claims “We can compute in polynomial time the (unique) database $db'_0$ with schema $S$ such that $\text{enc}_S(db'_0) = db_0$.” The flaw is that the database $db'_0$ does not generally exist, as shown next. Let $S = \{R, S\}$ and $q = \{R(x, y), S(y)\}$, as in Example 1. Then, $\text{enc}_S(q) = \{N(R, x, y), N(S, y, 0)\}$. A legal input to $\sharp\text{CERTAINTY}(\text{enc}_S(q))$ is $db_0 = \{N(R, b, c), N(S, c, 0), N(S, c, 1)\}$. However, there exists no database $db'_0$ such that $\text{enc}_S(db'_0) = db_0$. Indeed, for every database $db'_0$ with schema $S$, if $N(S, c, s) \in \text{enc}_S(db'_0)$, then $s = 0$.

2 The Solution

The following treatment is relative to a database schema $S$. Let $k, m$ be non-negative integers such that every relation name in $S$ has at most $k$ primary-key positions, and at most $m$ non-primary-key positions. We define a new function $\text{enc}'_S(q)$ which encodes Boolean conjunctive queries $q$ into unirelational Boolean conjunctive queries. For $\text{enc}'_S(q)$, we use a fresh relation name $N$ with $k + 1$ primary-key positions, and $m$ non-primary-key positions. For every atom $R(x, y)$ in $q$, the query $\text{enc}'_S(q)$ will contain some atom $N(R, x, 0, y, z)$, where $0$ is a sequence of padding zeros, and $z$ is a sequence of padding fresh variables, all distinct and not occurring elsewhere. This encoding is different from [MW14, Definition 3] where a sequence of padding zeros was used instead of $z$. 

1
Example 2. We illustrate the difference between the old encoding $\text{enc}_S(\cdot)$ of [MW14, Definition 3] and the newly proposed encoding $\text{enc}_S^*(\cdot)$. For $q_0 = \{R(x, y), S(y)\}$, we have

\[
\begin{align*}
\text{enc}_S(q_0) &= \{N(R, x, y), N(S, y, 0)\}, \\
\text{enc}_S^*(q_0) &= \{N(R, x, y), N(S, y, z)\}.
\end{align*}
\]

We recall from [MW14, p. 156] that the complex part of a Boolean conjunctive query contains every atom $F \in q$ such that some non-primary-key position in $F$ contains either a variable with two or more occurrences in $q$ or a constant. Note that $N(S, y, 0)$ belongs to the complex part of $\text{enc}_S(q_0)$, while $N(S, y, z)$ is not in the complex part of $\text{enc}_S^*(q_0)$.

Definition 1. We define $\text{skBCQ}$ as the class of Boolean conjunctive queries in which all relation names are simple-key. We say that a query $q \in \text{skBCQ}$ is minimal if both

- $q$ contains no two distinct atoms $R_1(x_1, y_1)$, $R_2(x_2, y_2)$ such that $R_1 = R_2$ and $x_1 = x_2$; and
- there exists no substitution $\theta$ over $\text{vars}(q)$ such that $\theta(q) \subsetneq q$.

We define $\text{cxBCQ}$ as the class of unirelational Boolean conjunctive queries $q$ whose relation name has signature $[n, 2]$ (for some $n \geq 2$) such that for every $F \in q$, the first position of $F$ is a constant.

Definition 2. The intersection graph of a Boolean conjunctive query is an undirected graph whose vertices are the atoms of $q$. There is an undirected edge between any two atoms that have a variable in common.

Lemma 1. Assume $\sharp\text{P} \neq \text{FP}$. For every minimal query $q$ in $\text{skBCQ}$, if $\sharp\text{CERTAINTY}(\text{enc}_S^*(q))$ is $\sharp\text{P}$-hard, then so is $\sharp\text{CERTAINTY}(q)$.

Proof. Let $q$ be a minimal query in $\text{skBCQ}$ such that $\sharp\text{CERTAINTY}(\text{enc}_S^*(q))$ is $\sharp\text{P}$-hard. Note that $q$ does not need to be unirelational or self-join-free. The query $\text{enc}_S^*(q)$, which is unirelational, is a legal input to the function $\text{IsEasy}$ of [MW14, p. 163] since $\sharp\text{CERTAINTY}(\text{enc}_S^*(q))$ is $\sharp\text{P}$-hard, the function $\text{IsEasy}$ will return $\text{false}$ on input $\text{enc}_S^*(q)$. This function will repeat, as long as possible, the following step: pick some atom $N(R, c, y)$ and some variable $y \in \text{vars}(y)$, with $R$ some relation name (treated as a constant) and $c$ some constant, and replace all occurrences of $y$ with an arbitrary constant. Let $\bar{q}$ be the query that results from these steps. Clearly, for every atom $N(R, s, \bar{t})$ in $\bar{q}$, either $s$ is a constant or $\bar{t}$ is variable-free. Since $\text{IsEasy}$ returns $\text{false}$ on input $\bar{q}$, it follows that $\bar{q}$ does not satisfy the premise of [MW14, Lemma 5]. Therefore, it must be the case that $\bar{q}$ contains two distinct atoms $N(R, x, \bar{u})$ and $N(S, y, \bar{w})$ that are connected in the intersection graph of $\bar{q}$ such that

- $R$ and $S$ are relation names (serving as constants), not necessarily distinct;
- $x$ and $y$ are distinct variables; and
- neither $\bar{u}$ nor $\bar{w}$ is exclusively composed of variables occurring only once in the query. That is, $N(R, x, \bar{u})$ and $N(S, y, \bar{w})$ belong to the complex part of $\bar{q}$.\footnote{For uniformity of notation, we will assume that the unirelational query uses relation name $N$.}
For every relation name $R$ that appears in $q$, we assume fresh relation names $R_1, R_2, R_3, \ldots$ with the same signature as $R$. Using these relation names, we can construct a self-join-free Boolean conjunctive query $q'$ such that $|q'| = |q|$ and for every atom $R(x, y)$ in $q$, the query $q$ contains some atom $R_1(x, y)$. For example, if $q = \{R(x, y), R(y, z), S(z, x)\}$, then we can let $q' = \{R_1(x, y), R_2(y, z), S_1(z, x)\}$. It can now be shown that the function IsSafe in [MW14, p. 158] will return false on input $q'$, and thus $\zeta\text{CERTAINTY}(q')$ is $\sharp\text{P}$-hard. Indeed, whenever IsEasy picked $N(R, c, \overline{y})$ and some variable $y \in \text{vars}(\overline{y}) \cap \text{vars}(q)$, the function IsSafe can execute SE3 on the corresponding $R$-atom of $q'$. This eventually leads to a query whose complex part contains two atoms $R_i(\overline{x}, \overline{w})$ and $S_j(\overline{y}, \overline{w})$, $x \neq y$, that are connected in the intersection graph, at which point IsSafe will return false. In this reasoning, one needs that non-primary-key positions are padded with fresh variables occurring only once, as can be seen from Example 2

In the remainder of this proof, we show the existence of a polynomial-time many-one reduction from $\zeta\text{CERTAINTY}(q')$ to $\zeta\text{CERTAINTY}(q)$. We incidentally note that the remaining reasoning, which generalizes the proof of [MW14, Lemma 1], does not require that relation names are simple-key.

Let $f$ be a mapping from facts to facts such that for every atom $R_i(x_1, \ldots, x_n) \in q'$, for every $R_i$-fact $A := \langle a_1, \ldots, a_n \rangle$, $f(A) := R(\langle a_1, x_1 \rangle, \ldots, \langle a_n, x_n \rangle)$. Notice that $f$ maps $R_i$-facts to $R$-facts. Here, every couple $\langle a_i, x_i \rangle$ denotes a constant such that $\langle a_i, x_i \rangle = \langle a_j, x_j \rangle$ if and only if both $a_i = a_j$ and $x_i = x_j$. Moreover, if $c$ is a constant, then $\langle c, c \rangle := c$. Since no two distinct atoms of $q$ agree on both their relation name and primary key, it will be the case that for all facts $A$ and $B$, $A \sim B$ if and only if $f(A) \sim f(B)$, where $\sim$ denotes “is key-equal-to.”

We extend the function $f$ in the natural way to databases $\text{db}$ that use only relation names from $q'$: $f(\text{db}) := \{f(A) \mid A \in \text{db}\}$. Clearly, $f(\text{db})$ can be computed in polynomial time in the size of $\text{db}$. Let $\text{db}$ be a set of facts with relation names in $q'$. It can be easily seen that $|\text{rset}(\text{db})| = |\text{rset}(f(\text{db}))|$ and $\text{rset}(f(\text{db})) = \{f(r) \mid r \in \text{rset}(\text{db})\}$. Let $r$ be an arbitrary repair of $\text{db}$. It suffices to show that

$$r \models q' \iff f(r) \models q.$$  

For the implication $\implies$, assume that $r \models q'$. We can assume a valuation $\theta$ over $\text{vars}(q')$ such that $\theta(q') \subseteq r$. Let $\mu$ be the valuation such that for every variable $x \in \text{vars}(q')$, $\mu(x) = \langle \theta(x), x \rangle$. By our construction of $q'$ and $f$, it will be the case that $\mu(q) \subseteq f(r)$, thus $f(r) \models q$.

For the implication $\impliedby$, assume that $f(r) \models q$. We can assume a valuation $\theta$ over $\text{vars}(q)$ such that $\theta(q) \subseteq f(r)$. Notice that if $c$ is a constant in $q$, then it must be the case that $\theta(c) = \langle c, c \rangle := c$. We define $\theta_L$ as the substitution that maps every variable $x$ in $\text{vars}(q)$ to the first coordinate of $\theta(x)$; and $\theta_R$ maps every $x$ to the second coordinate of $\theta(x)$. It is convenient to think of $L$ and $R$ as references to the Left and the Right coordinates, respectively. Thus, by definition, $\theta(x) = \langle \theta_L(x), \theta_R(x) \rangle$.

By inspecting the right-hand coordinates of couples $\langle a_i, x_i \rangle$ in $f(r)$, it can be easily seen that $\theta(q) \subseteq f(r)$ implies $\theta_R(q) \subseteq q$. Since the query $q$ is minimal, it follows that $\theta_R(q) = q$, i.e., $\theta_R$ is an automorphism. Since the inverse of an automorphism is an automorphism, $\theta_R^{-1}$ is an automorphism as well. Note that $\theta_R$ will be the identity on constants that appear in $q$. We now define $\mu := \theta_L \circ \theta_R^{-1}$ (i.e., $\mu$ is the composed function $\theta_L$ after the inverse of $\theta_R$), and show that $\mu(q') \subseteq r$, which implies the desired result that $r \models q'$. To this extent, let $R_i(x_1, \ldots, x_n)$ be an arbitrary atom of $q'$. It suffices to show $R_i(\mu(x_1), \ldots, \mu(x_n)) \in r$, which can be proved as follows:

$$R_i(\mu(x_1), \ldots, \mu(x_n)) \in q.$$

Thus, since $\theta_R^{-1}$ is an automorphism,
Since $\theta(q) \subseteq f(r)$,
\[ R \left( \theta(\theta_R^{-1}(x_1)), \ldots, \theta(\theta_R^{-1}(x_n)) \right) \in f(r). \]
Since, for every symbol $s$, $\theta(s) = \langle \theta_L(s), \theta_R(s) \rangle$ and $\theta_R(\theta_R^{-1}(s)) = s$, we obtain
\[ R \left( \langle \theta_L(\theta_R^{-1}(x_1)), x_1 \rangle, \ldots, \langle \theta_L(\theta_R^{-1}(x_n)), x_n \rangle \right) \in f(r). \]
That is, by our definition of $\mu$,
\[ R \left( \langle \mu(x_1), x_1 \rangle, \ldots, \langle \mu(x_n), x_n \rangle \right) \in f(r). \]
From this, it is correct to conclude that $R_i(\mu(x_1), \ldots, \mu(x_n)) \in r$. This concludes the proof. 

**Lemma 2.** For every Boolean conjunctive query $q$, there exists a polynomial-time many-one reduction from $\sharpCERTAINTY(q)$ to $\sharpCERTAINTY(\text{enc}_S(q))$.

**Proof.** Let $q$ be a Boolean conjunctive query. Let $R$ be a relation name that occurs in $q$. Let \( \{R(\bar{x}_i, \bar{y}_i)\}_{i=1}^m \) be the set of $R$-atoms of $q$. Then, $\text{enc}_S(q)$ will contain, for every $i \in \{1, \ldots, m\}$, some atom $N(R, \bar{x}_i, \bar{y}_i, \bar{z}_i)$, where $\bar{z}_i$ is a (possibly empty) sequence of distinct fresh variables not occurring elsewhere. For every $R$-fact $A := R(\bar{a}, \bar{b})$, we define $f(A) := N(R, \bar{a}, \bar{0}, \bar{b}, \bar{0})$. Note here that $f(A)$ depends on the signatures of $R$ and $N$, but not on the $R$-atoms of $q$. The mapping $f$ is defined similarly for all relation names that appear in $q$. It can be easily seen that for all facts $A$ and $B$ whose relation names appear in $q$, $A \sim B$ if and only if $f(A) \sim f(B)$.

If $db$ is an instance of $\sharpCERTAINTY(q)$, we can assume without loss of generality that every relation name in $db$ also appears in $q$. We extend the function $f$ to such instances $db$ of $\sharpCERTAINTY(q)$: $f(db) := \{f(A) \mid A \in db\}$. Obviously, $f(db)$ can be computed in polynomial time in the size of $db$. It is also obvious that $|\text{set}(db)| = |\text{set}(f(db))|$ and $\text{set}(f(db)) = \{f(r) \mid r \in \text{set}(db)\}$. It suffices to show that for every repair $r$ of $db$,
\[ r \models q \iff f(r) \models \text{enc}_S(q). \]

For the implication $\implies$, assume $r \models q$. We can assume a valuation $\theta$ over $\text{vars}(q)$ such that $\theta(q) \subseteq r$. Let $\theta'$ be the valuation that extends $\theta$ from $\text{vars}(q)$ to $\text{vars}(\text{enc}_S(q))$ such that $\theta'(z) = 0$ for every variable $z$ that appears in $\text{enc}_S(q)$ but not in $q$. By the construction of $f$, it will be the case that $\theta'(\text{enc}_S(q)) \subseteq f(r)$. Indeed, if $\text{enc}_S(q)$ contains $N(R, \bar{x}_i, \bar{0}, \bar{y}_i, \bar{z}_i)$, then $r$ will contain $R(\theta(\bar{x}_i), \theta(\bar{y}_i))$, hence $f(r)$ will contain $N(R, \theta'(\bar{x}_i), \bar{0}, \theta'(\bar{y}_i), \bar{0})$ and $\theta'(\bar{z}_i) = 0$.

For the implication $\impliedby$, assume $f(r) \models \text{enc}_S(q)$. We can assume a valuation $\theta$ over $\text{vars}(\text{enc}_S(q))$ such that $\theta(\text{enc}_S(q)) \subseteq f(r)$. It is straightforward to see that $\theta(q) \subseteq r$. 

We now give the new proof for Theorem 3 in [MW14].

**Theorem 1 ([MW14 Theorem 3]).** The set \{\$\sharpCERTAINTY(q) \mid q \in \text{skBCQ}\} exhibits an effective \$\text{FP-\sharpP}$-dichotomy.

**New proof.** Let $q \in \text{skBCQ}$. It can be decided whether $q$ can be satisfied by a consistent database. If $q$ cannot be satisfied by a consistent database, then for every database $db$, the number of repairs of $db$ satisfying $q$ is 0. An example is $q = \{R(x, 0), R(x, 1)\}$. Assume next that $q$ can be satisfied by a consistent database. Then, we can compute a minimal query $q_m$ such that for every database,
the number of repairs satisfying $q_m$ is equal to the number of repairs satisfying $q$. That is, the problems $\sharp\text{CERTAINTY}(q_m)$ and $\sharp\text{CERTAINTY}(q)$ are identical.

Then, $\text{enc}_S^*(q_m)$ belongs to $\text{cxBCQ}$. By [MW14, Lemma 8], the set $\{\sharp\text{CERTAINTY}(q) \mid q \in \text{cxBCQ}\}$ exhibits an effective $\text{FP}$-$\text{♯P}$-hard dichotomy. If the problem $\sharp\text{CERTAINTY}(\text{enc}_S^*(q_m))$ is in $\text{FP}$, then $\sharp\text{CERTAINTY}(q)$ is in $\text{FP}$ by Lemma 2 and if $\sharp\text{CERTAINTY}(\text{enc}_S^*(q_m))$ is $\text{♯P}$-hard, then $\sharp\text{CERTAINTY}(q)$ is $\text{♯P}$-hard by Lemma 1. Consequently, $\sharp\text{CERTAINTY}(q)$ is in $\text{FP}$ or $\text{♯P}$-hard, and it is is decidable which of the two cases applies.

References

[MW13] Dany Maslowski and Jef Wijsen. A dichotomy in the complexity of counting database repairs. J. Comput. Syst. Sci., 79(6):958–983, 2013.

[MW14] Dany Maslowski and Jef Wijsen. Counting database repairs that satisfy conjunctive queries with self-joins. In Nicole Schweikardt, Vassilis Christophides, and Vincent Leroy, editors, Proc. 17th International Conference on Database Theory (ICDT), Athens, Greece, March 24-28, 2014., pages 155–164. OpenProceedings.org, 2014.