Sensitivity of Counting Queries

Myrto Arapinis, Diego Figueira, Marco Gaboardi

Sensitivity of Counting Queries

Myrto Arapinis[1], Diego Figueira[2], and Marco Gaboardi[3]

1 **University of Edinburgh, UK**
2 **CNRS, LaBRI, France**
3 **University at Buffalo, USA**

Abstract

In the context of statistical databases, the release of accurate statistical information about the collected data often puts at risk the privacy of the individual contributors. The goal of differential privacy is to maximise the utility of a query while protecting the individual records in the database. A natural way to achieve differential privacy is to add statistical noise to the result of the query. In this context, a mechanism for releasing statistical information is thus a trade-off between utility and privacy. In order to balance these two "conflicting" requirements, privacy preserving mechanisms calibrate the added noise to the so-called *sensitivity* of the query, and thus a precise estimate of the sensitivity of the query is necessary to determine the amplitude of the noise to be added.

In this paper, we initiate a systematic study of sensitivity of counting queries over relational databases. We first observe that the sensitivity of a Relational Algebra query with counting is not computable in general, and that while the sensitivity of Conjunctive Queries with counting is computable, it becomes unbounded as soon as the query includes a join. We then consider restricted classes of databases (databases with constraints), and study the problem of computing the sensitivity of a query given such constraints. We are able to establish bounds on the sensitivity of counting conjunctive queries over constrained databases. The kind of constraints studied here are: functional dependencies and cardinality dependencies. The latter is a natural generalisation of functional dependencies that allows us to provide tight bounds on the sensitivity of counting conjunctive queries.

1 Introduction

With the emergence of new systems and services such as eHealth, electronic tickets (*e.g.*, London Oyster card), mobile phones, or social networks, important amounts of information concerning our everyday activities are collected in various databases. Statistical analysis of such datasets could be very useful for improving services, or enabling research and market studies for example. But at the same time, the collection and storage of all this data puts at risk our individual privacy. A solution to address this problem is not to release the *exact result* of any query on a sensitive dataset, but rather to perturb the released results by adding some noise. Differential privacy [3, 10] precisely characterises the level of privacy provided by such randomized mechanisms. It offers a worst-case statistical guarantee on the increase in harm that an individual can be exposed to, if deciding to contribute her data to the dataset.

The concept of differential privacy is rooted in the notion of *neighboring databases*, that is, databases that differ in the presence or not of the information regarding one participant. More precisely, a mechanism $\mathcal{M}$ is ε-differentially private, for $\varepsilon \geq 0$ if for any two neighboring databases D and D' and for any subset $S \subseteq \mathcal{R}$ of possible outputs we have:

$$\Pr[\mathcal{M}(D) \in S] \quad \leq \quad \mathbf{e}^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S].$$

That is, the probability that $\mathcal{M}$ releases an element of S on D is almost the same as the probability that $\mathcal{M}$ releases an element of S on D'. In the definition of differential privacy

the parameter ε plays a central role. It gives the concrete bound on the increase in *harm* that an individual I can be exposed to, by contributing her data to the database.

Several mechanisms have been proposed to turn a deterministic query into a differentially private one, like the Laplace mechanism, the exponential mechanism, the Gaussian mechanism, etc. An extended introduction to these and other mechanisms (and more generally to differential privacy) is the recent monograph by Dwork and Roth [11]. In order to provide a good balance between privacy and utility, such ε-differential private mechanisms calibrate the added noise to the so-called *sensitivity* of the query. The sensitivity of a query Q captures the influence that an individual's data can have on the output of the query. More precisely, let us denote by $D \sim D'$ the fact that two databases D and D' are *neighbors*. The sensitivity of a numeric query Q is then

$$\max_{D \sim D'} |Q(D) - Q(D')|.$$

This measure is generally referred to as the *global sensitivity* of the query to distinguish it from other notions of local or smooth sensitivity [25].

To avoid adding too much noise and thus sacrificing too much utility to achieve the intended level of differential privacy, the sensitivity of the query needs to be computed as accurately as possible. However, this problem is undecidable in general as we shall see. In this paper we propose algorithms for computing upper bounds of the sensitivity of queries. Our results hold in a rather general setting: we consider counting conjunctive queries over multi-table databases. Further, our results are not tied to any particular neighboring relation, but hold for any relation of *bounded order*. This work is a first step towards understanding the class of queries and neighboring relations that are amenable to differential privacy.

Relational databases. Most of the works on differential privacy assume the simplified situation where the database is a monolithic table [11]. However, real life databases consist of not one, but many tables containing the information scattered. Of course, one could build a unique table from all these tables, by simply producing the cartesian product of all the tables in the database. Nevertheless, this immediately raises two problems. First, materialising the cartesian product of many—possibly big—tables is impractical, and often plain impossible due to space and time requirements. Second, the notion of neighboring databases now becomes *unbounded* which makes queries have unbounded sensitivity, and thus not amenable to differential privacy mechanisms. For example, given two tables (T_1, T_2) and a neighbor $T_1' = T_1 \setminus \{\bar{t}\}$ of T_1 for some record $\bar{t} \in T_1$, we have that, whereas (T_1', T_2) is the neighbor of (T_1, T_2) resulting from removing *one* record, $T_1 \times T_2$ differs from $T_1' \times T_2$ in a number of records equal to $|T_2|$. This in general makes it impossible for non-trivial queries to have bounded sensitivity, unless further restrictions on the databases are assumed.

Neighboring relation. Most works on differential privacy define neighboring databases as those that that differ in exactly one record. This corresponds to assuming that each individual contributes at most one record in the database. However, as pointed out in [29] this assumption does not hold for many applications such as social networks or tabular data. So the definition of neighboring databases needs to be tailored to the application at hand with privacy in mind. Indeed, neighboring databases should, strictly speaking, differ in the complete set of information pertaining to one individual, which could mean more than one record. Alternative definitions of neighboring have been proposed [29, 8]. In particular, our results are not tied to any particular definition of the neighboring relation.

SQL. SQL is arguably the prevalent query language for relational databases. It is equivalent to first order logic (FO) over relational structures and to Relational Algebra (RA). Here, we focus on SQL with aggregation, and study the static analysis problem of computing

the sensitivity of SQL queries. As a first step in the larger programme of studying aggregate queries, we study the *counting* operator. We concentrate our investigation on one of the most prominent fragments of SQL, namely the Conjunctive Queries, corresponding to positive "select-from-where" queries [1].

Contributions. We first establish, in Section 2, that finding the sensitivity of a SQL query with counting is not computable in general. In the remaining sections we restrict our study to counting Conjunctive Queries. Section 3 shows that the sensitivity for this fragment is computable, although the characterisation shows that sensitivity becomes unbounded as soon as we have a 'non-trivial' join.

Now, in most scenarios the class of databases of interest for the application at hand are restricted (or constrained), and oftentimes the sensitivity of a query Q restricted to a constrained class of databases can become bounded. Following this idea, we then study the problem of computing global sensitivity restricted to databases from a constrained class. In Section 4, we focus on *Functional Dependencies* (FD), that allow constraining databases by rules of the form "in the table T, the i-th column determines the j-th column", in other words, "there are no 2 rows of T with the same datum in the i-th column but distinct data in their j-th columns". Further, in Section 5 we study *Cardinality Dependencies*, which are a generalisation of FDs, with rules of the form "there are no more than k rows of T with the same datum in the i-th column but pairwise distinct data in their j-th columns". Finally, Section 6 concludes and discusses future work. Due to space limitation, all proofs are contained in the appendix.

Related work. Several works have studied methods for computing the sensitivity of a given query or program. The work most related to ours is the one of Palamidessi and Stronati [26]. They study the problem of computing the sensitivity of queries in relational algebra. Their approach is based on the use of constraints on attributes: every attribute comes with a bounded range, e.g. $0 \leq \text{age} \leq 100$. They are able to provide tight bounds on the sensitivity of the query Q. This approach can be applied to general SQL queries but it has the drawback that it requires to constrain the ranges for all the attributes. In this paper, instead, we focus on counting queries and on more lax semantic restrictions, namely functional dependencies and cardinality dependencies.

Pierce and Reed [27] and Gaboardi et al. [14] use relational algebra operations with a fixed, predetermined sensitivity, and a linear type system to track the use of the data in programs. This combination permits to have sensitivity analyses that extend, beyond SQL, to a full functional programming language. Their approach can provide "bad" estimates on the sensitivity of given queries due to the use of fixed sensitivity for relational operations. Our approach could provide a kernel query language providing more precise estimates that could be then combined with their type systems.

Chaudhuri et al. in [7] study automatic program analyses that provide bounds on the sensitivity of numerical imperative programs. Their approach is not directly related to specific query languages but our work could, in principle, be combined with their techniques to design a general purpose programming language for differential privacy.

Several works have pointed out and studied the problem of providing a bound to the sensitivity of queries in disconnected structures. McSherry [22], in the setting of tabular data, considers a restricted form of join where the data of the two tables are grouped by their join keys, and then groups are joined using their group keys. The same solution has been used also in [27, 14]. A similar approach, with different restrictions, has also been used by Palamidessi and Stronati in [26]. This approach limits the situations where differential privacy can be used with a good utility. To overcome this problem, several approaches considered alternative

notions of sensitivity such as *local sensitivity* [25] or *emprirical sensitivity* [8].

2 Preliminaries

Let $\mathbb{N} = \{0, 1, 2, \dots\}$ and let $\underline{n} = \{1, \dots n\}$ for every $n \in \mathbb{N}$. We write $\bar{a}$ to denote a vector of elements, whose i-th element is denoted by $\bar{a}[i]$. We write A^* [resp. A^+, A^n] for the set of strings [resp. non-empty strings, length-n strings] over A, and ε for the empty string.

2.1 Relational structures

A **relational vocabulary** $\sigma = (\mathcal{K}, \mathcal{R})$ consists of a collection $\mathcal{K}$ of **constants** (usually denoted by $c_1, c_2, \dots$), and a collection $\mathcal{R}$ of **relation symbols**, each with a specified **arity**. By σ_n we denote a vocabulary $\sigma_n = (\mathcal{K}, \mathcal{R})$ where $\mathcal{K} = \{c_1, \dots, c_n\}$. For a relation R we write $arity(R) \in \mathbb{N}$ to denote its arity; and we sometimes write $R^{(r)}$ to specify that R has arity r. A **σ-structure** $\mathbb{A}$ consists of a universe A containing $\mathcal{K}$, or **domain**, and an **interpretation** which associates to each relation symbol $R \in \mathcal{R}$, a relation $R^{\mathbb{A}} \subseteq A^{arity(R)}$, and for each constant $c \in \mathcal{K}$, $c^{\mathbb{A}} = c$. An **isolated element** of $\mathbb{A}$ is an element $a \in A$ which does not appear in any interpretation. Let STR be the set of all finite structures (we write $STR[\sigma]$ to make explicit the vocabulary). We use $\mathbb{A}, \mathbb{B}, \mathbb{C}, \mathbb{A}', \mathbb{B}', \dots$ to denote relational structures from STR, and $A, B, C, A', B', \dots$ to denote their respective domains. In the context of a signature $(\mathcal{K}, \mathcal{R})$ we will refer to a relation $R \in \sigma$ and an index $i \in \underline{arity(R)}$ with "$R[i]$".

▶ **Example 1.** As our running example, we will consider a database of patients, doctors and hospitals, with tables

- *Hos(id, loc)*, containing the hospitals with its location,
- *Pat(id, sex, hos)*, listing the patients with an identifier, gender and the hospital where they are being treated,
- *Doc(id, specialty, hos)*, listing the doctors with their identifier, their specialty and the hospital where they practice,
- *PatDoc(pat, doc)*, containing each patient and its current attending doctor.

Such a database can be described over the vocabulary $\sigma = (\mathcal{K}, \mathcal{R})$ containing relations $\mathcal{R} = \{Hos^{(2)}, Pat^{(3)}, Doc^{(3)}, PatDoc^{(2)}\}$ and some constants such as $\mathcal{K} = \{c_{\mathrm{F}}, c_{\mathrm{O}}\}$.

A **graph** is a structure $G = (V, E)$, where E is a binary relation that is symmetric and irreflexive. Thus, our graphs are undirected, loopless, and without parallel edges. The **Gaifman graph** of a σ-structure $\mathbb{A}$, denoted by $\mathcal{G}(\mathbb{A})$, is the (undirected) graph whose set of nodes is the universe of $\mathbb{A}$, and whose set of edges consists of all pairs (a, a') of distinct elements of A such that a and a' appear together in some tuple of a relation in $\mathbb{A}$. Recall that the **distance** between two vertices u, v of a graph is the length of the shortest path from u to v. We define the distance between two elements a, b of a structure $\mathbb{A}$ as their distance in $\mathcal{G}(\mathbb{A})$, which we denote by $dist_{\mathbb{A}}(a, b)$. We write $A \sqcup B$ for the disjoint union of A and B.

A **homomorphism** from a $(\mathcal{K}, \mathcal{R})$-structure $\mathbb{A}$ and a $(\mathcal{K}', \mathcal{R}')$-structure $\mathbb{B}$ of so that $\mathcal{K} \subseteq \mathcal{K}'$ and $\mathcal{R} \subseteq \mathcal{R}'$ is a mapping $h : A \to B$ so that for each relation symbol $R \in \mathcal{R}$, if $(a_1, \dots, a_r) \in R^{\mathbb{A}}$, then $(h(a_1), \dots, h(a_r)) \in R^{\mathbb{B}}$, and for every constant $c \in \mathcal{K}$, $h(c) = c$. We will sometimes write $h(a_1, \dots, a_r)$ as short for $(h(a_1), \dots, h(a_r))$. We write $\mathbb{A} \to \mathbb{B}$ to denote that there is a homomorphism from $\mathbb{A}$ to $\mathbb{B}$, and we write $h : \mathbb{A} \to \mathbb{B}$ to denote that h is a homomorphism from $\mathbb{A}$ to $\mathbb{B}$. If $\mathbb{A} \to \mathbb{B}$ and $\mathbb{B} \to \mathbb{A}$ we say that $\mathbb{A}$ and $\mathbb{B}$ are **hom-equivalent**. We use $\cong$ for the isomorphism relation. Given a σ-structure $\mathbb{A}$ and a set $B \subseteq A$ there is (up to isomorphism) a unique structure $\mathbb{A}'$ so that

- it is hom-equivalent to $\mathbb{A}$, that is, there are $h : \mathbb{A} \to \mathbb{A}'$ and $h' : \mathbb{A}' \to \mathbb{A}$,

- $h(a) = h'(a) = a$ for all $a \in B$,
- it has the minimal number of elements.

Such a structure $\mathbb{A}'$ is called the **core preserving** B (or simply **core** if $B = \emptyset$). We write $core(\mathbb{A}, B)$ [resp. $core(\mathbb{A})$] to denote the core of $\mathbb{A}$ preserving B [resp. the core of $\mathbb{A}$].

2.2 Logic

Let $\mathcal{V}$ be a collection of first-order variables equipped with a linear order $<$. Let σ be a relational vocabulary. A **term** is either a first order variable $x \in \mathcal{V}$ or a constant from σ. The **atomic formulas** of σ are those of the form $R(t_1, \ldots, t_n)$, where $R \in \sigma$ is a relation symbol of arity r, and $t_1, \ldots, t_r$ are terms. Formulas of the form $t = t'$ are also atomic formulas, and we refer to them as **equalities**. The collection of **first-order formulas** (FO formulas) is obtained by closing the atomic formulas under negation, conjunction, disjunction, universal and existential first-order quantification. The semantics of first-order logic is standard. The set of variables of φ is denoted by $var(\varphi)$, and the set of free variables by $free(\varphi)$. We often write $\varphi(x_1, \ldots, x_n)$ where $\{x_1, \ldots, x_n\} = free(\varphi)$ and $x_1 < \cdots < x_n$, to stress the free variables. If $\mathbb{A}$ is a σ-structure and $\varphi(\bar{x})$ is a first-order formula, we use the notation $\mathbb{A} \models \varphi[\bar{a}]$ to denote the fact that φ is true in $\mathbb{A}$ when its free variables $\bar{x}$ are interpreted by the tuple of elements $\bar{a}$. When φ contains no free variables, we say that it is a **sentence**, and in this case we simply write $\mathbb{A} \models \varphi$. For any formula $\varphi(x_1, \ldots, x_n)$ and structure $\mathbb{A}$, we write $\varphi(\mathbb{A})$ to denote $\{(a_1, \ldots, a_n) \in A^n \mid \mathbb{A} \models \varphi[a_1, \ldots, a_n]\}$. We use '()' to denote the 0-ary tuple of elements. Hence, if φ has no free variables we interpret $\varphi(\mathbb{A})$ as $\{()\}$ if $\mathbb{A} \models \varphi$ or $\emptyset$ otherwise. Note that, in this case, $|\varphi(\mathbb{A})| = 1$ iff $\mathbb{A} \models \varphi$. We use $\equiv$ for the logical equivalence relation and $\equiv_{\mathcal{C}}$ for the equivalence relation restricted to a class of structures $\mathcal{C}$.

Given a class of FO formulas $\mathcal{L}$, by $\mathcal{L}^{\#}$ we denote the class of **counting queries** $\{\#\varphi \mid \varphi \in \mathcal{L}\}$. The evaluation of $\#\varphi$ in $\mathbb{A}$, denoted $\#\varphi(\mathbb{A})$, is defined as $|\varphi(\mathbb{A})|$, that is, as the number of distinct tuples making φ true in $\mathbb{A}$.

▶ **Example 2.** Continuing our running example, we consider the query that counts the number of oncology doctors that are treating female patients in the same hospital as they practice:

```
SELECT count distinct Doc.id
FROM Pat, Doc, PatDoc
WHERE Doc.specialty = 'O' and
      Pat.sex = 'F' and
      Pat.hos = Doc.hos and
      PatDoc.pat = Pat.id and
      PatDoc.doc = Doc.id
```

This can be equivalently expressed with the formula $\#\varphi$, where

$$\varphi(x_{doc}) = \exists x_{pat}, x_{hos} \,.\, Doc(x_{doc}, c_{\mathrm{O}}, x_{hos}) \wedge Pat(x_{pat}, c_{\mathrm{F}}, x_{hos}) \wedge PatDoc(x_{pat}, x_{doc})$$

2.3 Global sensitivity

In its standard formulation, Differential Privacy requires the privacy bound to be valid for every pair of structures that differ in one record. However, it is possible that an individual contributes more than a single record to the database. Further it may be that the database contains tables with public information. For this reason we do not set for our study a particular neighboring relation. Our results hold for any **neighboring relation** $\mathcal{N} \subseteq STR[\sigma] \times STR[\sigma]$.

Having said that, a specific neighboring relation, called **1-neighboring**, will be particularily useful for our proofs. Given two σ-structures $\mathbb{A}, \mathbb{B}$ with $\sigma = (\mathcal{K}, \mathcal{R})$, we say that $\mathbb{A}$

is a **substructure** of $\mathbb{B}$ (noted $\mathbb{A} \subseteq \mathbb{B}$) if $A \subseteq B$, and $R^{\mathbb{A}} \subseteq R^{\mathbb{B}}$ for all $R \in \sigma$. We write $\mathbb{A} \prec \mathbb{A}'$ if $\mathbb{A} \subsetneq \mathbb{A}'$ and there is no $\mathbb{B}$ so that $\mathbb{A} \subsetneq \mathbb{B} \subsetneq \mathbb{A}'$. We say that $\mathbb{A}, \mathbb{B}$ are 1-neighboring structures, noted $\mathbb{A} \sim_1 \mathbb{B}$, if $\mathbb{A} \prec \mathbb{B}$ or $\mathbb{B} \prec \mathbb{A}$. In other words, $\mathbb{A} \sim_1 \mathbb{B}$ if $\mathbb{A}$ can be obtained from $\mathbb{B}$ (and B from A) by removing/adding a tuple or an isolated noted.

We say that the neighboring relation $\mathcal{N}$ is **of order** $k \in \mathbb{N}$, if any two neighboring relational structures differ in at most k elements. More formally, $\mathcal{N}$ is of order k if for any $(\mathbb{A}, \mathbb{B}) \in \mathcal{N}$, there exist $\mathbb{A}_0, \ldots, \mathbb{A}_\ell$ such that $\ell \leq k$, $\mathbb{A} = \mathbb{A}_0$, $\mathbb{B} = \mathbb{A}_\ell$ and $\mathbb{A}_{i-1} \sim_1 \mathbb{A}_i$ for all $i \in \underline{\ell}$. We say that the neighboring relation is unbounded if no such k exists.

The **global sensitivity** of a function $f : STR \to \mathbb{N}$ over a class of models $\mathcal{C} \subseteq STR$ with respect to a neighboring relation $\mathcal{N} \subseteq \mathcal{C} \times \mathcal{C}$ is:

$$\mathrm{GS}^{\mathcal{N}}_{\mathcal{C}}(f) \stackrel{def}{=} \max_{(\mathbb{A},\mathbb{A}') \in \mathcal{N}} |f(\mathbb{A}) - f(\mathbb{A}')|.$$

▶ **Example 3.** Suppose now that we want to find out the number of oncological patients in the state of New York with the query

$$\varphi(x_{pat}) = \exists x_{pat}, x_{hos}, x_{doc}, x_{sex} \ . \\ Doc(x_{doc}, c_{\mathrm{O}}, x_{hos}) \wedge Pat(x_{pat}, x_{sex}, x_{hos}) \wedge PatDoc(x_{pat}, x_{doc}) \wedge Hos(x_{hos}, x_{loc})$$

It is not hard to see that this query has unbounded global sensitivity when all relations are considered sensitive, and thus all databases that differ in any one element are neighbors. Indeed changing the location of a hospital from Indiana to New-York can increase the number of ontological patients in the state of New York by any number.

▶ Observation 1. For any neighboring relation $\mathcal{N}$ of order k and any class of databases $\mathcal{C}$, the global sensitivity of a query Q is bounded with respect to $\mathcal{N}$ over $\mathcal{C}$ iff it is bounded with respect to $\sim_1$ over $\mathcal{C}$. Further, the global sensitivity with respect to $\mathcal{N}$ and relative to the class $\mathcal{C}$ is bounded by $k \cdot \mathrm{GS}^{\sim_1}_{\mathcal{C}}(Q)$. So in the remaining of the paper we focus on 1-neighboring.

We will study the following problem, given a query language $\mathcal{L}$, and a class of relational structures $\mathcal{C}$

Problem:	GlobalSensitivity($\mathcal{L}$, $\mathcal{C}$)
Input:	$Q \in \mathcal{L}$
Output:	$\mathrm{GS}^{\sim_1}_{\mathcal{C}}(Q)$

Unfortunately, this problem is undecidable already for counting first-order logic (and therefore for counting Relational Algebra [1]).

▶ **Theorem 4.** GlobalSensitivity*($FO^{\#}$, STR) is non-computable.*

The fact that the global sensitivity problem for FO is undecidable is not really surprising since most static analysis problems for FO on unrestricted structures are undecidable. This is why in the next sections we will focus on Conjunctive Queries.

3 Conjunctive queries

One of the most studied fragments of FO in relation to database queries is the fragment of *Conjunctive Queries* (CQ). We now, and for the rest of the paper, restrict our study to counting conjunctive queries, and show that sensitivity for this fragment is computable.

The class of Conjunctive Queries (also known as Primitive Positive Logic, or Existential Positive FO) is the fragment of FO corresponding to positive '*select-project-join*' queries of the Relational Algebra or to positive '*select-from-where*' queries of SQL, where by 'positive' we mean that there are no inequalities in the *select* [resp. *where*] conditions (we refer the reader to [1, §4] for more details). These are formulae of the form

$$\varphi(x_1, \ldots, x_n) = \exists y_1, \ldots, y_m\, \theta, \tag{†}$$

where θ is a conjunction of atomic formulae. Since we deal with constants, and, in future sections, with constrained databases, a conjunctive query can also be *false* (noted $\bot$). However, all the results that we show will assume that the input formula is not equivalent to $\bot$ (*i.e.*, that it is satisfiable, which can be checked in polynomial time)—for the particular case where formulae are unsatisfiable all the results are trivial, and this will avoid lengthy statements. For simplicity, and without any loss of generality, we assume that the formulae do not contain equalities.[1]

Every conjunctive query of the form (†) over a relational vocabulary σ_k gives rise to a **canonical structure** (sometimes called *tableau*) $\mathbb{C}_\varphi$ with $n + m + k$ elements, where the elements of $\mathbb{C}_\varphi$ are the variables $x_1, \ldots, x_n, y_1, \ldots, y_m$ plus the constants $c_1, \ldots, c_k$, the relations of $\mathbb{C}_\varphi$ consist of the tuples of terms in the conjuncts of θ. Given a CQ φ, we write $\mathbb{C}_\varphi$ for the canonical structure of φ, and C_φ for its domain (*i.e.*, the variables $x_1, \ldots, x_n, y_1, \ldots, y_m$ and constants $c_1, \ldots, c_k$). We also define $\mathbb{C}_\varphi^-$ as the result of removing all isolated constants from $\mathbb{C}_\varphi$ (note that $\mathbb{C}_\varphi^-$ may not necessarily be a structure over the same vocabulary of φ due to the absence of some constants). Likewise, any σ_k-structure $\mathbb{A}$ with domain $A = \{x_1, \ldots, x_n\} \cup \{c_1, \ldots, c_k\}$ gives rise to a canonical CQ $\varphi(x_1, \ldots, x_n)$ where $\mathit{var}(\varphi) = \mathit{free}(\varphi) = \{x_1, \ldots, x_n\}$, and φ has a conjunct $R(\bar{t})$ iff $\bar{t} \in R^{\mathbb{A}}$. Note that for every σ_k-structure $\mathbb{A}$ there is $\mathbb{A}' \cong \mathbb{A}$ and φ so that φ is the canonical query of $\mathbb{A}'$.

A CQ φ is **acyclic** if $\mathcal{G}(\mathbb{C}_\varphi)$ is acyclic. We say that a CQ φ is **connected** if $\mathcal{G}(\mathbb{C}_\varphi^-)$ is connected, otherwise it is **disconnected**. Note that every disconnected CQ φ so that $\mathcal{G}(\mathbb{C}_\varphi)$ has n connected components can be equivalently written in the form $\varphi = \bigwedge_{i \in \underline{n}} \psi_i(\bar{x}_i)$ so that $\psi_i(\bar{x}_i)$ is a connected CQ for every i, and for all $i \neq j$, $\bar{x}_i$ and $\bar{x}_j$ have no variables in common. We say that ψ_i is a **connected conjunct** of φ, and we say that ψ_i is a **sentential connected conjunct** if it is a sentence (*i.e.*, $\bar{x}_i = ()$). Given $\varphi = \bigwedge_{i \in \underline{n}} \psi_i(\bar{x}_i)$ a disconnected CQ with each ψ_i being a connected conjunct, we further define $\overline{\varphi}^j$ as the conjunction of all the ψ_s's but ψ_j.

▶ **Example 5** (Cont. from Ex. 2)**.** The canonical σ-structure $\mathbb{C}_\varphi$ has universe $\{x_{pat}, x_{hos}, x_{doc}, c_{\mathrm{O}}, c_{\mathrm{F}}\}$ and relations (shown in Figure 1):

$$\mathit{Doc}^{\mathbb{C}_\varphi} = \{(x_{doc}, c_{\mathrm{O}}, x_{hos})\}, \quad \mathit{Pat}^{\mathbb{C}_\varphi} = \{(x_{pat}, c_{\mathrm{F}}, x_{hos})\}, \quad \mathit{PatDoc}^{\mathbb{C}_\varphi} = \{(x_{pat}, x_{doc})\}.$$

Core of CQ's. For a CQ query $\varphi(\bar{x}) = \exists \bar{y}.\theta$ over σ_k we define $\mathit{core}(\varphi)$ as the CQ query $\varphi'(\bar{x}) = \exists \bar{y}'.\theta'$ where θ' is the canonical query of $\mathit{core}(\mathbb{C}_\varphi, \bar{x})$ and $\bar{y}'$ is the set of all non-constant elements of $\mathit{core}(\mathbb{C}_\varphi, \bar{x})$ that are not in $\bar{x}$. Note that $\mathbb{C}_{\mathit{core}(\varphi)} \cong \mathit{core}(\mathbb{C}_\varphi, \bar{x})$. We say that $\varphi(\bar{x})$ is a **core-CQ** if $\mathbb{C}_{\mathit{core}(\varphi)} \cong \mathit{core}(\mathbb{C}_\varphi)$, and we write $\mathrm{CQ}_{\mathit{core}}$ for the class of all core-CQ's. We define $\mathit{core}(\#\varphi)$ as $\#\mathit{core}(\varphi)$ for every CQ φ.

[1] Observe that for every CQ using equalities, there is an equivalent CQ that does not use equalities, computable in polynomial time.

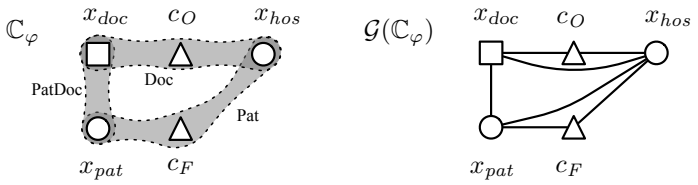

Figure 1 Depiction of the canonical structure of φ as defined in Example 2 as well as its Gaifman graph. Square vertices denote free variables and triangle vertices denote constants.

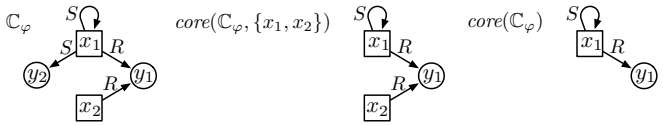

Figure 2 Core of CQ's.

▶ **Example 6.** Given $\varphi(x_1, x_2) = \exists y_1, y_2.S(x_1, x_1) \wedge S(x_1, y_2) \wedge R(x_1, y_1) \wedge R(x_2, y_1)$, whose canonical structure is depicted in Figure 2, we have that $core(\varphi) \equiv \exists y_1.S(x_1, x_1) \wedge R(x_1, y_1) \wedge R(x_2, y_1)$, and that φ is not a core-CQ since $core(\mathbb{C}_\varphi, \{x_1, x_2\})$ is not isomorphic to $core(\mathbb{C}_\varphi)$, as shown in the figure below.

Given a connected CQ φ, let us define

$$\Delta_{STR}(\#\varphi) = \begin{cases} \infty & \text{if } \exists x \in free(\varphi).\ \exists R \in \mathcal{R}.\ \exists \bar{a} \in R^{core(\varphi)}.\ x \notin \bar{a} \\ 1 & \text{otherwise} \end{cases}$$

▶ Proposition 1. For every connected CQ$^\#$ Q, we have $\mathrm{GS}^{\sim_1}_{STR}(Q) = \Delta_{STR}(Q)$.

▶ **Example 7** (Cont. from Ex. 5)**.** Note that we have $\Delta_{STR}(\#\varphi) = \infty$ since $core(\varphi) = \varphi$ and x_{doc} is not in the tuple (x_{pat}, c_F, x_{hos}) of the relation $Pat^{\mathbb{C}_\varphi}$, and thus that $\mathrm{GS}^{\sim_1}_{STR}(Q) = \infty$.

We extend the definition of Δ_{STR} to disconnected CQ$^\#$ as follows. For any $\varphi = \bigwedge_{i \in \underline{n}} \varphi_i$ disconnected CQ so that each φ_i is a connected conjunct, we define

$$\Delta_{STR}(\#\varphi) = \begin{cases} \Delta_{STR}(\#\varphi_k) & \text{if } \exists k \in \underline{n}.\ free(\varphi) = free(\varphi_k) \ \wedge\ \mathbb{C}_{\overline{\varphi}^k} \to \mathbb{C}_{\varphi_k} \\ \infty & \text{otherwise} \end{cases}$$

▶ **Theorem 8.** *For every CQ$^\#$ Q, we have $GS^{\sim_1}_{STR}(Q) = \Delta_{STR}(Q)$.*

The above characterization shows that, even when we deal with *connected* CQ's (arguably the most common), we obtain unbounded sensitivity very easily. Indeed, as soon as one has a 'join' with a free variable which is not the joining attribute, such as $\#\varphi(x) = \#\,\exists y, z\,.\,R(x, y) \wedge S(y, z)$ the global sensitivity is unbounded. Although this means that for every $N \in \mathbb{N}$ there are structures $\mathbb{A} \sim_1 \mathbb{A}'$ so that $\#\varphi(\mathbb{A}) - \#\varphi(\mathbb{A}') > N$, it may be that $\mathbb{A}, \mathbb{A}'$ do not correspond to databases that could arise in the domain of application at hand. However, when restricting the set of considered structures to ones satisfying some constraints, it may well be that the sensitivity becomes bounded. The next two sections will focus on evaluating sensitivity of queries over constrained structures.

4 Functional Dependencies

In this section we show bounds for the sensitivity of queries in the presence of what are called *functional dependencies*.

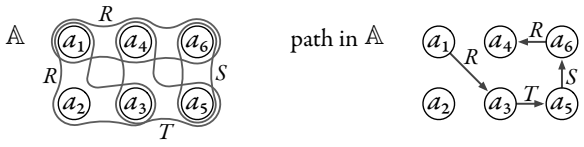

Figure 3 A path in a structure.

▶ **Example 9** (Cont. from Ex. 2)**.** Note that, the global sensitivity of $\#\varphi$ is unbounded. Indeed, this is a consequence of the possibility of having patients with unbounded number of attending doctors and doctors working in any number of hospitals. However, this does not correspond to databases that could occur in practice, since patients have normally one attending doctor and doctors work in at most one hospital. This is why the use of database constraints becomes useful, to restrict the collection of databases we are interested in, and thus to improve the bounds of the sensitivity of queries.

We write $R[i \to j]$ to denote a **functional dependency of a relation** R **of arity** n **between components** $i \in \underline{n}$ **and** $j \in \underline{n}$. A structure $\mathbb{A}$ satisfies a functional dependency (henceforth "FD") $R[i \to j]$ if $\max_{a \in A}(|\{\bar{b}[j] \mid \bar{b} \in R^{\mathbb{A}}, \bar{b}[i] = a\}|) \leq 1$. We use the symbol Σ to denote a set of FDs, and we write $\#_\Sigma R[i \to j]$ to denote 1 if $R[i \to j] \in \Sigma$, or ∞ otherwise. We write $\mathcal{C}_\Sigma$ for the class of all relational structures satisfying all FDs in Σ.

Given a CQ query φ and a set of FDs Σ we define the **Σ-chase** [21, 2] of φ, noted $chase_\Sigma(\varphi)$, as the closure of the application of the following rule:

- For every $R[i \to j] \in \Sigma$ and every pair of conjuncts $R(\bar{t})$ and $R(\bar{s})$ of φ so that $\bar{t}[i] = \bar{s}[i]$ and $\bar{t}[j] \neq \bar{s}[j]$,
  - if $\bar{s}[j]$ is a variable, replace every occurrence of $\bar{s}[j]$ with $\bar{t}[j]$;
  - if $\bar{s}[j]$ and $\bar{t}[j]$ are constants, output $\bot$.

It can be seen that the application of these rules is terminating and Church-Rosser confluent, up to renaming of variables [1].

The following result shows that, as soon as we have a disconnected query, the sensitivity is likely to be unbounded.

▶ Proposition 2. For every disconnected CQ query φ containing a conjunct without constants and at least one free variable, and for every set Σ of FD's, we have $\mathrm{GS}^{\sim 1}_{\mathcal{C}_\Sigma}(\#\varphi) = \infty$.

Paths. A **path** of a $(\mathcal{K}, \mathcal{R})$-structure $\mathbb{A}$ between an element $a \in A$ and $b \in A$, is a string

$$p = (R_1, i_1, a_1, j_1, b_1) \cdots (R_n, i_n, a_n, j_n, b_n) \in (\mathcal{R} \times \mathbb{N} \times A \times \mathbb{N} \times A)^* \qquad (\star)$$

so that either $p = \varepsilon$ and $a = b$ (*i.e.*, the empty path); or $a_1 = a$, $b_n = b$, $a_i = b_{i-1}$ for all $1 < i \leq n$, and for every $\ell \in \underline{n}$ we have $i_\ell, j_\ell \in \underline{arity(R_i)}$ and there is $\bar{a} \in R^{\mathbb{A}}_\ell$ so that $\bar{a}[i_\ell] = a_\ell$ and $\bar{a}[j_\ell] = b_\ell$. A path of the form $(\star)$ is **simple** if $a_i \neq b_i \neq b_j$ for all $1 \leq i < j \leq n$. Note that in particular the empty path ε is simple. We write $p : A_1 \rightsquigarrow_{\mathbb{A}} A_2$ to denote that p is a simple path of $\mathbb{A}$ from an element of $A_1 \subseteq A$ to an element of $A_2 \subseteq A$. We write $a \rightsquigarrow_{\mathbb{A}} b$, $A_1 \rightsquigarrow_{\mathbb{A}} b$, $a \rightsquigarrow_{\mathbb{A}} A_2$ as short for $\{a\} \rightsquigarrow_{\mathbb{A}} \{b\}$, $A_1 \rightsquigarrow_{\mathbb{A}} \{b\}$, $\{a\} \rightsquigarrow_{\mathbb{A}} A_2$ respectively.

▶ **Example 10.** For a structure $\mathbb{A}$ with relations $R^{\mathbb{A}} = \{(a_1, a_2, a_3), (a_1, a_4, a_6)\}$, $S^{\mathbb{A}} = \{(a_5, a_6)\}$, and $T^{\mathbb{A}} = \{(a_3, a_5, a_4)\}$, we have that $p : a_1 \rightsquigarrow_{\mathbb{A}} a_4$ for $p = (R, 1, a_1, 3, a_3)$ $(T, 1, a_3, 2, a_5)$ $(S, 1, a_5, 2, a_6)$ $(R, 3, a_6, 2, a_4)$, as depicted in Figure 3

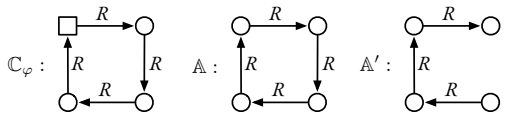

Figure 4 Structures of Example 12. Square vertices denote free variables.

Given a vocabulary $\sigma = (\mathcal{K}, \mathcal{R})$ and a path p of the form $(\star)$, let $m \in \underline{n}$ be the greatest index m so that $b_m \in \mathcal{K}$, or 0 otherwise. We define the **cardinality of path** p as

$$\#_\Sigma(p) \stackrel{def}{=} \prod_{m<\ell\leq n} \#_\Sigma R_\ell[i_\ell \to j_\ell] \tag{1}$$

where as usual the product of the empty sequence is 1, and ∞ is absorbing wrt the product ($\infty \cdot N = N \cdot \infty = \infty$). Note that $\#_\Sigma(\varepsilon) = 1$. The intuition is that $\#_\Sigma(p)$ gives a bound on how many different elements b can be reached from a through p on any structure $\mathbb{A} \in \mathcal{C}_\Sigma$ (*i.e.*, so that $p : a \rightsquigarrow_{\mathbb{A}} b$).

Let $Q = \#\psi(x_1, \ldots, x_n)$ be a connected CQ$^\#$ over a vocabulary $\sigma = (\mathcal{K}, \mathcal{R})$, and let $\varphi = \mathit{core}(\mathit{chase}_\Sigma(\psi))$. We define

$$\Delta^+_\Sigma(Q) \stackrel{def}{=} \max_{R\in\mathcal{R}} \sum_{\bar{a}\in R^{\mathbb{C}_\varphi}} \max_{i\in\underline{n}} \left(\min_{p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i} \#_\Sigma(p_i) \right)$$

$$\Delta^-_\Sigma(Q) \stackrel{def}{=} \max_{R\in\mathcal{R}} \max_{\bar{a}\in R^{\mathbb{C}_\varphi}} \max_{i\in\underline{n}} \left(\min_{p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i} \#_\Sigma(p_i) \right).$$

▶ Observation 2. Note that $\Delta^-_\Sigma(Q)$ is either 1 or ∞ and that $\Delta^-_\Sigma(Q) = \infty$ iff $\Delta^+_\Sigma(Q) = \infty$. Further, observe that $\Delta^+_\Sigma(Q) - \Delta^-_\Sigma(Q) \leq n_Q - 1$, where n_Q is the maximum number of elements in a relation of the canonical structure of $\mathit{core}(\mathit{chase}_\Sigma(\psi))$, assuming $Q = \#\psi$.

▶ **Theorem 11.** *Given a set Σ of functional dependencies and a connected CQ$^\#$ query Q, we have that $GS^{\sim^1}_{\mathcal{C}_\Sigma}(Q) \leq \Delta^+_\Sigma(Q)$. Further, if $Q \in \mathrm{CQ}^\#_{core}$, we have $GS^{\sim^1}_{\mathcal{C}_\Sigma}(Q) \geq \Delta^-_\Sigma(Q)$.*

▶ **Example 12.** Take for instance the CQ with one free variable of Figure 4. Observe that, for $\Sigma = \{R[1 \to 2],\ R[2 \to 1]\}$, we have that $\mathrm{GS}^{\sim^1}_{\mathcal{C}_\Sigma}(\#\varphi) \leq \Delta^+_\Sigma(\#\varphi) = 4$, which is tight since $\#\varphi(\mathbb{A}) = 4$, and $\#\varphi(\mathbb{A}') = 0$. Further, this example can be easily generalized, obtaining that for every $n \in \mathbb{N}$ there is a CQ Q so that $\mathrm{GS}^{\sim^1}_{\mathcal{C}_\Sigma}(Q) = n = \Delta^+_\Sigma(Q)$.

▶ **Example 13** (Cont. from Ex. 2)**.** As noted in Example 9, $\#\varphi$ has unbounded global sensitivity. However, if every patient has no more than one attending doctor, the sensitivity of $\#\varphi$ becomes bounded. Indeed, if $\Sigma = \{\mathit{PatDoc}[1 \to 2]\}$, then

$$\Delta^-_\Sigma(\#\varphi) \leq \mathrm{GS}^{\sim^1}_{\mathcal{C}_\Sigma}(\#\varphi) \leq \Delta^+_\Sigma(\#\varphi)$$

by Theorem 11—observe that $\varphi \in \mathrm{CQ}_{core}$ since it is unary, *cf.* Lemma 23. Since $\Delta^-_\Sigma(\#\varphi) = \Delta^+_\Sigma(\#\varphi) = 1$, it thus follows that $\mathrm{GS}^{\sim^1}_{\mathcal{C}_\Sigma}(\#\varphi) = 1$.

As we have shown, adding functional dependencies immediately improves the global sensitivity of queries. However, functional dependencies are often very restrictive, and it may not always be possible to impose such restrictions. This leads to a more general notion of dependencies, that we call *cardinality dependencies*. These dependencies bound the number of elements associated with component i of a relation R for each fixed element of a component j. This will be the object of study of our next section.

5 Cardinality Dependencies

While functionality constraints are a very natural restriction of databases, there are many scenarios in which, although we don't have an attribute i functionally determining an attribute j in a relation, we have a **cardinality dependency** nonetheless. This is a dependency of the form "there are at most n different attributes j sharing the same attribute i in the relation R"—functional dependencies being the special case when $n = 1$.

These dependencies arise naturally when modelling relations between entities (such as in ER modelling [17]). For example, the business rules underlying a company database may allow that an employee has more than one manager, but no more than 2. Another example is for bounded domain attributes: whereas the name of a person does not determine the gender, there cannot be more than two possibilities of gender for any given name. As we will see next, cardinality dependencies provide further means to give tighter bounds for the global sensitivity of CQ's.

▶ **Example 14** (Cont. from Ex. 13)**.** We already noticed that constraining each patient to have at most one attending doctor, brings the sensitivity of $\#\varphi$ down to 1. However, it may be that a patient can have more than one attending doctor, although it can't have an *unbounded* number of attending doctors. For example, a scenario in which a patient has *at most* 3 attending doctors.

More formally, we write $R[i \xrightarrow{k} j]$ to denote a **k-cardinality dependency of a relation R of arity n between components $i \in \underline{n}$ and $j \in \underline{n}$**. A structure $\mathbb{A}$ satisfies a cardinality dependency (henceforth "CD") $R[i \xrightarrow{k} j]$ if $\max_{a \in A}(|\{\bar{b}[j] \mid \bar{b} \in R^{\mathbb{A}}, \bar{b}[i] = a\}|) \leq k$. For the particular case where $k = 1$, note that $R[i \xrightarrow{k} j]$ is a *functional* dependency. We use the symbol Σ to denote a set of CD's, and we write $\#_\Sigma R[i \to j]$ to denote the minimum k so that $R[i \xrightarrow{k} j] \in \Sigma$, or ∞ otherwise. As before, we write $\mathcal{C}_\Sigma$ for the class of all relational structures satisfying all CDs in Σ. We define the cardinality of a path $\#_\Sigma(p)$ as in (1), where now Σ is a set of CD's, and in the definition $\#_\Sigma R[i \to j]$ is interpreted as defined above, over CD's.

Upper bound. Given a connected CQ$^\#$ query Q over a vocabulary $\sigma = (\mathcal{R}, \mathcal{K})$ so that $core(Q) = \#\varphi(x_1, \ldots, x_n)$, let us define

$$\Delta^+_\Sigma(Q) \stackrel{def}{=} \max_{R \in \mathcal{R}} \left(\sum_{\bar{a} \in R^{\mathbb{C}_\varphi}} \left(\min_{\substack{p_1,\ldots,p_n \text{ s.t.} \\ p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i \text{ for } i \in \underline{n}}} \left(\prod_i \#_\Sigma(p_i) \right) \right) \right).$$

▶ **Theorem 15.** *Given a set of cardinality dependencies Σ, for all connected CQ$^\#$ queries Q we have $GS^{\sim^1}_{\mathcal{C}_\Sigma}(Q) \leq \Delta^+_\Sigma(Q)$.*

▶ **Example 16** (Cont. from Ex. 14)**.** If every patient has at most 3 attending doctors, the sensitivity of $\#\varphi$ becomes bounded. Indeed, if $\Sigma = \{PatDoc[1 \xrightarrow{3} 2]\}$, then $\mathrm{GS}^{\mathcal{R}}_{\mathcal{C}_\Sigma}(\#\varphi) \leq \Delta^+_{\mathcal{R},\Sigma}(\#\varphi) = 3$ by Theorem 15.

Lower bound. Let us now define

$$\Delta^-_\Sigma(Q) \stackrel{def}{=} \max_{R \in \mathcal{R}, \bar{a} \in R^{\mathbb{C}_\varphi}} \left(\min_{\substack{p_1,\ldots,p_n \text{ s.t.} \\ p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i \text{ for } i \in \underline{n}}} \left(\prod_i \#_\Sigma(p_i) \right) \right)$$

We use the symbol π to denote permutations $\pi : \underline{n} \to \underline{n}$, and given $\bar{a} \in A^n$ we write $\bar{a}_\pi \in A^n$ to denote the vector whose i-th element is $\bar{a}[\pi(i)]$.

Let us say that a CQ$^{\#}$ query $Q = \#\varphi(\bar{x})$ has **no repeated joins** if for every binary relation $R \in \mathcal{R}$, any variable or constant appears in at most one record $\bar{a} \in R^{\mathbb{C}_\varphi}$.

▶ **Theorem 17.** *For all acyclic connected CQ$^{\#}$ queries Q with no repeated joins and with all relations of arity ≤ 2, we have that $\Delta^{-}_{\Sigma}(Q) \leq GS^{\sim_1}_{\mathcal{C}_\Sigma}(Q)$.*

6 Conclusion

We have given bounds for the global sensitivity of counting Conjunctive Queries under the functionality or cardinality constraints. These bounds can be used to turn those queries in differentially private ones by using mechanisms like the Laplacian or the Gaussian mechanisms without adding too much noise.

There are several interesting directions that we will pursue in future work. We will study other aggregation operations already present in SQL such as *average* or *sum*. We will also investigate sensitivity of queries with negation, where one can ask for example for the number of patients that are *not* treated by a given doctor. Further, we have focused here on global sensitivity but there are other notions of sensitivity that have been proposed. In particular, the so-called *local sensitivity* is studied in [25]. The local sensitivity is defined by quantifying not over all possible databases but only over the ones in the neighborhood of the particular database under analysis. The local sensitivity is often lower than the global sensitivity, but adding noise proportional to the local sensitivity does not ensure differential privacy. Nevertheless, adding the noise proportional to a smooth approximation of the local sensitivity permits to recover differential privacy.

References

1 S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
2 A. V. Aho, C. Beeri, and J. D. Ullman. The theory of joins in relational databases. *ACM TODS*, 4(3):297–314, 1979.
3 A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS* 2005.
4 T.-H. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. *ACM TISSEC*, 14(3):26, 2011.
5 A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In STOC, p. 77–90, 1977. ACM.
6 K Chatzikokolakis, M. Andrés, N. E. Bordenabe and C. Palamidessi. Broadening the Scope of Differential Privacy Using Metrics. In PETS 2013.
7 S. Chaudhuri, S. Gulwani, R. Lublinerman, and S. NavidPour. Proving programs robust. In *SIGSOFT/FSE'11*, p. 102–112, 2011.
8 S. Chen and S. Zhou. Recursive mechanism: Towards node differential privacy and unrestricted joins. In SIGMOD, p. 653–664, 2013. ACM.
9 S. Cohen, W. Nutt, and Y. Sagiv. Deciding equivalences among conjunctive aggregate queries. *JACM*, 54(2), 2007.
10 C. Dwork. Differential privacy. In *ICALP*, 2006.
11 C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
12 H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Perspectives in Mathematical Logic. Springer, 1995.
13 M. Gaboardi, E. J. Gallego Arias, J. Hsu, A. Roth, and Z. S. Wu. Dual query: Practical private query release for high dimensional data. In *ICML*, 2014.

14 M. Gaboardi, A. Haeberlen, J. Hsu, A. Narayan, and B. C. Pierce. Linear dependent types for differential privacy. In *POPL*, p. 357–370, 2013.

15 A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar. Differentially private combinatorial optimization. In *SODA*, p. 1106–1125. 2010.

16 M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *NIPS*, p. 2348–2356, 2012.

17 S. Hartmann. On interactions of cardinality constraints, key, and functional dependencies. In FoIKS, p. 136–155, 2000. Springer-Verlag.

18 M. Jha and S. Raskhodnikova. Testing and reconstruction of Lipschitz functions with applications to data privacy. *SIAM J. Comput.*, 42(2):700–731, 2013.

19 N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE07*.

20 A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM TKDD*, 1(1), 2007.

21 D. Maier, A. O. Mendelzon, and Y. Sagiv. Testing implications of data dependencies. *ACM TODS.*, 4(4):455–469, 1979.

22 F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In SIGMOD, p. 19–30, 2009.

23 A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on S&P*, p. 111–125, 2008.

24 J. Nešetřil and P. Ossona de Mendez. Sparsity - Graphs, Structures, and Algorithms, vol. 28 of *Algorithms and combinatorics*. Springer, 2012.

25 K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In STOC '07, p. 75–84, 2007. ACM.

26 C. Palamidessi and M. Stronati. Differential privacy for relational algebra: Improving the sensitivity bounds via constraint systems. In *QAPL*, 2012.

27 J. Reed and B. C. Pierce. Distance makes the types grow stronger: a calculus for differential privacy. In *ICFP* 2010.

28 L. Sweeney. K-anonymity: A model for protecting privacy. In *IJUFKS*, 10(5):557–570, 2002. World Scientific.

29 D. Kifer, and A. Machanavajjhala. No Free Lunch in Data Privacy. In SIGMOD, p. 193–204, 2011.

A Proofs of Section 2

We show that the problem of computing the sensitivity of an FO query is, in general, undecidable. In fact, we show that this is already true for $\sim_1$. Our proof relies on a reduction from the satisfiability problem for first-order logic, via the contingency problem.

▶ Observation 3. For every pair of finite structures $\mathbb{A}, \mathbb{A}'$ differing in relations $\mathcal{S}$, there is $n \in \mathbb{N}$ and structures $\mathbb{B}_1, \ldots, \mathbb{B}_n$ so that $\mathbb{B}_1 = \mathbb{A}$, $\mathbb{B}_n = \mathbb{A}'$ and $\mathbb{B}_i \sim_{\mathcal{S}} \mathbb{B}_{i+1}$ for all $1 \leq i < n$. Thus, for any property P of structures, if there is one structure $\mathbb{A}$ verifying P and some other structure $\mathbb{A}'$ not verifying P, there must be necessarily two such structures so that $\mathbb{A} \sim_{\mathcal{S}} \mathbb{A}'$.

We say that a sentence φ of FO is a **contingency** if φ is satisfiable but not valid. That is, there are structures $\mathbb{A}, \mathbb{A}'$ so that $\mathbb{A} \models \varphi$ and $\mathbb{A}' \not\models \varphi$. The **contingency problem** is the problem of determining, given an FO sentence φ, whether it is a contingency or not.

▶ **Lemma 18.** *The contingency problem for FO is undecidable.*

Proof of Theorem 4. Let ψ be a sentence of FO and P be a monadic predicate not used in ψ. Consider the formula

$$\varphi_\psi(x) = \psi^P \vee P(x),$$

where ψ^P is the relativization of ψ to elements verifying P, that is, the result from replacing every occurrence of $\exists y\, \psi'$ with $\exists y\, (P(y) \wedge \psi')$ and every occurrence of $\forall y\, \psi'$ with $\forall y\, (P(y) \rightarrow \psi')$ in ψ, for every possible y, ψ'. Let $Q_\psi \in \mathrm{FO}^{\#}$ be the query that counts elements verifying φ_ψ, that is, $Q_\psi = \#\varphi_\psi(x)$. We show that

$$\mathrm{GS}^{\sim_1}_{STR}(Q_\psi) = \begin{cases} \infty & \text{if } \psi \text{ is a contingency, or} \\ 1 & \text{otherwise.} \end{cases}$$

Note that if ψ is unsatisfiable, then $\varphi_\psi(x) \equiv P(x)$, and it is easy to see that it has sensitivity of 1. If $\neg\psi$ is unsatisfiable, then $\varphi_\psi(x) \equiv \top$, which also has sensitivity of 1.

Suppose then that ψ is a contingency. Then, there are structures $\mathbb{A}, \mathbb{A}'$ so that $\mathbb{A} \models \psi^P$ and $\mathbb{A}' \not\models \psi^P$. Further, by Observation 3, we can assume that $\mathbb{A} \sim_1 \mathbb{A}'$. Let $\mathbb{B}_n$ be the structure consisting of n elements in no relations. Define $\mathbb{C}_n = \mathbb{A} \sqcup \mathbb{B}_n$ and $\mathbb{C}'_n = \mathbb{A}' \sqcup \mathbb{B}_n$. It is easy to see that $\mathbb{C}_n \sim_1 \mathbb{C}'_n$, $\mathbb{C}_n \models \psi^P$, and $\mathbb{C}'_n \not\models \psi^P$ for every n. Therefore, $Q_\psi(\mathbb{C}_n) = |A| + n$, and $0 \leq Q_\psi(\mathbb{C}'_n) \leq |A'|$, and hence

$$n - k \leq |Q_\psi(\mathbb{C}_n) - Q_\psi(\mathbb{C}'_n)| \leq n + k$$

where $k = \max(|A|, |A'|)$. Thus, $\lim_{n\to\infty} |Q_\psi(\mathbb{C}_n) - Q_\psi(\mathbb{C}'_n)| = \infty$ and therefore

$$\mathrm{GS}^{\sim_1}_{STR}(Q_\psi) = \infty.$$

Hence, ψ is a contingency iff $\mathrm{GS}^{\sim_1}_{STR}(Q_\psi) = \infty$. Since the contingency problem for FO is undecidable by Lemma 18, the statement follows. ◀

B Proofs of Section 3

It is known that the semantics of CQs can be described in terms of homomorphisms, as the following result evidences.

▶ **Lemma 19** (Chandra-Merlin [5]). *For every finite structure $\mathbb{A}$, $n \in \mathbb{N}$, $b_1, \ldots, b_n \in A$ and conjunctive query $\varphi(x_1, \ldots, x_n)$, the following are equivalent*

- $\mathbb{A} \models \varphi[\bar{b}]$,
- *there is a homomorphism* $h : \mathbb{C}_\varphi \to \mathbb{A}$ *so that* $h(x_i) = b_i$ *for every* $i \in \underline{n}$.

The following results are simple (but useful) consequences of the lemma above.

▶ **Lemma 20.** *For every CQ* φ *and* $\mathbb{A} \subseteq \mathbb{B}$ *we have* $\varphi(\mathbb{A}) \subseteq \varphi(\mathbb{B})$.

Proof. Immediate from Lemma 19 and the fact that $\mathbb{A} \to \mathbb{B}$. ◀

▶ **Lemma 21.** *Given a CQ*$^\#$ *query* $Q = \# \bigwedge_{i \in \underline{n}} \varphi_i$ *so that every* φ_i *is a connected conjunct, and given a structure* $\mathbb{A}$*, we have*

$$Q(\mathbb{A}) = \prod_{i \in \underline{n}} \#\varphi_i(\mathbb{A}).$$

Proof. Let $\varphi = \bigwedge_{i \in \underline{n}} \varphi_i(\bar{x}_i)$ and note that $\mathbb{C}^-_\varphi = \sqcup_i \mathbb{C}^-_{\varphi_i}$. Hence,

$$\begin{aligned}
\#\varphi(\mathbb{A}) &= |\{(h(\bar{x}_1), \ldots, h(\bar{x}_n)) \mid h : \mathbb{C}_\varphi \to \mathbb{A}\}| && \text{(by Lemma 19)} \\
&= |\{(h(\bar{x}_1), \ldots, h(\bar{x}_n)) \mid h : \mathbb{C}^-_\varphi \to \mathbb{A}\}| \\
&= |\{h_1(\bar{x}_1), \ldots, h_n(\bar{x}_n) \mid h_i : \mathbb{C}^-_{\varphi_i} \to \mathbb{A} \text{ for all } i \in \underline{n}\}| \\
&= |\{h_1(\bar{x}_1), \ldots, h_n(\bar{x}_n) \mid h_i : \mathbb{C}_{\varphi_i} \to \mathbb{A} \text{ for all } i \in \underline{n}\}| \\
&= \prod_{i \in \underline{n}} \#\varphi_i(\mathbb{A})
\end{aligned}$$

which concludes the proof. ◀

▶ **Lemma 22.** *For every CQ*$^\#$ *query* Q *over* σ_k *we have* $GS^{\sim 1}_{STR}(Q) \geq 1$.

Proof. This is a consequence of the more general Lemma 36 shown in the next section. ◀

▶ **Lemma 23.** *Every CQ query* φ *with at most one free variable is a core-CQ.*

Proof. This is because $core(\mathbb{A}) \cong core(\mathbb{A}, \{a\})$, for any singleton $\{a\}$. ◀

The following are well-known properties of cores.

▶ **Lemma 24** ([5, 9])**.** *For all* $\varphi, \psi \in CQ$*,*

1. $\varphi \equiv core(\varphi)$*;*
2. $\varphi \equiv \psi$ *iff* $core(\varphi) \equiv core(\psi)$ *iff* $\mathbb{C}_{core(\varphi)} \cong \mathbb{C}_{core(\psi)}$ *iff* $\#\varphi \equiv \#\psi$*.*

Proof of Proposition 1. Let $Q = \#\varphi$. We first show $\Delta_{STR}(Q) \leq \mathrm{GS}^{\sim 1}_{STR}(Q)$. Suppose first that $\Delta_{STR}(Q) = \infty$. Let $\mathbb{A} = core(\mathbb{C}_\varphi)$. By construction, there exists a homomorphism $h : \mathbb{C}_\varphi \to core(\mathbb{C}_\varphi)$. Let $x \in free(\varphi)$ and $\bar{t} \in T^{\mathbb{A}}$ so that $h(x) \notin \bar{t}$. Since φ is connected, there must be relations R, S of arity n_R and n_S respectively so that for some $i, j \in \underline{n_S}$, $j' \in \underline{n_R}$, $\bar{r} \in R^{\mathbb{A}}, \bar{s} \in S^{\mathbb{A}}$ we have $\bar{s}[i] = h(x)$, $\bar{s}[j] = \bar{r}[j']$, and $h(x) \notin \bar{r}$.

Let $\mathbb{A}_n$ be the structure resulting from adding n new elements $a_1, \ldots, a_n$ to the domain of $\mathbb{A}$ and the following new tuples to $S^{\mathbb{A}}$

$$\left\{ \bar{s_1}, \ldots, \bar{s_n} \;\middle|\; \begin{array}{l} \forall i \in \underline{n}.\ \forall j \in \underline{arity(S)} \\ \bar{s_i}[j] = a_i \text{ if } \bar{s}[j] = h(x) \text{ and} \\ \bar{s_i}[j] = \bar{s}[j] \text{ otherwise} \end{array} \right\}$$

Informally, the $\bar{s_i}$'s are obtained by replacing in $\bar{s}$ all the occurences of $h(x)$ by a_i. It then follows that $Q(\mathbb{A}_n) > n$.

Let $\mathbb{A}'_n$ be the result of removing $\bar{r}$ from $R^{\mathbb{A}_n}$ in $\mathbb{A}_n$, and let $\mathbb{A}'$ be the result of removing $\bar{r}$ from $R^{\mathbb{A}}$ in $\mathbb{A}$. Note that $\mathbb{A}_n \sim_1 \mathbb{A}'_n$. We now show that $Q(\mathbb{A}'_n) = 0$. Note first that $\mathbb{A}'_n \to \mathbb{A}'$. Therefore, $Q(\mathbb{A}'_n) > 0$ implies $Q(\mathbb{A}') > 0$, which is not true because $\mathbb{A}$ is a core and $\mathbb{A}' \subsetneq \mathbb{A}$ and thus $\mathbb{C}_\varphi \not\to \mathbb{A}'$ by the minimality condition of cores. Thus, $\mathrm{GS}^{\sim_1}_{STR}(Q) = \infty$ since $\lim_{n\to\infty} Q(\mathbb{A}_n) - Q(\mathbb{A}'_n) = \infty$.

Since $\mathrm{GS}^{\sim_1}_{STR}(Q) \geq 1$ by Lemma 22, we conclude $\Delta_{STR}(Q) \leq \mathrm{GS}^{\sim_1}_{STR}(Q)$.

On the other hand, the fact that $\Delta_{STR}(Q) \geq \mathrm{GS}^{\sim_1}_{STR}(Q)$ follows from Lemma 37 that we will show in the next section. ◀

▶ **Lemma 25.** *Let $Q = \#\varphi \in CQ^\#$. If Q has no free variables, then $GS^{\sim_1}_{STR}(\#\varphi) = 1$.*

Proof. For any $\mathbb{A}$, $Q(\mathbb{A}) \in \{0, 1\}$. Thus for any $\mathbb{A} \sim_1 \mathbb{A}'$, $|Q(\mathbb{A}) - Q(\mathbb{A}')| \in \{0, 1\}$. Now by Lemma 22 we know that $\mathrm{GS}^{\sim_1}_{STR}(\#\varphi) \geq 1$, which implies that $\mathrm{GS}^{\sim_1}_{STR}(\#\varphi) = 1$. ◀

▶ **Lemma 26.** *Let $\varphi = \bigwedge_{i\in\underline{n}} \varphi_i$ be a disconnected CQ with the φ_i's being its connected conjuncts. If there are $x, y \in free(\varphi)$ and $k_x, k_y \in \underline{n}$ so that $k_x \neq k_y$, $x \in free(\varphi_{k_x})$, and $y \in free(\varphi_{k_y})$, then $GS^{\sim_1}_{STR}(\#\varphi) = \infty$.*

Proof. Let $Q = \#\varphi$ and $\mathbb{A} = core(\mathbb{C}^-_{\varphi_1}) \sqcup \cdots \sqcup core(\mathbb{C}^-_{\varphi_n})$, and let $h_i : \mathbb{C}^-_{\varphi_i} \to core(\mathbb{C}^-_{\varphi_i})$ for $i \in \underline{n}$. By construction there exists $h : \mathbb{C}^-_\varphi \to \mathbb{A}$. Now note that it must either be that y is an isolated element in $\mathbb{C}^-_{\varphi_{k_y}}$ and $h_{k_y}(y) = y$; or that y appears in a tuple of $\mathbb{C}_{\varphi_{k_y}}$ and thus $h_{k_y}(y)$ appears in a tuple of $core(\mathbb{C}^-_{\varphi_{k_y}})$. Let $\mathbb{A}_m$ be the structure resulting from adding m fresh elements $a_1, \ldots, a_m$ to the domain of $\mathbb{A}$, and for all $R \in \mathcal{R}$ adding the following set of tuples to $R^{\mathbb{A}}$,

$$\left\{ \bar{r_1}, \ldots, \bar{r_m} \;\middle|\; \begin{array}{l} \exists \bar{s} \in R^{\mathbb{A}}.\ \forall i \in \underline{m}.\ \forall j \in \underline{arity(R)} \\ \bar{r_i}[j] = a_i \text{ if } \bar{s}[j] = h_{k_x}(x) \text{ and} \\ \bar{r_i}[j] = \bar{s}[j] \text{ otherwise} \end{array} \right\}$$

Informally, the $\bar{r_i}$'s are obtained by replacing in $\bar{s}$ all the occurences of $h_{k_x}(x)$ by a_i. It then follows by construction that $Q(\mathbb{A}_m) \geq m$. Indeed, there are at least m distinct homomorphisms $h^1, \ldots, h^m : \mathbb{A} \to \mathbb{A}_m$ such that $h^i(y) = h_{k_y}(y)$ and $h^i(x) = a_i$ for $i \in \underline{m}$, and thus m distinct homomorphisms $g^1, \ldots, g^m : \mathbb{C}^-_\varphi \to \mathbb{A}_m$. More precisely we consider $g^i = h^i \circ h$ for $i \in \underline{m}$. We now need to distinguish two cases.

Case y is isolated in $\mathbb{C}^-_{\varphi_{k_y}}$. We consider the structure $\mathbb{A}'_m$ resulting from removing $h_{k_y}(y) = y$ from the domain of $\mathbb{A}_m$. Note that $\mathbb{A}'_m \sim_1 \mathbb{A}_m$, and that by construction, for any $f : \mathbb{C}^-_\varphi \to \mathbb{A}_m$ with $f(a) = y$ for some element a in the domain of φ, then $f : \mathbb{C}^-_\varphi \not\to \mathbb{A}'_m$. But we just saw that there are at least m distinct such homomorphisms. Thus, by Lemma 20 we can conclude that $Q(\mathbb{A}'_m) \leq Q(\mathbb{A}_m) - m$, and thus that $Q(\mathbb{A}_m) - Q(\mathbb{A}'_m) \geq m$

Case y appears in a tuple $\bar{a}$ in $\mathbb{C}^-_{\varphi_{k_y}}$. In that case there exists a relation $R \in \mathcal{R}$ such that $\bar{a} \in R^{\mathbb{C}^-_{\varphi_{k_y}}}$. We consider the structure $\mathbb{A}'_m$ with the same domain as $\mathbb{A}_m$ and resulting from removing $h_{k_y}(\bar{a})$ from $R^{\mathbb{A}_m}$. Note that $\mathbb{A}'_m \sim_1 \mathbb{A}_m$, and that by construction, for any $g : \mathbb{C}^-_\varphi \to \mathbb{A}_m$ with $g(\bar{b}) = h_{k_y}(\bar{a})$ for some tuple $\bar{b}$ in $R^{\mathbb{C}^-_\varphi}$, $g : \mathbb{C}^-_\varphi \not\to \mathbb{A}'_m$, which in turn implies that for any $g : \mathbb{C}^-_\varphi \not\to \mathbb{A}'_m$, $g(y) \neq h_{k_y}(y)$. But we just saw that $h^i(y) = h_{k_y}(y)$ for $i \in \{0, \ldots, m\}$. Thus, by Lemma 20 we can conclude that $Q(\mathbb{A}'_m) \leq Q(\mathbb{A}_m) - m$, and thus that $Q(\mathbb{A}_m) - Q(\mathbb{A}'_m) \geq m$.

We can finally conclude our proof of $\mathrm{GS}^{\sim_1}_{STR}(Q) = \infty$ since $\lim_{m\to\infty} Q(\mathbb{A}_m) - Q(\mathbb{A}'_m) = \infty$. ◀

▶ **Lemma 27.** *Let $\varphi = \bigwedge_{i\in\underline{n}} \varphi_i$ be a disconnected CQ with the φ_i's being its connected conjuncts. If φ has at least one free variable x, and there exists $k \in \underline{n}$ so that for all $y \in free(\varphi)$, $y \in free(\varphi_k)$ and $\mathbb{C}_{\overline{\varphi}^k} \not\to \mathbb{C}_{\varphi_k}$, then $GS^{\sim_1}_{STR}(\#\varphi) = \infty$.*

Proof. Let $Q = \#\varphi$ and $\mathbb{A} = core(\mathbb{C}^-_{\overline{\varphi}^k}) \sqcup core(\mathbb{C}^-_{\varphi_k})$, and let $h_{\overline{k}} : \mathbb{C}^-_{\overline{\varphi}^k} \to core(\mathbb{C}^-_{\overline{\varphi}^k})$ and $h_k : \mathbb{C}^-_{\varphi_k} \to core(\mathbb{C}^-_{\varphi_k})$. By construction there exists $h : \mathbb{C}^-_{\varphi} \to \mathbb{A}$. Let $\mathbb{A}_m$ be the structure resulting from adding m fresh elements $a_1, \ldots, a_m$ to the domain of $\mathbb{A}$, and for all $R \in \mathcal{R}$ adding the following set of tuples to $R^{\mathbb{A}}$,

$$\left\{ \bar{r_1}, \ldots, \bar{r_m} \;\middle|\; \begin{array}{l} \exists \bar{s} \in R^{\mathbb{A}}.\ \forall i \in \underline{m}.\ \forall j \in \underline{arity(R)} \\ \bar{r_i}[j] = a_i \text{ if } \bar{s}[j] = h_k(x) \text{ and} \\ \bar{r_i}[j] = \bar{s}[j] \text{ otherwise} \end{array} \right\}$$

Informally, the $\bar{r_i}$'s are obtained by replacing in $\bar{s}$ all the occurences of $h_k(x)$ by a_i. It then follows by construction that $Q(\mathbb{A}_m) \geq m$. Indeed, there are at least m distinct homomorphisms $h^1, \ldots, h^m : \mathbb{A} \to \mathbb{A}_m$ such that and $h^i(x) = a_i$ and $h^i(y) = y$ for any $y \neq x$ and $i \in \underline{m}$, and thus m distinct homomorphisms $g^1, \ldots, g^m : \mathbb{C}^-_{\varphi} \to \mathbb{A}_m$. More precisely we consider $g^i = h^i \circ h$ for $i \in \underline{m}$. We now need to distinguish two cases.

Case $core(\mathbb{C}^-_{\overline{\varphi}^k})$ is an isolated element. This case cannot occur, as it would either be a constant but this contradicts the definition of $\mathbb{C}^-_{\varphi_k}$; or there would be a homomorphism $g : core(\mathbb{C}^-_{\overline{\varphi}^k}) \to core(\mathbb{C}^-_{\varphi_k})$ which contradicts our hypothesis that $\mathbb{C}_{\overline{\varphi}^k} \to \mathbb{C}_{\varphi_k}$.

Case there is a tuple $\bar{s}$ in $core(\mathbb{C}^-_{\overline{\varphi}^k})$. In that case, there exists $j \in \underline{n}$, $j \neq k$ such that $\mathbb{C}^-_{\varphi_j} \not\to \mathbb{C}^-_{\varphi_k}$, and with $\bar{t}$ in $R^{\mathbb{C}^-_{\varphi_j}}$ and $h_{\overline{k}}(\bar{t}) \in R^{\mathbb{C}^-_{\overline{\varphi}^k}}$. We consider the structure $\mathbb{A}'_m$ with the same domain as $\mathbb{A}_m$ and resulting from removing $h(\bar{t})$ from $R^{\mathbb{C}^-_{\overline{\varphi}^k}}$. Note that $\mathbb{A}'_m \sim_1 \mathbb{A}_m$, and that by construction, for any $g : \mathbb{C}^-_{\varphi} \to \mathbb{A}_m$ with $g(\bar{t}) = h(\bar{t})$, $g : \mathbb{C}^-_{\varphi} \not\to \mathbb{A}'_m$, which in turn implies that $g^i : \mathbb{C}^-_{\varphi} \not\to \mathbb{A}'_m$ for $i \in \underline{m}$, and thus that $Q(\mathbb{A}') = 0$. We can now conclude that $Q(\mathbb{A}_m) - Q(\mathbb{A}'_m) \geq m$. ◀

▶ **Lemma 28.** *Let $\varphi = \bigwedge_{i\in\underline{n}} \varphi_i$ be a disconnected CQ$^{\#}$ with the φ_i's being its connected conjuncts. If there exists $k \in \underline{n}$ so that for all $x \in free(\varphi)$, $x \in free(\varphi_k)$ and $\mathbb{C}_{\overline{\varphi}^k} \to \mathbb{C}_{\varphi_k}$, then $GS^{\sim_1}_{STR}(\#\varphi) = GS^{\sim_1}_{STR}(\#\varphi_k)$.*

Proof. Let $h : \mathbb{C}_{\overline{\varphi}^k} \to \mathbb{C}_{\varphi_k}$, and $\mathbb{A}$ any structure. For any $g : \mathbb{C}_{\varphi} \to \mathbb{A}$, we define $\overline{g}$ as follows:

$$\overline{g}(a) = \begin{cases} g(a) & \text{if } a \in \text{ domain } \mathbb{C}_{\varphi_k} \\ g \circ h(a) & \text{otherwise} \end{cases}$$

By construction, we have that $\overline{g} : \mathbb{C}_{\varphi_k} \to \mathbb{A}$ with $\overline{g}(a) = g(a)$ if $a \in$ domain $\mathbb{C}_{\varphi_k}$. But then it is necessarily the case that $\#\varphi(\mathbb{A}) = \#\varphi_k(\mathbb{A})$. Finally, for any $\mathbb{A} \sim_1 \mathbb{A}'$, $|\#\varphi(\mathbb{A}) - \#\varphi(\mathbb{A}')| = |\#\varphi_k(\mathbb{A}) - \#\varphi_k(\mathbb{A})|$, which implies that $\mathrm{GS}^{\sim_1}_{STR}(\#\varphi) = \mathrm{GS}^{\sim_1}_{STR}(\#\varphi_k)$. ◀

▶ **Theorem 29.** *For every CQ$^{\#}$ Q we have $GS^{\sim_1}_{STR}(Q) = \Delta_{STR}(Q)$.*

Proof of Theorem 8. The proof derives immediately from Proposition 1 and Lemmas 25, 26, 27, ans 28, that handle the five possibilities for Q separately. ◀

C Proofs of Section 4

▶ **Lemma 30.** *For any set of FDs Σ, $\mathcal{C}_\Sigma$ is closed under taking substructures and disjoint unions.*

Proof. Given structures $\mathbb{A}, \mathbb{A}'$ with disjoint set of constants, note that if $\mathbb{A} \sqcup \mathbb{A}'$ does not satisfy an FD $R[i \to j]$, then there must be a connected component that does not satisfy it. This means that either $\mathbb{A}$ or $\mathbb{A}'$ does not satisfy $R[i \to j]$. Therefore, $\mathbb{A} \sqcup \mathbb{A}' \in \mathcal{C}_\Sigma$ provided $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$.

Given structures $\mathbb{A} \subseteq \mathbb{A}'$ so that $\mathbb{A}' \in \mathcal{C}_\Sigma$ note that, for every element $a \in A$ and relation R, we have

$$\begin{aligned} |\{\bar{b}[j] \mid \bar{b} \in R^{\mathbb{A}}, \bar{b}[i] = a\}| &\leq |\{\bar{b}[j] \mid \bar{b} \in R^{\mathbb{A}'}, \bar{b}[i] = a\}| \\ &\leq 1 \end{aligned}$$

and thus $\mathbb{A} \in \mathcal{C}_\Sigma$. ◀

▶ **Lemma 31.** $\#\mathit{chase}_\Sigma(\varphi) \equiv_{\mathcal{C}_\Sigma} \#\varphi$.

Proof. First note that $\mathit{chase}_\Sigma(\varphi)$ and φ may not be equivalent. For example, for $\varphi(x, y, z) = R(x, y) \wedge R(x, z)$ and $\Sigma = \{R[1 \to 2]\}$, we have $\mathit{chase}_\Sigma(\varphi) = R(x, y)$ which has only two free variables as opposed to φ. However, note that any application of the rule of chase, preserves the number of solutions since we are replacing a variable x with another variable y that must be mapped to the same element for any homomorphism to a structure of $\mathcal{C}_\Sigma$. We therefore have that $\#\mathit{chase}_\Sigma(\varphi) \equiv_{\mathcal{C}_\Sigma} \#\varphi$. ◀

▶ **Lemma 32.** *For every CQ φ there is a homomorphism $\mathbb{C}_\varphi \to \mathbb{C}_{\mathit{chase}_\Sigma(\varphi)}$.*

Proof. Remember that we assuming that φ is satisfiable in $\mathcal{C}_\Sigma$ (see note in §Preliminaries) and thus that $\mathit{chase}_\Sigma(\varphi) \neq \bot$. Note that any application of a rule of chase to φ resulting in φ', corresponds to identifying two elements a, a' in $\mathbb{C}_\varphi$. Thus, the function $h(a') = a$ and $h(a'') = a''$ for any $a'' \neq a'$ in C_φ is in fact a homomorphism from $\mathbb{C}_\varphi$ to $\mathbb{C}'_\varphi$. By transitivity of homomorphisms, we obtain that there must be a homomorphism $\mathbb{C}_\varphi \to \mathbb{C}_{\mathit{chase}_\Sigma(\varphi)}$. ◀

▶ **Lemma 33.** $\mathbb{C}_{\mathit{chase}_\Sigma(\varphi)} \in \mathcal{C}_\Sigma$.

Proof. Remember that we assume that φ is satisfiable over $\mathcal{C}_\Sigma$ (see note in §Preliminaries), and thus that $\mathit{chase}_\Sigma(\varphi) \neq \bot$. If $\mathbb{C}_{\mathit{chase}_\Sigma(\varphi)}$ does not satisfy a FD $R[i \to j]$ in Σ, then there must be two tuples $\bar{t}, \bar{s} \in R^{\mathbb{C}_{\mathit{chase}_\Sigma(\varphi)}}$ so that $\bar{t}[i] = \bar{s}[i]$ but $\bar{t}[j] \neq \bar{s}[j]$. But this means that there is a conjunct $R(\bar{t})$ and another one $R(\bar{s})$ with the same properties, which means that $\mathit{chase}_\Sigma(\varphi)$ is not saturated by the rules of chase, which is a contradiction. Thus, $\mathbb{C}_{\mathit{chase}_\Sigma(\varphi)} \in \mathcal{C}_\Sigma$. ◀

▶ **Lemma 34.** $\mathit{chase}_\Sigma(\mathit{core}(\mathit{chase}_\Sigma(\varphi))) = \mathit{core}(\mathit{chase}_\Sigma(\varphi))$.

Proof. If there would be a pair of conjuncts $R(\bar{t})$ and $R(\bar{s})$ in $\mathit{core}(\mathit{chase}_\Sigma(\varphi))$ so that $\bar{t}[i] = \bar{s}[i]$ and $\bar{s}[j] \neq \bar{s}[j]$, this would mean that there are such conjuncts in $\mathit{chase}_\Sigma(\varphi)$ (since the core is isomorphic to a substructure). This cannot be because $\mathit{chase}_\Sigma(\varphi)$ is saturated by the rules of chase, and thus we must have that there is no chase rule applicable to $\mathit{core}(\mathit{chase}_\Sigma(\varphi))$. Hence, $\mathit{chase}_\Sigma(\mathit{core}(\mathit{chase}_\Sigma(\varphi))) = \mathit{core}(\mathit{chase}_\Sigma(\varphi))$. ◀

▶ **Lemma 35.** *core*(*core*(φ)) = *core*(φ).

Proof. By minimality of core, it is idempotent. ◀

▶ **Lemma 36.** *For every CQ query Q over σ_k and set of FD's Σ, we have $GS^{\sim_1}_{\mathcal{C}_\Sigma}(Q) > 0$. Further, there are σ_k-structures $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$ so that $\mathbb{A} \sim_1 \mathbb{A}'$, $\mathbb{A}' \subseteq \mathbb{A}$, $Q(\mathbb{A}) > 0$ and $Q(\mathbb{A}') = 0$.*

Proof. Let $Q = \#\varphi$. Note that $\mathbb{C}_\varphi \to \mathbb{C}_{chase_\Sigma(\varphi)}$ by Lemma 32 and $\mathbb{C}_\varphi \not\to \mathbb{B}_k$, where $\mathbb{B}_k$ is the structure with just constants $\{c_1, \ldots, c_k\}$ and no relations. Further, $\mathbb{C}_{chase_\Sigma(\varphi)} \in \mathcal{C}_\Sigma$ by Lemma 33, $\mathbb{B}_k \subseteq \mathbb{C}_{chase_\Sigma(\varphi)}$ and therefore $\mathbb{B}_k \in \mathcal{C}_\Sigma$ by Lemma 30. Thus, by Observation 3, there must be some $\mathbb{A} \sim_1 \mathbb{A}'$ so that $\mathbb{C}_\varphi \to \mathbb{A}$ and $\mathbb{C}_\varphi \not\to \mathbb{A}'$. Further, it is easy to see that there are such structures so that $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$. ◀

Proof of Proposition 2. Let $\varphi = \bigwedge_{i \in \underline{n}} \psi_i(\bar{x}_i)$ so that each ψ_i is a connected conjunct and assume $\bar{x}_1 \neq ()$. Let $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$ be so that $\mathbb{A} \sim_1 \mathbb{A}'$, $\mathbb{A}' \subseteq \mathbb{A}$, $\#(\bigwedge_{2 \leq i \leq n} \psi_i)(\mathbb{A}) - \#(\bigwedge_{2 \leq i \leq n} \psi_i)(\mathbb{A}') > 0$; by Lemma 36 these structures exist. Let $\mathbb{A}_N = \underbrace{\mathbb{C}^-_{chase_\Sigma(\psi_1)} \sqcup \cdots \sqcup \mathbb{C}^-_{chase_\Sigma(\psi_1)}}_{N \text{ times}}$ and let $\mathbb{B}_N = \mathbb{A}_N \sqcup \mathbb{A}$, $\mathbb{B}'_N = \mathbb{A}_N \sqcup \mathbb{A}'$. Note that $\mathbb{B}_N, \mathbb{B}'_N \in \mathcal{C}_\Sigma$ (as a consequence of Lemmas 30 and 33) and that $\mathbb{B}_N \sim_1 \mathbb{B}'_N$. Since each ψ_i is connected, we have $\#\psi_i(\mathbb{B}_N) = \#\psi_i(\mathbb{A}_N) + \#\psi_i(\mathbb{A})$ and $\#\psi_i(\mathbb{B}'_N) = \#\psi_i(\mathbb{A}_N) + \#\psi_i(\mathbb{A}')$. Let $f_i(N) \stackrel{def}{=} \#\psi_i(\mathbb{A}_N) = N \cdot \#\psi_i(\mathbb{C}^-_{chase_\Sigma(\psi_1)})$ (since ψ_i is connected), $M_i \stackrel{def}{=} \#\psi_i(\mathbb{A})$, $M'_i \stackrel{def}{=} \#\psi_i(\mathbb{A}')$ for $i \in \underline{2}$. Then,

$$
\begin{aligned}
&\#\varphi(\mathbb{B}_N) - \#\varphi(\mathbb{B}'_N) = \\
&= \prod_{i \in \underline{n}} \#\psi_i(\mathbb{B}_N) - \prod_{i \in \underline{n}} \#\psi_i(\mathbb{B}'_N) && \text{(by Lemma 21)} \\
&= \prod_{i \in \underline{n}} (\#\psi_i(\mathbb{A}_N) + \#\psi_i(\mathbb{A})) - \prod_{i \in \underline{n}} (\#\psi_i(\mathbb{A}_N) + \#\psi_i(\mathbb{A}')) \\
&= \prod_{i \in \underline{n}} (f_i(N) + M_i) - \prod_{i \in \underline{n}} (f_i(N) + M'_i)
\end{aligned}
$$

Since M_i, M'_i do not depend on N, and since

- for every i, f_i is non-decreasing by Lemma 20,
- for every i, $M_i \geq M'_i$ by Lemma 20,
- there is some $i \in \underline{n}$, $i > 1$ so that $M_i > M'_i$ since otherwise we would have

$$\#(\bigwedge_{2 \leq i \leq n} \psi_i)(\mathbb{A}) - \#(\bigwedge_{2 \leq i \leq n} \psi_i)(\mathbb{A}') = 0,$$

contradicting our hypothesis,

we obtain that

$$
\begin{aligned}
g(N) &\stackrel{def}{=} \prod_{2 \leq i \leq n} (f_i(N) + M_i) > \prod_{2 \leq i \leq n} (f_i(N) + M'_i) \\
&\stackrel{def}{=} g'(N)
\end{aligned}
$$

for every N. Note that f_1 is strictly increasing since $\mathbb{C}^-_{\psi_1} \to \mathbb{C}^-_{chase_\Sigma(\psi_1)}$ (as a consequence of Lemma 32) and ψ_1 has free variables. Therefore, we have

$$\begin{aligned}
&\lim_{N\to\infty} \#\varphi(\mathbb{B}_N) - \#\varphi(\mathbb{B}'_N) \\
&= \lim_{N\to\infty} (f_1(N) + M_1) \cdot g(N) - (f_1(N) + M'_1) \cdot g'(N) \\
&\geq \lim_{N\to\infty} (f_1(N) + M_1) \cdot g(N) - (f_1(N) + M_1) \cdot g'(N) && \text{(since } M_1 \geq M'_1\text{)} \\
&= \lim_{N\to\infty} (f_1(N) + M_1) \cdot (g(N) - g'(N)) \\
&= \infty && \text{(since } \lim_{N\to\infty} f_1(N) = \infty \text{ and } g(N) - g'(N) > 0\text{)}
\end{aligned}$$

and thus $\mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(\#\varphi) = \infty$. ◀

Proof of Theorem 11. This theorem is the immediate consequence of Lemmas 37 and 39 below. ◀

▶ **Lemma 37.** *Given a set Σ of functional dependencies and a CQ$^\#$ connected query Q, we have that*

$$GS^{\sim_1}_{\mathcal{C}_\Sigma}(Q) \leq \Delta^+_\Sigma(Q).$$

Proof. Let $Q = \#\varphi(\bar{x})$ be a connected CQ$^\#$ query, for $\bar{x} = (x_1, \ldots, x_n)$. By Lemmas 24 and 31 we can assume that $\varphi = core(chase_\Sigma(\psi))$ where we can further assume that $\psi = \varphi$ by Lemmas 34 and 35.

Let $1 < \mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q)$ (otherwise the statement is trivial). Thus, there are $\mathbb{A} \sim_1 \mathbb{A}'$ so that $Q(\mathbb{A}) - Q(\mathbb{A}') = N > 1$. Note that this means that there are N homomorphisms $h_1, \ldots, h_N : \mathbb{C}_\varphi \to \mathbb{A}$ so that $|\{h_i(\bar{x}) \mid i \in \underline{N}\}| = N$ and $h_i : \mathbb{C}_\varphi \not\to \mathbb{A}'$ for all $i \in \underline{N}$.

• If $A = A' \sqcup \{a\}$, then $N \leq 1$, because every $h : \mathbb{C}_\varphi \to \mathbb{A}$ so that $h : \mathbb{C}_\varphi \not\to \mathbb{A}'$ verifies $h(\bar{x}) = (a, \ldots, a)$, as a consequence of φ being connected. This is in contradiction with our assumption $N > 1$.

• Then, it must be that $A = A'$ where for some $R, \bar{b}$ we have $R^{\mathbb{A}} = R^{\mathbb{A}'} \sqcup \{\bar{b}\}$ and $S^{\mathbb{A}} = S^{\mathbb{A}'}$ for every other relational symbol $S \neq R$.

Given $\bar{a} \in R^{\mathbb{C}_\varphi}$ and $p_j : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_j$ for all $j \in \underline{n}$, note that there cannot be more than $\max_j \#_\Sigma(p_j)$ distinct h_i's so that $h_i(\bar{a}) = \bar{b}$, as otherwise $\mathbb{A}$ would not satisfy Σ. Therefore we obtain

$$|\{i \mid h_i(\bar{a}) = \bar{b}\}| \leq \max_{j\in\underline{n}} \min_{p_j : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_j} \#_\Sigma(p_j). \tag{2}$$

Summing over all possible $\bar{a}$ we obtain

$$\begin{aligned}
N &\leq \sum_{\bar{a}\in R^{\mathbb{C}_\varphi}} |\{i \mid h_i(\bar{a}) = \bar{b}\}| \\
&\leq \sum_{\bar{a}\in R^{\mathbb{C}_\varphi}} \max_{i\in\underline{n}} \left(\min_{p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i} \#_\Sigma(p_i) \right) && \text{(by (2))} \\
&\leq \Delta^+_\Sigma(Q).
\end{aligned}$$

Thus, $\mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q) \leq \Delta^+_\Sigma(Q)$. ◀

▶ **Corollary 38.** *For $\Sigma = \emptyset$, the above Lemma 37 implies that $\Delta_{STR}(Q) \geq \Delta^{+}_{\Sigma}(Q) \geq GS^{\sim 1}_{STR}(Q)$, concluding the proof of Theorem 1.*

▶ **Lemma 39.** *Given a set Σ of functional dependencies and a connected $CQ^{\#}_{core}$ query Q, we have that*

$$\Delta^{-}_{\Sigma}(Q) \leq GS^{\sim 1}_{\mathcal{C}_\Sigma}(Q).$$

Proof. Let $Q = \#\varphi(\bar{x})$ be a connected $\mathrm{CQ}^{\#}_{core}$ query, for $\bar{x} = (x_1, \ldots, x_n)$, over a vocabulary σ_k. By Lemma 36, Q has global sensitivity of at least 1. Hence, since $\Delta^{-}_{\Sigma}(Q) \in \{1, \infty\}$, let us suppose that $\Delta^{-}_{\Sigma}(Q) = \infty$ and let us show that $\mathrm{GS}^{\sim 1}_{\mathcal{C}_\Sigma}(Q) = \infty$. Due to Lemmas 24 and 31, we can assume that $\varphi = core(chase_\Sigma(\psi))$, so that $\psi = \varphi$ by Lemmas 34 and 35, and so that $\mathbb{C}_\varphi \cong core(\mathbb{C}_\varphi)$, since $\varphi \in \mathrm{CQ}_{core}$. Note that since $\mathbb{C}_{chase_\Sigma(\varphi)} \in \mathcal{C}_\Sigma$ by Lemma 33 and $\mathbb{C}_{core(chase_\Sigma(\varphi))} \subseteq \mathbb{C}_{chase_\Sigma(\varphi)}$, we have $\mathbb{C}_{core(chase_\Sigma(\varphi))} = \mathbb{C}_\varphi \in \mathcal{C}_\Sigma$ by Lemma 30.

We show that for any given $N \in \mathbb{N}$, $\mathrm{GS}^{\sim 1}_{\mathcal{C}_\Sigma}(Q) \geq N$. Let R and $\bar{a} \in R^{\mathbb{C}_\varphi}$ be as in the definition of $\Delta^{-}_{\Sigma}(Q)$. Let x_j be so that every path $p : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_j$ is so that $\#_\Sigma(p) = \infty$. For each such path p let t_p be the maximum index verifying $p[t_p] = (S, i, a, i', b)$ where

- b is not a constant, and for every $t' > t_p$, $p[t']$ contains no constants, and
- $\#_\Sigma S[i \to i'] = \infty$;

by definition of $\#_\Sigma(p) = \infty$ such t_p must exist. Let

$$K = \{(S, i, i', \bar{b}) \mid p : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_j, \\ p[t_p] = (S, i, \bar{b}[i], i', \bar{b}[i']), \bar{b} \in S^{\mathbb{C}_\varphi}\}.$$

Note that K is a cut-set of edges that induces a partition of $C_\varphi = A \sqcup A'$ —in the sense that if we remove all tuples in K from $\mathbb{C}_\varphi$ we obtain a connected component induced by A— so that $\bar{a} \cap A \neq \emptyset$ and $x_j \in A'$.

Let $\mathbb{B} = \mathbb{C}_\varphi|_A$, $\mathbb{B}' = \mathbb{C}_\varphi|_{A'}$, $\mathbb{B}'_N = \underbrace{\mathbb{B}' \sqcup \cdots \sqcup \mathbb{B}'}_{N \text{ times}}$. Note that, by construction, $\mathbb{B}'$ contains no constants and thus $\mathbb{B}'_N$ is a well-defined σ_k-structure. We build $\mathbb{A}$ as the minimal model so that

- it contains $\mathbb{B} \sqcup \mathbb{B}'_N$ as a substructure
- for every $(S, i, i', \bar{b}) \in K$ and $t \in \underline{N}$ we have $\bar{b}^t \in S^{\mathbb{A}}$, where

$$\bar{b}^t[i] = \begin{cases} \bar{b}[i] \text{ from } \mathbb{B} & \text{if } \bar{b}[i] \in A, \\ \bar{b}[i] \text{ from the } t\text{-th copy of } \mathbb{B}' & \text{if } \bar{b}[i] \in A'. \end{cases}$$

Note that, in particular, $\bar{b}^t[i']$ is in the t-th copy of $\mathbb{B}'$, and $\bar{b}^t[i]$ is in the $\mathbb{B}$ component of $\mathbb{A}$.

Figure 5 depicts the definitions of K, $\mathbb{B}$, $\mathbb{B}'$ and $\mathbb{A}$ in an example.

▶ Claim 1. $\mathbb{A} \in \mathcal{C}_\Sigma$.

Proof. The fact that $\mathbb{B} \sqcup \mathbb{B}'_N \in \mathcal{C}_\Sigma$ is a direct consequence of $\mathbb{C}_\varphi \in \mathcal{C}_\Sigma$ and Lemma 30. Note that if there is a violation of some $S[i \to i'] \in \Sigma$ by the presence of some $\bar{b}, \bar{b}' \in S^{\mathbb{A}}$ so that $\bar{b}[i] = \bar{b}'[i]$ and $\bar{b}[i'] \neq \bar{b}'[i']$, then either

- $\bar{b}, \bar{b}'$ were already in $\mathbb{C}_\varphi$ which is not possible since $\mathbb{C}_\varphi \in \mathcal{C}_\Sigma$, or
- $\bar{b}[i] \in A$ belongs to the $\mathbb{B}$ component of $\mathbb{A}$ and $\bar{b}[i'], \bar{b}'[i'] \in A'$ belong to the $\mathbb{B}'_N$ component of $\mathbb{A}$. In this case, there must be elements $(S, i, i', \bar{b})$ and $(S, i, i', \bar{b}')$ in K which implies that $\#_\Sigma S[i \to i'] = \infty$ and thus $S[i \to i'] \notin \Sigma$, which is in contradiction with our hypothesis.

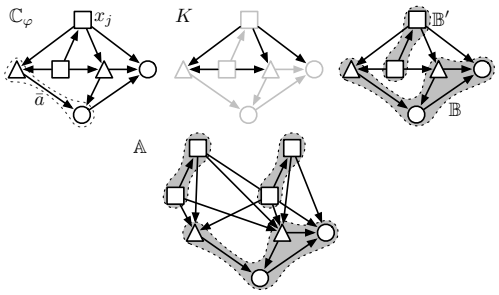

Figure 5 Example of construction of $\mathbb{B}, \mathbb{B}', \mathbb{A}$, assuming that the query uses only one binary relation R, the set of functional dependencies is $\Sigma = \{R[1 \to 2]\}$ and $N = 2$. Square vertices represent free variables, rounded vertices bound variables and triangles constants.

Hence, there are no such $\bar{b}, \bar{b}'$, and thus $\mathbb{A} \in \mathcal{C}_\Sigma$. ◀

▶ Claim 2. $Q(\mathbb{A}) \geq N$.

Proof. Note that if $\varphi(\mathbb{A})$ must contain one answer in which the j-th component is in each of the copies of $\mathbb{C}_\varphi|_{A_2}$. Since there are N copies in $\mathbb{A}$, $|\varphi(\mathbb{A})| \geq N$. ◀

Let $\mathbb{A}'$ be the result of removing $\bar{a}$ from $R^{\mathbb{A}}$ in $\mathbb{A}$. Note that $\mathbb{A}' \in \mathcal{C}_\Sigma$ by Lemma 30 since $\mathbb{A} \in \mathcal{C}_\Sigma$.

▶ Claim 3. $\mathbb{A} \sim_1 \mathbb{A}'$

▶ Claim 4. $Q(\mathbb{A}') = 0$.

Proof. This is because

- $\mathbb{C}_\varphi \cong \mathit{core}(\mathbb{C}_\varphi)$,
- $\mathbb{A}' \to \mathbb{C}_\varphi^-$

where $\mathbb{C}_\varphi^-$ is the result of removing $\bar{a}$ from $R^{\mathbb{C}_\varphi}$ in $\mathbb{C}_\varphi$. Note that $\mathbb{C}_\varphi \to \mathbb{A}'$ would imply $\mathbb{C}_\varphi \cong \mathit{core}(\mathbb{C}_\varphi) \to \mathbb{C}_\varphi^-$, and since $\mathbb{C}_\varphi^- \subsetneq \mathbb{C}_\varphi$ (and thus $\mathbb{C}_\varphi^- \to \mathbb{C}_\varphi$), this would mean that $\mathbb{C}_\varphi$ is not a core. Since this contradicts with our assumptions, we obtain $\mathbb{C}_\varphi \not\to \mathbb{A}'$. ◀

Therefore, $Q(\mathbb{A}) - Q(\mathbb{A}') = N$. ◀

D Proofs of Section 5

We say that $\mathbb{A}$ is an **induced $\mathcal{S}$-substructure** of $\mathbb{B}$ if it is an $\mathcal{S}$-substructure so that $S^{\mathbb{A}} = S^{\mathbb{B}} \cap A^{arity(S)}$ for all $S \in \mathcal{S}$. In this case we say that $\mathbb{A}$ is the **$\mathcal{S}$-substructure induced by** A and we denote it by $\mathbb{B}|_A$. It easy to see that the core of $\mathbb{A}$ is isomorphic to an induced substructure of $\mathbb{A}$.[2]

▶ **Lemma 40.** *For any set of CD's Σ, $\mathcal{C}_\Sigma$ is closed under taking substructures and disjoint unions.*

[2] For more details about the properties of cores, we refer the reader to [24].

Proof. By Lemmas 35 and 24 let us assume that $Q = \#\varphi(\bar{x})$ so that $core(\varphi) = \varphi$, and let $\bar{x} = (x_1, \ldots, x_n)$. Given $\mathbb{A} \sim_1 \mathbb{A}'$ with $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$, we show that $Q(\mathbb{A}) - Q(\mathbb{A}') \leq \Delta^+_\Sigma(Q)$.

• Suppose first that $A = A' \sqcup \{a\}$. By definition of $\sim_1$, it follows that a is in no relation in $\mathbb{A}'$. This, plus the fact that φ is connected, implies that every homomorphism $h : \mathbb{C}_\varphi \to \mathbb{A}$ verifies $h^{-1}(a) = \emptyset$. Thus, $\#\varphi(\mathbb{A}) = \#\varphi(\mathbb{A}')$ and $\#\varphi(\mathbb{A}') - \#\varphi(\mathbb{A}) = 0 \leq \Delta^+_\Sigma(Q)$.

• If $A = A'$, then there must be $R \in \mathcal{R}$ and $\bar{a}$ so that $R^{\mathbb{A}} = R^{\mathbb{A}'} \sqcup \{\bar{a}\}$, and $S^{\mathbb{A}} = S^{\mathbb{A}'}$ for any other $S \in \mathcal{R} \setminus \{R\}$. Let $Q(\mathbb{A}) - Q(\mathbb{A}') = N > 0$. In particular, this means that there are N homomorphisms $h_1, \ldots, h_N : \mathbb{C}_\varphi \to \mathbb{A}$ so that $|\{h_i(\bar{x}) \mid i \in \underline{N}\}| = N$ and $h_i : \mathbb{C}_\varphi \not\to \mathbb{A}'$.

Given $\bar{y} \in R^{\mathbb{C}_\varphi}$, let $p_1, \ldots, p_n$ be so that $p_j : \bar{y} \rightsquigarrow_{\mathbb{C}_\varphi} x_j$ for all $j \in \underline{n}$. This implies that there cannot be more than $\prod_{i \in \underline{n}} \#_\Sigma(p_i)$ distinct h_i's so that $h(\bar{y}) = \bar{a}$. Applying this for all possible $\hat{p}_1, \ldots, \hat{p}_n$ we obtain that there are no more than

$$\min_{\substack{p_1,\ldots,p_n \text{ s.t.} \\ p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i \text{ for } i \in \underline{n}}} \left(\prod_i \#_\Sigma(p_i) \right)$$

h_i's so that $h_i(\bar{y}) = \bar{a}$. Note that for every h_i there must be some $\bar{y} \in R^{\mathbb{C}_\varphi}$ so that $h_i(\bar{y}) = \bar{a}$, as otherwise we would have $h_i : \mathbb{C}_\varphi \to \mathbb{A}'$. Therefore,

$$\begin{aligned} N &\leq \sum_{\bar{a} \in R^{\mathbb{C}_\varphi}} \left(\min_{\substack{p_1,\ldots,p_n \text{ s.t.} \\ p_i : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} h(x_i) \text{ for } i \in \underline{n}}} \left(\prod_i \#_\Sigma(p_i) \right) \right) \\ &\leq \Delta^+_\Sigma(Q). \end{aligned}$$

Thus, $\mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q) \leq \Delta^+_\Sigma(Q)$. ◀

Proof of Theorem 17. We show $\Delta^-_\Sigma(Q) \leq \mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q)$. To this end, we show that there are $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$ so that $\mathbb{A} \sim_1 \mathbb{A}'$ and $|Q(\mathbb{A}) - Q(\mathbb{A}')| \geq \Delta^-_\Sigma(Q)$.

For any structure $\mathbb{A}$ and $a \in A$ we define the **expansion at** a of $\mathbb{A}$, denoted by $\exp(\mathbb{A}, a)$, as the structure with domain $\mathrm{dom}(\exp(\mathbb{A}, a))$ equal to

$$\begin{aligned} \{(\varepsilon, a)\} \cup \{(w, b) \in \mathbb{N}^+ \times A \mid\ & p : a \rightsquigarrow_{\mathbb{A}} b, |p| = |w|, \\ & \text{and for all } \ell \in \underline{|w|} \text{ we have} \\ & 1 \leq w[\ell] \leq \#_\Sigma T[i \to j] \text{ for} \\ & p[\ell] = (T, i, c, j, c')\} \end{aligned}$$

and with an interpretation for $T \in \sigma$

$$T^{\exp(\mathbb{A},a)} = \begin{cases} \{(w, b) \mid b \in T^{\mathbb{A}}\} & \text{if } T \text{ is unary,} \\ \{((w, b), (w', b')) \mid & \\ \quad (b, b') \in T^{\mathbb{A}} \wedge & \\ \quad (w \sim_1 w' \vee w = w')\} & \text{if } T \text{ is binary,} \end{cases}$$

where $w \sim_1 w'$ if for some $t \in \mathbb{N}$ we have $w = w' \cdot t$ or $w \cdot t = w'$. See Figure 6-(b) and -(c) for an example.

Let $R, \bar{a}, p_1, \ldots, p_n$ be the witnesses of the max's and min in the definition of $\Delta^-_\Sigma(Q)$. Let us assume that R is binary and that

$$\bar{a}[1] \neq \bar{a}[2] \tag{3}$$

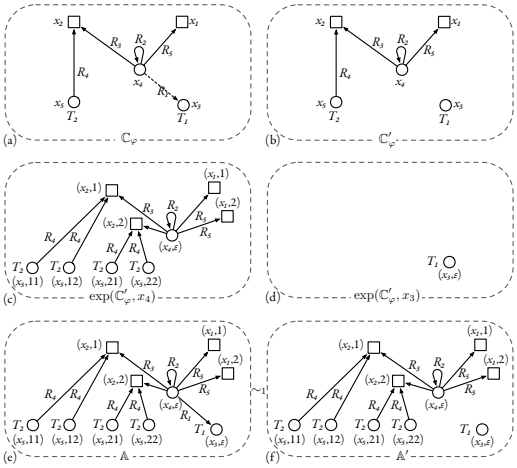

Figure 6 Example of construction of $\mathbb{A}, \mathbb{A}'$, assuming $R = R_1$, $\bar{a} = (x_4, x_3)$ and $\Sigma = \{R_1[1 \xrightarrow{1} 2],\ R_1[2 \xrightarrow{1} 1],\ R_3[1 \xrightarrow{2} 2],\ R_3[2 \xrightarrow{1} 1],\ R_4[1 \xrightarrow{1} 2],\ R_4[2 \xrightarrow{2} 1],\ R_5[1 \xrightarrow{2} 2],\ R_5[2 \xrightarrow{1} 1]\}$, and $\varphi(x_1, x_2) = \exists x_3, x_4, x_5\ .\ R_2(x_4, x_4) \wedge R_1(x_4, x_3) \wedge R_5(x_4, x_1) \wedge R_3(x_4, x_2) \wedge R_4(x_5, x_2) \wedge T_1(x_3) \wedge T_2(x_5)$.

since the cases where this does not hold are only easier. Let $\mathbb{C}'_\varphi \sim_1 \mathbb{C}_\varphi$ be the result of removing $\bar{a}$ from $R^{\mathbb{C}_\varphi}$ in $\mathbb{C}_\varphi$. Let $\mathbb{A}' = \exp(\mathbb{C}'_\varphi, \bar{a}[1]) \cup \exp(\mathbb{C}'_\varphi, \bar{a}[2])$ and let $\mathbb{A}$ be the result of adding $((\varepsilon, \bar{a}[1]), (\varepsilon, \bar{a}[2]))$ to $R^{\mathbb{A}'}$. Figure 6 contains an example.

▶ Observation 4. Note that $\mathbb{A} \sim_1 \mathbb{A}'$.

▶ Observation 5. For every $(w, a), (w', a') \in A$ so that $a = a'$ we have $|w| = |w'|$.

▶ Claim 5. $\mathbb{A}, \mathbb{A}' \in \mathcal{C}_\Sigma$.

Proof. By Lemma 40, it suffices to show that $\mathbb{A} \in \mathcal{C}_\Sigma$ since $\mathbb{A}' \subseteq \mathbb{A}$.

Let $T[i \to j]k \in \Sigma$ and let (w, b) be an element of $\mathbb{A}$. We show that

$$|\{\bar{b}[j] \mid \bar{b} \in T^{\mathbb{A}}, \bar{b}[i] = (w, b)\}| \leq k. \tag{4}$$

Wlog let us fix $i = 1$ and $j = 2$. By definition of $\mathbb{A}$ it follows

$$\begin{aligned} &\{\bar{b}[j] \mid \bar{b} \in T^{\mathbb{A}}, \bar{b}[i] = (w, b)\} = \\ &\{(w', b') \mid (b, b') \in T^{\mathbb{C}_\varphi} \wedge (w \sim_1 w' \vee w = w')\}. \end{aligned} \tag{5}$$

Let us define types of binary relations:

(a) the one contaning all non-looping relations of $\mathbb{C}'_\varphi$ (*i.e.*, $\{S \mid \exists a \in C_\varphi, \pi : \underline{2} \to \underline{2}$ s.t. $a \neq b \wedge (a, b)_\pi \in S^{\mathbb{C}'_\varphi}\}$);
(b) the one contaning all looping relations of $\mathbb{C}_\varphi$ (*i.e.*, $\{S \mid (b, b) \in S^{\mathbb{C}'_\varphi}\}$); and
(c) $\{R \mid \bar{a}[1] = b \vee \bar{a}[2] = b\}$.

By the no repeated joins assumption, these three sets are disjoint. In fact, they are a partition of the binary relations containing b in $\mathbb{C}_\varphi$. We now show that (4) holds for T in any of these sets of relations.

(a) Let us first asume that T is of the type (a). The following are direct consequences of the equality (5) above.

- By the hypothesis of no repeated joins, for every element a of $\mathbb{C}_\varphi$ there is at most one b in $\mathbb{C}_\varphi$ so that $(a,b) \in T^{\mathbb{C}_\varphi}$ or $(b,a) \in T^{\mathbb{C}_\varphi}$. Therefore, by definition of $\mathrm{dom}(\exp(\mathbb{C}_\varphi, \bar{a}[1]))$ [resp. $\mathrm{dom}(\exp(\mathbb{C}_\varphi, \bar{a}[2]))$] there are no more than $\#_\Sigma T[i \to j]$ (and thus no more than k) elements (b', w') so that $w \sim_1 w'$ and $(b, b') \in T^{\mathbb{C}_\varphi}$ in $\exp(\mathbb{C}_\varphi, \bar{a}[1])$ [resp. $\exp(\mathbb{C}_\varphi, \bar{a}[2])$].
- The images of the interpretation of binary relations of $\exp(\mathbb{C}_\varphi, \bar{a}[1])$ and of $\exp(\mathbb{C}_\varphi, \bar{a}[2])$ are disjoint, due to the acyclicity of $\mathbb{C}_\varphi$ and the assumption (3) that $\bar{a}[1] \neq \bar{a}[2]$. In particular, there is no element present simultaneously in $T^{\exp(\mathbb{C}_\varphi, \bar{a}[1])}$ and $T^{\exp(\mathbb{C}_\varphi, \bar{a}[2])}$.

Thus, there are no more than k elements (w', b') in $\mathbb{A}$ so that $(b, b') \in T^{\mathbb{C}_\varphi}$ with $w' \sim_1 w$.

(b) If T is of type (b), then there is at most one element (w', b') so that $(b, b') \in T^{\mathbb{C}_\varphi}$ and $(w \sim_1 w' \vee w = w')$, which is the case where $(w, b) = (w', b')$.

(c) If $T = R$, then there is at most one element (w', b') so that $(b, b') \in T^{\mathbb{C}_\varphi}$ and $(w \sim_1 w' \vee w = w')$, which is when $b = \bar{a}[1]$, $b' = \bar{a}[2]$ (or viceversa), and $w = w' = \varepsilon$.

Thus, we have that (4) holds, which proves the statement. ◀

▶ Claim 6. $Q(\mathbb{A}') = 0$.

Proof. Note that there is no homomorphism $h : \mathbb{C}_\varphi \to \mathbb{A}'$ since otherwise the homomorphic image of $\bar{a}$ would need to verify $h(\bar{a}) \in R^{\mathbb{A}'}$, while we know by construction that $R^{\mathbb{A}'} = \emptyset$. Then, there is no $\bar{b}$ in $\mathbb{A}'$ so that $\mathbb{A}' \models \varphi[\bar{b}]$ and thus $|\varphi(\mathbb{A}')| = 0$. ◀

▶ Claim 7. For every $\bar{b} = ((w_1, b_1), \ldots, (w_n, b_n))$ so that $(w_i, b_i) \in A$ and $b_i = x_i$ for all $i \in \underline{n}$ we have that $\mathbb{A} \models \varphi[\bar{b}]$.

Proof. Given $\bar{b}$ as above, we define the homomorphism $h : \mathbb{C}_\varphi \to \mathbb{A}$ as follows. We define $h(\bar{a}[1]) = (\varepsilon, \bar{a}[1])$, $h(\bar{a}[2]) = (\varepsilon, \bar{a}[2])$. For every b lying in the (unique) simple path of $\mathbb{A}'$ between $\bar{a}[1]$ and b_i for some i, we define $h(b) = (w, b)$, where w is a prefix of w_i. Note that there is only one such w for a given b due to Observation 5.

Finally, we define inductively all the remaining elements. For every b' already defined, and for every b with $dist_{\mathbb{C}_\varphi}(b, b') = 1$, we define $h(b) = (w, b)$, where $h(b') = (w', b')$ and $w = w'1$.

It is not hard to see that h is indeed a homomorphism from $\mathbb{C}_\varphi$ to $\mathbb{A}$, and that $(h(x_1), \ldots, h(x_n)) = \bar{b}$. Hence, $\mathbb{A} \models \varphi[\bar{b}]$. ◀

▶ Claim 8. There are $\Delta_\Sigma^-(Q)$ many tuples of the form $((w_1, b_1), \ldots, (w_n, b_n)) \in A^n$ with $b_i = x_i$ for all $i \in \underline{n}$.

Proof. Let $i \in \underline{n}$, and let $r \in \{1, 2\}$ be so that there is a path from $\bar{a}[r]$ to x_i in $\mathbb{C}'_\varphi$, let $p_i : \bar{a}[r] \rightsquigarrow_{\mathbb{C}_\varphi} x_i$ and let $t = |p_i|$. The number of elements (w, x_i) in A is then given by the cardinality of

$$S = \{v \in \mathbb{N}^t \mid p_i[j] = (T, s, a, s', b) \wedge 1 \leq v[j] \leq \#_\Sigma T[s \to s'] \text{ for all } j \in \underline{t}\},$$

which is equal to $\#_\Sigma(p_i)$. Thus, the number of tuples of the form above is given by $N \stackrel{def}{=} \prod_{i \in \underline{n}} \#_\Sigma(p_i)$.

▶ Observation 6. Note that if we would take $p'_i : \bar{a}[3-r] \rightsquigarrow_{\mathbb{C}_\varphi} x_i$ then $\#_\Sigma(p_i) \leq \#_\Sigma(p'_i)$. Thus, $\#_\Sigma(p_i) = \min_{p : \bar{a} \rightsquigarrow_{\mathbb{C}_\varphi} x_i} \#_\Sigma(p)$.

By the previous Observation 6, we have

$$N = \min_{p_1,\ldots,p_n \text{ s.t. } p_i:\bar{a}\rightsquigarrow_{\mathbb{A}}\{x_i\} \text{ for all } i} \left(\prod_i \#_\Sigma(p_i)\right)$$

and by hypothesis on R and $\bar{a}$ we obtain

$$N = \max_{R\in\sigma}\left(\max_{\bar{a}\in R^{\mathbb{A}}}\left(\min_{\substack{p_1,\ldots,p_n \text{ s.t.}\\ p_i:\bar{a}\rightsquigarrow_{\mathbb{A}}\{x_i\} \text{ for all } i}}\left(\prod_i \#_\Sigma(p_i)\right)\right)\right)$$

which proves the statement. ◀

Therefore, by Claims 7 and 8 we have that $Q(\mathbb{A}) \geq \Delta^-_\Sigma(Q)$, and due to Claim 6 we then obtain that $Q(\mathbb{A}') - Q(\mathbb{A}) \geq \Delta^-_\Sigma(Q)$. Hence, by Claim 5 with Observation 4 we obtain that $\mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q) \geq \Delta^-_\Sigma(Q)$.

Finally, if $\bar{a}[1] \neq \bar{a}[2]$ [resp. if R is unary] we define $\mathbb{A}' = \exp(\mathbb{C}'_\varphi, \bar{a}[1])$ and $\mathbb{A}$ as the result of adding $((\varepsilon, \bar{a}[1]), (\varepsilon, \bar{a}[1]))$ [resp. $(\varepsilon, \bar{a}[1])$] to $R^{\mathbb{A}'}$. It is easy to check that the same argument works in this case. ◀

Proof of Corollary 38. Note that $\Delta_{STR}(Q) \geq \Delta^+_\Sigma(Q)$ for $\Sigma = \emptyset$. Indeed, if $\Delta_{STR}(Q) = 1$ this means that every free variable $x \in \mathit{free}(\varphi)$ is included in every tuple. Therefore, $\min_{p:\bar{a}\rightsquigarrow_{\mathbb{C}_\varphi} x} \#_\Sigma(p) = 1$ for every free variable x and tuple $\bar{a}$ and thus $\Delta^+_\Sigma(Q) = 1$. Thus, since $\Delta_{STR}(Q) \in \{1, \infty\}$, the above Lemma 37 implies that $\Delta_{STR}(Q) \geq \Delta^+_\Sigma(Q) \geq \mathrm{GS}^{\sim_1}_{\mathcal{C}_\Sigma}(Q) = \mathrm{GS}^{\sim_1}_{STR}(Q)$ since $\mathcal{C}_\Sigma = STR$, concluding the proof of Theorem 1. ◀