RESEARCH-ARTICLE

Query Answering Under Volume-Based Diversity Functions

MARCELO ARENAS, Pontifical Catholic University of Chile, Santiago, RM, Chile

TIMO CAMILLO MERKL, Vienna University of Technology, Vienna, Vienna, Austria

REINHARD PICHLER, Vienna University of Technology, Vienna, Vienna, Austria

CRISTIAN RIVEROS, Pontifical Catholic University of Chile, Santiago, RM, Chile

Proceedings of the ACM on Management of Data, Volume 3, Issue 5 (November 2025)
https://doi.org/10.1145/3767717
EISSN: 2836-6573

Query Answering Under Volume-Based Diversity Functions

MARCELO ARENAS, Pontificia Universidad Católica de Chile, Chile, IMFD, Chile, and RelationalAI, USA
TIMO CAMILLO MERKL, TU Wien, Austria
REINHARD PICHLER, TU Wien, Austria
CRISTIAN RIVEROS, Pontificia Universidad Católica de Chile, Chile and IMFD, Chile

When query evaluation produces too many tuples, a new approach in query answering is to retrieve a diverse subset of them. The standard approach for measuring the diversity of a set of tuples is to use a distance function between tuples, which measures the dissimilarity between them, to then aggregate the pairwise distances of the set into a score (e.g., by using sum or min aggregation). However, as we will point out in this work, the resulting diversity measures may display some unintuitive behavior. Moreover, even in very simple settings, finding a maximally diverse subset of the answers of fixed size is, in general, intractable and little is known about approximations apart from some hand-picked distance-aggregator pairs.

In this work, we introduce a novel approach for computing the diversity of tuples based on volume instead of distance. We present a framework for defining volume-based diversity functions and provide several examples of these measures applied to relational data. Although query answering of conjunctive queries (CQ) under this setting is intractable in general, we show that one can always compute a (1-1/e)-approximation for any volume-based diversity function. Furthermore, in terms of combined complexity, we connect the evaluation of CQs under volume-based diversity functions with the ranked enumeration of solutions, finding general conditions under which a (1-1/e)-approximation can be computed in polynomial time.

CCS Concepts: • **Theory of computation** → **Database theory**.

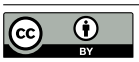
Additional Key Words and Phrases: Query evaluation, diversity, conjunctive queries.

ACM Reference Format:
Marcelo Arenas, Timo Camillo Merkl, Reinhard Pichler, and Cristian Riveros. 2025. Query Answering Under Volume-Based Diversity Functions. *Proc. ACM Manag. Data* 3, 5 (PODS), Article 281 (November 2025), 18 pages. https://doi.org/10.1145/3767717

1 Introduction

When the set of answers to a query gets too big, a user might be better served by being presented a meaningful subset of the answers rather than being overwhelmed with the entire set. Clearly, sampling might provide one way of selecting a "representative" subset of the answers. However, as was pointed out in [26], such an approach typically misses interesting but rarely occurring answers. An alternative approach, which has recently received increased attention by the database community, is to aim at a small, *diverse* set of answers [1, 2, 15, 20, 23]. For instance (following an example given in [12]), in a car dealership setting, the number of models satisfying the constraints expressed by the customer may be huge. Therefore, rather than presenting all solutions to this constraint satisfaction problem (which is a well-known equivalent problem to conjunctive query

Authors' Contact Information: Marcelo Arenas, Pontificia Universidad Católica de Chile, Santiago, Chile and IMFD, Santiago, Chile and RelationalAI, Berkeley, USA, marenas@uc.cl; Timo Camillo Merkl, TU Wien, Vienna, Austria, timo.merkl@tuwien.ac.at; Reinhard Pichler, TU Wien, Vienna, Austria, reinhard.pichler@tuwien.ac.at; Cristian Riveros, Pontificia Universidad Católica de Chile, Santiago, Chile and IMFD, Santiago, Chile, cristian.riveros@uc.cl.



ACM 2836-6573/2025/11-ART281
https://doi.org/10.1145/3767717

answering [17]), it would be more useful to come up with a small, *diverse* set of solutions and let the customer decide on which type of models to focus further discussions.

The most common approach of assigning a diversity score δ to a subset of the universe $\mathcal{U}$ (e.g., the answers to a query) is to first define a distance measure d between any two distinct elements of $\mathcal{U}$ and then define the diversity $\delta(S)$ of any subset S of $\mathcal{U}$ by applying some aggregation to the pairwise distances of the elements in S [14]. Typical distance measures in the database context are the Hamming distance [1, 6, 20] (i.e., counting the positions in which two tuples differ), an ultrametric [25, 26] (i.e., imposing an order on the attributes and considering tuples farther apart if they differ on an attribute further up in this order), or the Euclidean distance for numeric attributes [1]. Typical aggregate functions are the sum and min operators [1, 6, 20].

While sum and min are natural and familiar aggregate functions, they may lead to some anomalies of the resulting diversity measure: In case of the sum operator, consider a setting where a set S contains two elements t_1, t_2 with high distance $d(t_1, t_2)$. Then adding to S another element t_1' very close to t_1 but again with high distance from t_2 seemingly leads to a significant increase of diversity, even though t_1' is almost a "copy" of t_1. In [27] several desiderata on diversity measures are presented – including the *twin property*, i.e., adding an (almost) identical copy should not increase the diversity, and *monotonicity*, i.e., adding a new element to a set S never decreases the diversity. Clearly, the sum operator violates the first fundamental property while the min operator violates this second property. Consequently, Weitzman [27] introduced a diversity measure (henceforth referred to as δ_{W}) based on a more sophisticated aggregation of pairwise distances. However, as was shown in [2], even the basic task of determining the diversity $\delta_{\mathrm{W}}(S)$ of a given set S of elements is NP-complete.

The goal of our work is to introduce a novel framework for defining diversity measures, such that this framework is generally applicable but, at the same time, particularly well suited for defining natural diversity measures in the (relational) database world. We will thus introduce a two-staged approach which, in the first place, assigns to each element of the universe (e.g., a tuple in a relation or in an entire database) a *volume* in the form of some measurable set. As will be illustrated in Section 3, for a set S of tuples, there are many ways of choosing such a volume. In the simplest case, we could just collect the set of values occurring in S. Various other options, such as considering k-ary balls of a pre-specified radius r around a k-tuple of numerical attributes are presented in Section 3. The second stage then consists in assigning values to the unions of these measurable sets. For the basic case of collecting the set V of values occurring in S, we could simply take the cardinality of V. For the case of k-ary balls associated with each tuple, we would take the volume of the union of the balls associated with the tuples in S. A formal definition of our *volume-based* approach to diversity will be given in Section 3.

We will then study interesting properties of this approach. In particular, we will analyze its relationship with previous approaches – in particular, Weitzman's approach [27] and the multi-attribute approach of Nehring and Puppe [21] (in Section 4, we will formally define that approach and also point at its major shortcoming, namely the conceptual and computational complexity caused by having to deal with the powerset of the powerset of the universe). Somewhat surprisingly, we will show that diversity functions defined via the multi-attribute approach can also be defined in our framework and, for a finite universe, also the converse holds. In other words, while avoiding the negative computational properties of the multi-attribute approach, our volume-based diversity measures share the favorable properties shown in [21]. One of them is *submodularity*, which formalizes the intuition that adding a new element to a smaller set potentially leads to a bigger increase of diversity than adding the same element to a bigger set.

When analyzing computational properties of volume-based diversity measures, submodularity will prove beneficial. Concretely, we study the problem of searching for a subset of the answers to a conjunctive query which, for a given size k, maximizes the diversity. This problem has been studied

before for various diversity functions and, even in very simple settings (e.g., considering the sum or the minimum of the pairwise Hamming distances) this problem was shown to be intractable [20] – even for data complexity. We will show that also for the natural volume-based diversity measures of sets of tuples presented in Section 3, intractability holds. We therefore study the search for a maximally diverse set of answers to a conjunctive query from an approximation point of view. For data complexity, we prove a tractable $(1 - 1/e)$-approximation (where e is the Euler number) of the maximum diversity score for arbitrary volume-based diversity measures by making use of a classical approximation result for submodular set functions by Nemhauser et al. [22]. We also show that, in general, a better tractable approximation can be excluded unless P = NP. Clearly, combined complexity requires further restrictions since even query evaluation of Boolean conjunctive queries (without paying any attention to diversity) is NP-complete [7]. However, by restricting our attention to CQs of bounded fractional hypertreewidth [11] and establishing a relationship with ranked enumeration [9], we manage to achieve tractable $(1 - 1/e)$-approximation of the maximum diversity score also for combined complexity.

Structure of the paper and summary of results. After recalling some basic notions in Section 2, we will formally introduce our volume-based framework of defining diversity functions in Section 3. By presenting some examples of natural diversity functions for sets of tuples, we illustrate the suitability of this framework in the database context. We then study the relationship of our volume-based approach of defining diversity measures with previous approaches, namely with the multi-attribute approach of [21] in Section 4 and with distance-based approaches (above all Weitzman's diversity measure [27]) in Section 5. The search for a maximally diverse set of k answers to a conjunctive query Q over a given database D is studied in Sections 6 and 7. As mentioned above, a tractable exact solution to this maximization problem is out of reach. We therefore settle for an approximation. In Section 6, we study data complexity and establish a tractable $(1 - 1/e)$-approximation of the maximum diversity score of k-element subsets of the answers to first-order queries by virtue of the submodularity of volume-based diversity measures. In Section 7, we study combined complexity and identify a sufficient condition on the queries to achieve the same quality of approximation. We conclude with Section 8. Proof details can be found in the full version of this paper [3].

2 Preliminaries

Sets and sequences. We denote by $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_{\geq 0}$ the set of natural, real and non-negative real numbers, respectively. Given a set A, we denote by $\text{finite}(A)$ the set of all non-empty finite subsets of A. For $k \in \mathbb{N}$, we say that $B \in \text{finite}(A)$ is a *k-subset* if $|B| = k$. We usually use a, b, or c to denote elements, and $\bar{a}$, $\bar{b}$, or $\bar{c}$ to denote sequences of such elements. For $\bar{a} = a_1, \ldots, a_k$, we write $\bar{a}[i] := a_i$ to denote the i-th element of $\bar{a}$ and $|\bar{a}| := k$ to denote the length of $\bar{a}$. Further, given a function f we write $f(\bar{a}) := f(a_1), \ldots, f(a_k)$ to denote the function applied to each element of $\bar{a}$.

Conjunctive queries. Fix a set $\mathbb{D}$ of data values. A relational schema Σ (or just schema) is a pair $(\mathcal{R}, \text{arity})$, where $\mathcal{R}$ is a set of relation names and $\text{arity} : \mathcal{R} \to \mathbb{N}$ assigns each name to a number. An R-tuple of Σ (or just a tuple) is a syntactic object $R(a_1, \ldots, a_k)$ such that $R \in \mathcal{R}$, $a_i \in \mathbb{D}$ for every i, and $k = \text{arity}(R)$. We will write $R(\bar{a})$ to denote a tuple with values $\bar{a}$. Given a schema Σ, we denote by $\mathbb{T}_\Sigma$ the set of all tuples over Σ with values in $\mathbb{D}$. A *relational database* D over Σ is a finite set of tuples over Σ. For a schema $\Sigma = (\mathcal{R}, \text{arity})$ and a set of variables $\mathcal{X}$ disjoint from $\mathbb{D}$, a *Conjunctive Query* (CQ) over Σ is a syntactic structure of the form:

$$Q(\bar{x}) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$$

such that Q denotes the answer relation, each R_i is a relation name in $\mathcal{R}$, $\bar{x}_i$ is a sequence of variables in $\mathcal{X}$, $|\bar{x}| = \text{arity}(Q)$, and $|\bar{x}_i| = \text{arity}(R_i)$ for every $i \leq m$. Further, $\bar{x}$ is a sequence of variables

appearing in $\bar{x}_1, \ldots, \bar{x}_m$. We refer to such a CQ simply as Q, where $Q(\bar{x})$ and $R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$ are called the *head* and the *body* of Q, respectively. Furthermore, we call each $R_i(\bar{x}_i)$ an *atom* of Q, and we say that Q is a *full* CQ if each variable occurring in the body of Q also appears in the head of Q.

Let Q be a CQ of the above form, and D be a database over the same schema Σ. A *homomorphism* from Q to D is a function $h : \mathcal{X} \to \mathbb{D}$ such that $R_i(h(\bar{x}_i)) \in D$ for every $i \leq m$. We define the *answers* of Q over D as the set of Q-tuples $[\![Q]\!](D) := \{Q(h(\bar{x})) \mid h \text{ is a homomorphism from } Q \text{ to } D\}$.

Distance-based diversity. Let $\mathcal{U}$ be an infinite set. We see $\mathcal{U}$ as a *universe* of possible solutions and $S \in \text{finite}(\mathcal{U})$ as a candidate finite set of solutions. In its most general form, a diversity function over $\mathcal{U}$ is a function $\delta : \text{finite}(\mathcal{U}) \to \mathbb{R}_{\geq 0} \cup \{\infty\}$. The standard approach, that we call here *distance-based diversity* functions, is to first define a *distance function* $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ (typically d is a metric on $\mathcal{U}$) and to define the diversity δ as an extension of d from pairs to arbitrary subsets of $\mathcal{U}$ setting $\delta(S) = 0$ if $|S| \leq 1$. As proposed in [14], one way of defining δ for a given distance function d is to define an aggregator f that combines the pairwise distances. That is, we set $\delta(S) := f(d(a,b)_{a,b \in S})$. The most common aggregators are sum and min, which give rise to the following diversity functions:

$$\delta_{\text{sum}}(S) := \sum_{a,b \in S} d(a,b) \quad \text{and} \quad \delta_{\text{min}}(S) := \min_{a,b \in S \,:\, a \neq b} d(a,b).$$

3 Volume-based Diversity Framework

In this section, we introduce a general volume-based framework for measuring the diversity of sets of tuples. We begin by recalling the definitions of σ-algebra and measures. Then, we introduce the main definitions of the framework, present several examples to motivate the use of volume-based diversity measures over relational data, and prove that volume-based diversity functions satisfy two fundamental properties expected of diversity measures.

Measures. We recall here the standard definitions of σ-algebra and measures (see e.g. [4] for further details). Let Ω be a set (possibly infinite). A *σ-algebra* over Ω is a family $\mathcal{S}$ of subsets of Ω (i.e., $\mathcal{S} \subseteq 2^\Omega$) such that (1) $\varnothing \in \mathcal{S}$, (2) if $X \in \mathcal{S}$, then $\Omega \setminus X \in \mathcal{S}$, and (3) if $X_i \in \mathcal{S}$ for every $i \in \mathbb{N}$, then $\bigcup_{i \in \mathbb{N}} X_i \in \mathcal{S}$. Given a σ-algebra $\mathcal{S}$, a *measure* for $\mathcal{S}$ is a function $\mu : \mathcal{S} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that (1) $\mu(\varnothing) = 0$ and (2) if $X_i \in \mathcal{S}$ for every $i \in \mathbb{N}$ and $X_i \cap X_j = \varnothing$ for every $i, j \in \mathbb{N}$ with $i \neq j$, then:

$$\mu\Big(\bigcup_{i \in \mathbb{N}} X_i\Big) \;=\; \sum_{i \in \mathbb{N}} \mu(X_i).$$

For example, assuming that Ω is a countable set, one can check that 2^Ω is a σ-algebra and $\mu_{\text{count}} : 2^\Omega \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ that maps $\mu_{\text{count}}(X) = |X|$ if X is finite and $\mu_0(X) = \infty$, otherwise, is a measure, called the *counting measure*. Another example is the *weighted measure*, where we consider a weight function $w : \Omega \to \mathbb{R}_{\geq 0}$ over Ω and define $\mu_w(X) = \sum_{a \in X} w(a)$ where the sum is defined as the supremum of $\sum_{a \in Y} w(a)$ over all finite subsets $Y \subseteq X$. A particular case here is a probability distribution over a σ-algebra $\mathcal{S}$ where μ assigns a probability in $[0, 1]$ to each subset X of Ω.

The volume-based framework. Assume that $\mathcal{U}$ is the *universe of possible solutions* over which we want to measure diversity. A *volume assignment* $\mathcal{V}$ over $\mathcal{U}$ is a tuple $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ such that $\mathcal{S}$ is a σ-algebra over a set Ω (that may be different from $\mathcal{U}$), μ is a measure for $\mathcal{S}$ and $\beta : \mathcal{U} \to \mathcal{S}$. Intuitively, the function β, called the *ball function*, is a function that assigns a *ball* in $\mathcal{S}$ to each element a of the universe $\mathcal{U}$, namely, it assigns a volume to a.

We now introduce our framework for defining diversity functions over volume assignments as follows. Given a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ over $\mathcal{U}$, a function $\delta_{\mathcal{V}} : \text{finite}(\mathcal{U}) \to \mathbb{R}_{\geq 0} \cup \{\infty\}$

D_1 :

R	
a	a
a	b
b	a

D_2 :

R	
a	a
a	b
b	a
b	b

D_3 :

R	
a	b
a	c

Fig. 1. Databases D_1, D_2 and D_3 consisting of a binary relation R with data values a, b, and c.

is a *volume-based diversity function* over $\mathcal{U}$ if for every $S \in \text{finite}(\mathcal{U})$:

$$\delta_{\mathcal{V}}(S) \;=\; \mu\big(\bigcup_{a \in S} \beta(a)\big).$$

Intuitively, each element of S contributes with different characteristics to the diversity of the group (i.e., its volume), and when we add all these characteristics together, the intersection only adds once. In particular, when two elements a_1 and a_2 of $\mathcal{U}$ are totally different with respect to the diversity (i.e. $\beta(a_1) \cap \beta(a_2) = \varnothing$), we have that $\delta_{\mathcal{V}}(\{a_1, a_2\}) = \delta_{\mathcal{V}}(\{a_1\}) + \delta_{\mathcal{V}}(\{a_2\})$. Also, note that, contrary to distance-based diversity functions (defined in Section 2), it is not necessary that $\delta_{\mathcal{V}}(\{a\}) = 0$ (indeed, $\delta_{\mathcal{V}}(\{a\}) \neq 0$ almost surely). Depending on the application context, positive diversity of singletons might actually be the desired behavior. For instance, suppose that the universe $\mathcal{U}$ denotes the set of employees, Ω a set of skills, and β assigns to each employee her skills. Then the diversity $\delta(S)$ assigns a measure (via μ) to the skill set present in a team S of employees. In this case, we clearly want δ applied to a singleton to reflect the value of the skills possessed by each individual.

In the following, we provide several examples of volume-based diversity functions applied to relational data (i.e., tuples). For this purpose, recall that we use $\mathbb{D}$ to denote a set of data values, Σ to denote an arbitrary relational schema, and $R(a_1, \ldots, a_k)$ to denote an R-tuple of Σ where $a_i \in \mathbb{D}$ for every i. In the following examples, we use $\mathbb{T}_\Sigma$ to denote our universe (i.e., $\mathcal{U}$) of all possible tuples.

Example 3.1. Let $\mathcal{V}_{\text{elem}} = (2^{\mathbb{D}}, \mu_{\text{count}}, \beta_{\text{elem}})$ be the volume assignment such that $\mathcal{S} = 2^{\mathbb{D}}$ is the σ-algebra (over $\mathbb{D}$), μ_{count} is the counting measure and β_{elem} is the ball function defined as:

$$\beta_{\text{elem}}(R(a_1, \ldots, a_k)) \;:=\; \{a_1, \ldots, a_k\}$$

for every tuple $R(a_1, \ldots, a_k)$. For every finite set of tuples $S \subseteq \mathbb{T}_\Sigma$, we have that $\delta_{\mathcal{V}_{\text{elem}}}(S)$ measures the number of different data values contained in the tuples in S. That is, the more different data values the tuples have, the more diverse they are.

For instance, consider the databases D_1, D_2 and D_3 in Figure 1 consisting of a binary relation R. We have that $\beta_{\text{elem}}(R(a,a)) = \{a\}$, $\beta_{\text{elem}}(R(a,b)) = \{a,b\}$, $\beta_{\text{elem}}(R(b,a)) = \{a,b\}$ and

$$\delta_{\mathcal{V}_{\text{elem}}}(D_1) = \mu_{\text{count}}(\beta_{\text{elem}}(R(a,a)) \cup \beta_{\text{elem}}(R(a,b)) \cup \beta_{\text{elem}}(R(b,a))) = \mu_{\text{count}}(\{a,b\}) = |\{a,b\}| = 2.$$

In the same way, we conclude that $\delta_{\mathcal{V}_{\text{elem}}}(D_2) = 2$ and $\delta_{\mathcal{V}_{\text{elem}}}(D_3) = 3$. Hence, D_1 and D_2 are equally diverse under the measure $\delta_{\mathcal{V}_{\text{elem}}}$, while D_3 is considered more diverse than these two databases, as it contains an extra value. □

Example 3.2. In addition to measuring the diversity in data values, we now also want to consider the position where these data values occur, i.e., it is different whether a appears in the first or second component of a tuple. We thus capture the intuition that different attributes, even if they have the same data type, have a different semantics (e.g., in a car-relation, the number 6 occurring both in the "gears" and in the "cylinders" attribute does not reduce the diversity). For this purpose, consider the volume assignment $\mathcal{V}_{\text{pos}} = (2^{\mathbb{D}\times\mathbb{N}}, \mu_{\text{count}}, \beta_{\text{pos}})$ where we use the σ-algebra $2^{\mathbb{D}\times\mathbb{N}}$ (over

$\mathbb{D} \times \mathbb{N}$), the counting measure μ_{count} and the ball function β_{pos} such that:

$$\beta_{\text{pos}}(R(a_1, \ldots, a_k)) \;:=\; \{(a_1, 1), \ldots, (a_k, k)\}$$

for every tuple $R(a_1, \ldots, a_k) \in \mathbb{T}_\Sigma$. Then, the diversity $\delta_{\mathcal{V}_{\text{pos}}}(S)$ measures the number of different values that appear in different positions of the tuples in $S \subseteq \mathbb{T}_\Sigma$. For instance, consider again the databases D_1, D_2 and D_3 given in Figure 1. Then we have that $\beta_{\text{pos}}(R(a,a)) = \{(a,1),(a,2)\}$, $\beta_{\text{pos}}(R(a,b)) = \{(a,1),(b,2)\}$, $\beta_{\text{pos}}(R(b,a)) = \{(b,1),(a,2)\}$ and

$$\delta_{\mathcal{V}_{\text{pos}}}(D_1) = \mu_{\text{count}}(\beta_{\text{pos}}(R(a,a)) \cup \beta_{\text{pos}}(R(a,b)) \cup \beta_{\text{pos}}(R(b,a))) = \mu_{\text{count}}(\{(a,1),(a,2),(b,2),(b,1)\}) = 4.$$

In the same way, we conclude that $\delta_{\mathcal{V}_{\text{pos}}}(D_2) = 4$ and $\delta_{\mathcal{V}_{\text{pos}}}(D_3) = 3$. Hence, as opposed to the diversity measurements given in Example 3.1, D_1 and D_2 are equally diverse under the measure $\delta_{\mathcal{V}_{\text{pos}}}$, while D_3 is considered less diverse than these two databases, as it contains a smaller number of values in different positions. □

Example 3.3. Another practical example of volume-based diversity functions is considering a weight function $w : \mathbb{D} \to \mathbb{R}_{\geq 0}$. For instance, if the relational data considers animals in $\mathbb{D}$, then a user could use a weight function where $w(\text{'dog'})$ will weigh less than $w(\text{'dodo'})$ given that dodo is a less common animal than a dog. Then one can consider the volume assignment $\mathcal{V}^w_{\text{elem}} = (2^{\mathbb{D}}, \mu_w, \beta_{\text{elem}})$ where the σ-algebra and ball functions are the same as in $\mathcal{V}_{\text{elem}}$ (see Example 3.1) and the measure μ_w is the weighted measure defined above. Then $\delta_{\mathcal{V}^w_{\text{elem}}}(S)$ measures the weight of the data values appearing in tuples, assigning more diversity to tuples where a dodo appears versus a dog.

One can naturally extend this example to also consider the positions of the data values (denoted by $\mathcal{V}^w_{\text{pos}}$) as in Example 3.2 and, instead of a weight function w, one can use a probability function that assigns a probability to each data value. To showcase $\mathcal{V}^w_{\text{pos}}$, consider again the databases D_1, D_2 and D_3 given in Example 3.1. Moreover, assume that c is an uncommon value for the second attribute of R, which is represented by the following weight function: $w((a,1)) = w((a,2)) = w((b,1)) = w((b,2)) = w((c,1)) = 1$, and $w((c,2)) = 3$. Then we have that:

$$\delta_{\mathcal{V}^w_{\text{pos}}}(D_3) = \mu_w(\beta_{\text{pos}}(R(a,b)) \cup \beta_{\text{pos}}(R(a,c))) = \mu_w(\{(a,1),(b,2),(c,2)\}) = w((a,1)) + w((b,2)) + w((c,2)) = 5.$$

In the same way, we conclude that $\delta_{\mathcal{V}^w_{\text{pos}}}(D_1) = 4$ and $\delta_{\mathcal{V}_{\text{pos}}}(D_2) = 4$. Hence, in this case D_3 is considered as the most diverse database given the occurrence of c in the second column of R. □

Example 3.4. Let D be a relational database over a schema Σ and consider a CQ $Q(\bar{x}) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$. A user may want to measure the diversity of a subset $S \subseteq [\![Q]\!](D)$ concerning the provenance of each tuple, namely, which are the tuples in D that contribute to the outputs in S (cf. the "which provenance" studied in [8]). One way to formalize this is as follows. Let $[\![Q]\!](D)$ be the universe of possible solutions. Consider the volume assignment $\mathcal{V}_{Q,D} = (2^D, \mu_{\text{count}}, \beta_{Q,D})$ where 2^D is the σ-algebra (i.e., all subsets of tuples in D), μ_{count} is the counting measure, and $\beta_{Q,D} : [\![Q]\!](D) \to 2^D$ is the ball function such that for every answer $Q(\bar{a}) \in [\![Q]\!](D)$:

$$\beta_{Q,D}(Q(\bar{a})) \;:=\; \{R_i(h(\bar{x}_i)) \mid 1 \leq i \leq m \text{ and } h \text{ is a homomorphism from } Q \text{ to } D \text{ with } h(\bar{x}) = \bar{a}\}$$

In other words, $\beta_{Q,D}$ maps $Q(\bar{a})$ to all the tuples that contribute to it, that is, its provenance. For $S \subseteq [\![Q]\!](D)$, the value $\delta_{\mathcal{V}_{Q,D}}(S)$ counts the number of different tuples in D that support the outputs in S. Then, the more different tuples support S, the more diverse they are. For instance, consider again the database D_1 given in Example 3.1, and let $Q_1(x,y)$ be the conjunctive query $\exists z\, R(x,z) \wedge R(z,y)$.

Then the tuples (a, a) and (b, b) are both answers to $Q_1(x, y)$ over D_1. However, we have that $\beta_{Q_1,D_1}(Q_1(a, a)) = \{R(a, a), R(a, b), R(b, a)\}$, $\beta_{Q_1,D_1}(\{Q_1(b, b)\}) = \{R(b, a), R(a, b)\}$ and

$$\delta_{\mathcal{V}_{Q_1,D_1}}(\{Q_1(a, a)\}) = \mu_{\text{count}}(\beta_{Q_1,D_1}(Q_1(a, a))) = \mu_{\text{count}}(\{R(a, a), R(a, b), R(b, a)\}) = 3,$$
$$\delta_{\mathcal{V}_{Q_1,D_1}}(\{Q_1(b, b)\}) = \mu_{\text{count}}(\beta_{Q_1,D_1}(Q_1(b, b))) = \mu_{\text{count}}(\{R(b, a), R(a, b)\}) = 2.$$

Hence, in this case, $Q_1(a, a)$ is considered a more diverse answer than $Q_1(b, b)$, as there is a larger number of ways in which (a, a) can be obtained as an answer to $Q_1(x, y)$ over D_1. □

Example 3.5. We now consider a more geometrical scenario where $\mathbb{D} = \mathbb{R}$ and a tuple $R(a_1, \ldots, a_k)$ represents points in the $\mathbb{R}^k$-space. Then, given a radius $r > 0$ we can define the volume assignment $\mathcal{V}_r = (\mathcal{B}^k, \mu, \beta_r)$ where $\mathcal{B}^k$ are the Borel sets of $\mathbb{R}^k$ (i.e., measurable sets), μ is the Lebesgue measure (i.e., measures the volume of a measurable set in $\mathcal{B}^k$), and β_r is the ball function such that:

$$\beta_r(R(a_1, \ldots, a_k)) \;:=\; \{(b_1, \ldots, b_k) \in \mathbb{R}^k \mid \sqrt{(a_1 - b_1)^2 + \ldots + (a_k - b_k)^2} \leq r\}$$

namely, β_r assigns a ball of radius r under euclidean distance around $(a_1, \ldots, a_k)$. Then the volume-based diversity function $\delta_{\mathcal{V}_r}(S)$ measures the volume of r-balls around points in S. In particular, the farther apart (up to radius r) the points in S, the more diverse they are. □

Example 3.6. As our last example, we adapt the previous example to have points closer to the tuples $R(a_1, \ldots, a_k)$ contribute more to the diversity than points further away by adding Gaussian functions. To that end, again let $\mathbb{D} = \mathbb{R}$ and a tuple $R(a_1, \ldots, a_k)$ represents points in the $\mathbb{R}^k$-space. Then, we can define the volume assignment $\mathcal{V}_g = (\mathcal{B}^{k+1}, \mu, \beta_g)$ where $\mathcal{B}^{k+1}$ are the Borel sets of $\mathbb{R}^{k+1}$ (note that we added a dimension), μ is the Lebesgue measure, and β_g is the ball function with

$$\beta_g(R(a_1, \ldots, a_k)) \;:=\; \{(b_1, \ldots, b_k, d) \in \mathbb{R}^{k+1} \mid 0 \leq d \leq e^{-(a_1-b_1)^2-\ldots-(a_k-b_k)^2}\}$$

namely, β_g assigns the area under the Gaussian function centered around $(a_1, \ldots, a_k)$. Then the volume-based diversity function $\delta_{\mathcal{V}_g}(S)$ measures the collective volume under the Gaussian functions, i.e., the integral $\int_{\mathbb{R}^k} \max_{s \in S} e^{\|x-s\|_2^2}\, dx$. A benefit of using Gaussian over simple boxes is that adding a new element will always increase the diversity at least a bit.

Monotonicity and submodularity. We conclude this section by introducing two fundamental properties of diversity functions, advocated for in [21, 27].

Fix a universe $\mathcal{U}$ of possible solutions and a volume assignment $\mathcal{V}$ over $\mathcal{U}$. A first desirable property of diversity functions is that of *monotonicity*: adding an element to a set cannot decrease the diversity of the set. Formally, a diversity function δ is monotone if $\delta(S \cup \{a\}) \geq \delta(S)$ for every $S \in \text{finite}(\mathcal{U})$ and $a \in \mathcal{U}$.

A second desirable property of diversity functions is *submodularity*[1], which means that, for every $a \in \mathcal{U}$ and $S_1, S_2 \in \text{finite}(\mathcal{U})$ with $S_1 \subseteq S_2$, the property $\delta(S_1 \cup \{a\}) - \delta(S_1) \geq \delta(S_2 \cup \{a\}) - \delta(S_2)$ holds. As mentioned in Section 1, submodularity captures the intuition that adding an element a to the smaller set S_1 should result in a greater increase in diversity than adding it to S_2.

In the next proposition, we show that both properties are satisfied by volume-based diversity functions, thereby providing evidence of the naturalness of our approach.

Proposition 3.7. *Let $\mathcal{V}$ be any volume assignment over a universe $\mathcal{U}$ of possible solutions. Then $\delta_{\mathcal{V}}$ is always monotone and submodular.*

[1]We note that, in contrast to Nehring and Puppe [21], Weitzman [27] does not explicitly propose *submodularity* as a desideratum. However, he mentions that, ideally, the increase of diversity when adding a new element a to $\mathcal{U}$, should correspond to the minimum distance of a from the already existing elements in $\mathcal{U}$. Clearly, this property implies submodularity.

4 Characterizing Volume-based Diversity Functions through the Multi-Attribute Model

In [21], Nehring and Puppe proposed a different and novel approach by introducing "multi-attribute" diversity functions. The idea here is to consider *attributes* of the elements in a universe $\mathcal{U}$ as subsets of $\mathcal{U}$, i.e., each attribute is characterized by the elements that share this attribute. Like Weitzman [27], the authors drew the motivation for their approach above all from the diversity of species in biodiversity and the study of acts that ensure high (expected) diversity among them. Basic attributes could then be, for instance, "being a mammal" or "living in the ocean", etc., and each of these attributes is then represented by the set of the corresponding animals.

For a finite set X, Nehring and Puppe [21] formally define diversity functions as follows. Let λ be a non-negative measure (an additive set function) on 2^{2^X}. For $A \subseteq X$, we write λ_A rather than $\lambda(\{A\})$. Then, for $S \subseteq X$, the diversity v_λ is defined as

$$v_\lambda(S) \;=\; \lambda(\{A \subseteq X \mid A \cap S \neq \varnothing\}) \;=\; \sum_{A \subseteq X \,:\, A \cap S \neq \varnothing} \lambda_A.$$

Intuitively, this definition considers each subset $A \subseteq X$ as an attribute (or a "feature class") that may contribute to the diversity of S. The weight λ_A quantifies the relevance or distinctiveness of that attribute. A subset S is then considered diverse if it collectively touches many of these informative subsets A, each with non-negative weight.

We now establish the relationship between this notion and our volume-based approach.

Theorem 4.1. *Let X be a finite set. If v_λ is a multi-attribute diversity function, then there exists a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ over the universe $\mathcal{U} = X$, such that $v_\lambda = \delta_\mathcal{V}$. Likewise, if $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ is a volume assignment over some* finite *universe $\mathcal{U}$, then there exists a non-negative measure λ on 2^{2^X} with $X = \mathcal{U}$, such that $\delta_\mathcal{V} = v_\lambda$.*

Proof sketch. For given multi-attribute diversity function v_λ, defining an equivalent volume-based diversity function $v_\mathcal{V}$ is straightforward. More precisely, we set $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ with $\mathcal{S} = 2^{2^X}$, $\beta(x) = \{A \subseteq X \mid x \in A\}$, and $\mu(\mathcal{B}) = \sum_{A \in \mathcal{B}} \lambda_A$ for $\mathcal{B} \in \mathcal{S}$. The other direction is more involved and only works for finite universe $\mathcal{U}$. In particular, we set $\lambda_A = \mu\big(\bigcap_{a \in A} \beta(a) \setminus \bigcup_{x \in X \setminus A} \beta(x)\big)$. □

This characterization is important for several reasons. First, it confirms that volume-based diversity functions are at least as expressive as multi-attribute ones, thereby unifying two frameworks under a common perspective. Second, the volume-based framework avoids the computational burden of working directly over the power set of the power set in the multi-attribute formulation, and instead operates over a more intuitive geometric or set-based representation of diversity. Moreover, by the correspondence with the multi-attribute diversity model, our volume-based diversity functions inherit all favorable properties proved for the former in [21]. In particular, the fact that volume-based diversity functions are monotone and submodular, as shown in Proposition 3.7, follows directly from this equivalence.

Finally, a fundamental advantage of the volume-based framework is its suitability for relational data. In this setting, tuples from a relation can be mapped to measurable regions in a space defined by the attributes occurring in a tuple or its provenance, allowing the use of volume as a principled measure of diversity. For example, the balls $\beta(t)$ assigned to tuples t can reflect their attribute values or provenance sets, while the measure μ can reflect weighted or count-based semantics over these regions. This enables a natural and scalable representation of diversity across query answers without requiring explicit enumeration of exponentially many subsets, as is needed in the multi-attribute approach. In contrast, the latter becomes infeasible in large relational domains due to its dependence on attribute power sets. Volume-based diversity is thus more aligned with the semantics and structure of relational databases.

5 Distance-Based versus Volume-Based Diversity Functions

As has already been mentioned in Section 1, a common way of defining diversity of outputs in the database area is by using the distance-based approach. In contrast, we have defined the diversity $\delta_{\mathcal{V}}$ via a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ over $\mathcal{U}$. This raises the question of what are the differences or similarities between the two approaches, and how can we compare them.

Comparison by properties. A direct way to compare the two approaches is in terms of properties. As we already noticed, volume-based diversity functions are always monotone and submodular. In contrast, almost all distance-based diversity functions are not submodular, and some are not even monotone.

Proposition 5.1. *There exists a metric such that its corresponding diversity functions δ_{sum} and δ_{min} are not submodular. Further, δ_{min} is not even monotone.*

Although this fact is direct, it provides evidence that the two approaches differ considerably for δ_{sum} and δ_{min}. In the following, we provide further evidence of their differences and similarities.

Volume-based as distance-based. Another way to compare the two approaches is to try to encode volume-based diversity functions by using a distance-based approach. As we will see, in general, this is not possible. To that end, we first discuss two natural approaches to define a distance function given a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$.

Specifically, for one, we can define $d^{\triangle}_{\mathcal{V}}(a, b)$ as the measure μ of the symmetric difference of $\beta(a)$ and $\beta(b)$, i.e., $d^{\triangle}_{\mathcal{V}}(a, b) = \mu((\beta(a) \setminus \beta(b)) \cup (\beta(b) \setminus \beta(a)))$. We observe that any distance measure $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_{\geq 0}$ defined from a volume-based diversity $\delta_{\mathcal{V}}$ as $d := d^{\triangle}_{\mathcal{V}}$ is a *pseudo-metric*, i.e., it satisfies *non-negativity* (i.e., $d(a, b) \geq 0$ for all $a, b \in \mathcal{U}$), *symmetry* (i.e., $d(a, b) = d(b, a)$), *identity* ($d(a, a) = 0$ for all $a \in \mathcal{U}$), and the *triangle inequality* (i.e., $d(a, c) \leq d(a, b) + d(b, c)$ for all $a, b, c \in \mathcal{U}$). If in addition, $d(a, b) = 0$ implies $a = b$ for all $a, b \in \mathcal{U}$, then d is actually a metric.

A second option (essentially considered in [21] in the context of the multi-attribute approach) is to define the distance function $d^{M}_{\mathcal{V}}$ as the marginal $d^{M}_{\mathcal{V}}(a, b) := \delta_{\mathcal{V}}(\{a, b\}) - \delta_{\mathcal{V}}(\{b\})$. However, in that case, we give up symmetry. Note that this can be recovered when the diversity of all singletons are the same. In that case, $d^{M}_{\mathcal{V}}$ again becomes a pseudo-metric.

Now, the hope could be that $d^{\triangle}_{\mathcal{V}}$ or $d^{M}_{\mathcal{V}}$ (or any other pseudo-metric) combined with an appropriate aggregator can recover the expressiveness of $\delta_{\mathcal{V}}$. To that end, we denote by $\delta_{agg,d}$ a distance-based diversity function defined through an aggregator function agg and a pseudo-metric d, namely, $\delta_{agg,d}(S) := agg(d(a, b)_{a,b \in S})$. We say that agg is *monotone* if $agg((d_i)_i) \leq agg((d'_i)_i)$ when $d_i \leq d'_i$ for all i. Further, we say that a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ is *oblivious to data values* if for any bijection $f : \mathbb{D} \to \mathbb{D}$ and for any set of tuples $S \subseteq \mathbb{T}_{\Sigma}$ we have:

$$\mu\Big(\bigcup_{R(\bar{a}) \in S} \beta(R(\bar{a}))\Big) = \mu\Big(\bigcup_{R(\bar{a}) \in S} \beta(R(f(\bar{a})))\Big).$$

We also say that pseudo-metric d is *oblivious to data values* if $d(R(\bar{a}), R(\bar{a}')) = d(R(f(\bar{a})), R(f(\bar{a}')))$. Essentially, this means that the diversity functions should not depend on the concrete data values that appear as constants in the tuples but instead only on whether constants are equal or not. Clearly, from the examples presented in Section 3, the volume assignments $\mathcal{V}_{\text{elem}}$ and $\mathcal{V}_{\text{pos}}$ are oblivious to data values while $\mathcal{V}^{w}_{\text{elem}}$, $\mathcal{V}^{w}_{\text{pos}}$, $\mathcal{V}_{Q,D}$, $\mathcal{V}_r$, and $\mathcal{V}_g$ are in general not oblivious to data values. When it comes to metrics, naturally, the Hamming-distance is an example of a metric oblivious to data values while the Euclidean-distance is not.

Theorem 5.2. *There exists a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ (e.g., $\mathcal{V}_{\text{elem}}$) oblivious to data values over tuples $\mathbb{T}_{\Sigma}$ such that there does not exist a monotone aggregator agg and pseudo-metric d over $\mathbb{T}_{\Sigma}$*

that is oblivious to data values and can distinguish the same sets as $\mathcal{V}$. In other words, no matter the agg, d, there are two k-subsets $S, S' \subseteq \mathbb{T}_\Sigma$ such that $\delta_{\mathcal{V}}(S) \neq \delta_{\mathcal{V}}(S')$ while $\delta_{agg,d}(S) = \delta_{agg,d}(S')$.

Distance-based as volume-based. We now consider the other direction and see if one can understand the distance-based approach in terms of volumes. Of course, this is not possible in general as volume-based diversity functions are always monotone and submodular (Proposition 3.7) while natural distance-based diversity functions are neither (Proposition 5.1). But this leaves open the question if more sophisticated distance-based diversity functions like Weitzman's δ_{W} can be captured by volumes. Below we give a partially positive answer to this question: In general, this is not possible, as we can show that Weitzman's diversity function δ_{W} is, in general, not submodular. However, in the most important special case considered in [27], namely if the distance function d underlying δ_{W} is an ultrametric, then δ_{W} is essentially a volume-based diversity function.

For a distance function d, The diversity function δ_{W} is defined recursively as follows:

$$\delta_W(S) := \max_{a \in S} \left(\delta_W(S \setminus \{a\}) + d(a, S \setminus \{a\})\right),$$

with base case $\delta_W(\{a\}) := 0$. The distance $d(a, S)$ is defined as $\min_{x \in S} d(a, x)$.

Weitzman's diversity function is motivated by applications to species hierarchies. However, one shortcoming is that δ_W is not generally submodular:

Proposition 5.3. *Weitzman's diversity measure δ_W is, in general, not submodular.*

Another shortcoming of δ_W is its computational complexity: even computing $\delta_W(S)$ for a given S is, in general, intractable [2]. However, if d is an ultrametric (i.e., it satisfies the strong triangle inequality $d(a, c) \leq \max(\{d(a, b), d(b, c)\})$) then the computation becomes tractable [27]. Moreover, in this case, δ_W becomes essentially volume-based:

Theorem 5.4. *Let Weitzman's diversity measure be defined over a distance function d that is an ultrametric over some finite set X. Then there exists a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ such that $\delta_{\mathcal{V}} = \delta_W + r$, where r denotes the radius of the ultrametric (i.e., the max. distance between any two elements in X).*

The above result illustrates a key advantage of our volume-based framework: it subsumes and generalizes the best-performing cases of the distance-based approach. In particular, ultrametrics have been identified as a desirable form of distance for diversity due to their favorable computational properties [2] and their suitability for modeling hierarchical systems [27]. Note that a hierarchical notion of distance naturally fits relationally structured data as is illustrated in [25, 26], where the distance between two tuples is based on the first position at which they differ: tuples with longer common prefixes are considered closer. A typical example is a car relation with attributes such as 'make', 'model', 'color', and 'year'. Under this ultrametric, diversification is done according to the attribute order: first one tries to diversify 'make' , then 'model', then 'color', and finally 'year'.

By Theorem 5.4, our framework naturally captures ultrametric diversity functions as a special case – up to an additive constant – through an appropriate volume assignment. This demonstrates that volume-based diversity not only provides a broader modeling language for diversity but also inherits and extends the desirable theoretical guarantees associated with ultrametric distances. As such, it offers a principled and unified framework for defining well-behaved diversity measures.

6 Query Evaluation Under Volume-Based Diversity Functions

In this section, we start our study of CQ evaluation under volume-based diversity functions in data complexity (i.e., the query is fixed). We start by showing that this problem is hard in general for most of the volume assignments $\mathcal{V}$ presented in Section 3. Despite this negative result, we show

that under some reasonable assumptions on $\mathcal{V}$, we can always find a $(1 - 1/e)$-approximation of a maximally diverse k-subset of the solutions in polynomial time under data complexity.

Hardness of exact computation. Let Σ be a schema and $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ be a volume assignment over $\mathbb{T}_\Sigma$. Further, let Q be a CQ over Σ. We are interested in the following computational problem:

Problem:	CQEval$[\Sigma, \mathcal{V}, Q]$
Input:	A database D over Σ and $k \geq 1$
Output:	$\arg\max_{S \subseteq [\![Q]\!](D) : \lvert S\rvert = k} \delta_{\mathcal{V}}(S)$

In other words, given a database D and a number $k \geq 1$, we want to compute a k-subset S of $[\![Q]\!](D)$ that maximizes the volume diversity $\delta_{\mathcal{V}}(S)$ over all k-subsets. Note that Σ and Q are fixed; namely, we measure the computational resources of the problem in data complexity. Furthermore, the volume assignment $\mathcal{V}$ and, thus, the diversity function $\delta_{\mathcal{V}}$ are also fixed. We implicitly assume that if $k > |[\![Q]\!](D)|$, then we output all the tuples in $[\![Q]\!](D)$. In particular, if $[\![Q]\!](D) = \varnothing$, then an algorithm for CQEval$[\Sigma, \mathcal{V}, Q]$ outputs $\varnothing$. By slight abuse of notation, we will formulate intractability results of CQEval$[\Sigma, \mathcal{V}, Q]$ in the form of "NP-hardness". Strictly speaking, the NP-hardness applies to the decision variant of the problem CQEval$[\Sigma, \mathcal{V}, Q]$, i.e., deciding if $\delta_{\mathcal{V}}(S)$ is above a given threshold *th* for some $S \subseteq [\![Q]\!](D)$ subject to $|S| = k$.

We will always assume that $\mathcal{V}$ and $\delta_{\mathcal{V}}$ are fixed in all query evaluation problems studied in this paper (see also Section 7). Moreover, for the sake of simplification, in this section we will assume that for any volume assignment $\mathcal{V}$ and any set S of tuples, computing $\delta_{\mathcal{V}}(S)$ takes constant time[2]. Intuitively, one can consider $\delta_{\mathcal{V}}$ as a black box in the system that can be evaluated efficiently for a set of tuples whose complexity does not considerably affect the query evaluation process. Clearly, if we show that CQEval$[\Sigma, \mathcal{V}, Q]$ is hard, then it is even harder if the cost of computing $\delta_{\mathcal{V}}$ is included. The other way around, if we show that CQEval$[\Sigma, \mathcal{V}, Q]$ can be evaluated in polynomial time, this result will be subjected that $\delta_{\mathcal{V}}$ can also be efficiently evaluated (which is typically the case for natural volume assignments $\mathcal{V}$).

Unfortunately, similar to previous work on query evaluation under diversity functions, we can show that CQEval is NP-hard for most of the volume assignments $\mathcal{V}$ presented in Section 3.

Theorem 6.1. *The problem CQEval$[\Sigma, \mathcal{V}, Q]$ is NP-hard if* $\mathcal{V} \in \{\mathcal{V}_{\text{elem}}, \mathcal{V}_{\text{pos}}, \mathcal{V}^{w}_{\text{elem}}, \mathcal{V}^{w}_{\text{pos}}, \mathcal{V}_{Q,D}\}$.

Given that for simple volume assignments like $\mathcal{V}_{\text{elem}}$ and $\mathcal{V}_{\text{pos}}$, the query evaluation problem is hard, we move in the rest of this section to provide good approximations to CQEval$[\Sigma, \mathcal{V}, Q]$.

Approximation of optimal solutions. Recall that Σ is a schema, Q is a CQ over Σ, and $\mathcal{V}$ is a volume assignment over $\mathbb{T}_\Sigma$. We say that $S^* \subseteq [\![Q]\!](D)$ with $|S^*| = k$ is an $(1 - \epsilon)$-*approximation* of CQEval$[\Sigma, \mathcal{V}, Q]$ on a database D and a number $k \geq 1$ if, and only if:

$$\delta_{\mathcal{V}}(S^*) \;\geq\; (1 - \epsilon) \cdot \max_{S \subseteq [\![Q]\!](D) : |S| = k} \delta_{\mathcal{V}}(S)$$

In other words, the diversity of S^* with respect to $\delta_{\mathcal{V}}$ is not worse than $(1 - \epsilon)$ times the diversity of the best solution, where the smaller $\epsilon \geq 0$, the better the approximation.

Since CQEval$[\Sigma, \mathcal{V}, Q]$ is NP-hard, we strive to find an $(1 - \epsilon)$-approximation for some $\epsilon \geq 0$. Given that $\delta_{\mathcal{V}}$ is monotone and submodular by Proposition 3.7, we can take advantage of the algorithmic theory of submodular set functions to find the following approximation [22].

Theorem 6.2. *One can compute an $(1 - 1/e)$-approximation of CQEval$[\Sigma, \mathcal{V}, Q]$ for every database D and $k \geq 1$ in polynomial time in $|D|$, where e is the Euler number.*

[2]We are only making statements on tractability in this section. Section 7 then focuses on finer analysis and does not make this assumption.

Algorithm 1: Greedy algorithm for finding a $(1 - 1/e)$-approximation of the problem CQEval$[\Sigma, \mathcal{V}, Q]$ for a schema Σ, a volume assignment $\mathcal{V}$, and a CQ Q over Σ.

Input: A database D and a value $k \geq 1$.
Output: A k-diversity set $S \subseteq [\![Q]\!](D)$ with respect to $\delta_{\mathcal{V}}$.

```
S ← ∅
for i = 1 to k do
    t* ← arg max_{t ∈ [[Q]](D)} δ_V(S ∪ {t})
    S ← S ∪ {t*}
return S
```

Proof. In [22], Nemhauser, Wolsey, and Fisher showed that for every monotone submodular set function $f : \text{finite}(\mathcal{U}) \to \mathbb{R}$ and $k \geq 1$ one can compute in polynomial time a k-subset A of $\mathcal{U}$ such that $f(A) \geq (1 - 1/e) \cdot \max_{B \subseteq \mathcal{U} : |B| = k} f(B)$. Since $\delta_{\mathcal{V}}$ is submodular and monotone and Q is fixed, one can compute the set $[\![Q]\!](D)$ in polynomial time over D and then apply the result in [22] to retrieve a $(1 - 1/e)$-approximation of $\delta_{\mathcal{V}}$ over $[\![Q]\!](D)$ restricted to subsets of size k. In Algorithm 1, we depict this procedure for $\delta_{\mathcal{V}}$ which follows a greedy strategy: starting from $S = \emptyset$; in every iteration it finds a tuple $t \in [\![Q]\!](D)$ that maximizes the *marginal diversity* of $\delta_{\mathcal{V}}$, namely, $\delta_{\mathcal{V}}(S \cup \{t\}) \setminus \delta_{\mathcal{V}}(S)$. After the k-th iteration, it outputs S. By [22], this procedure achieves a $(1 - 1/e)$-approximation of CQEval$[\Sigma, \mathcal{V}, Q]$ for every database D and $k \geq 1$, and runs in polynomial time. □

The previous result is indeed a direct consequence of Nemhauser et al. techniques on the maximization of submodular set functions. Nevertheless, one must compare the approximation ratio obtained for volume-based diversity functions with that of the best approximation found for distance-based analogs. Recently, approximation algorithms were proposed in [1] for CQ evaluation under distance-based diversity functions. For $\delta_{\min}$ under the Hamming or Euclidean metrics, the best approximation ratio is $(1 - (1/2 + \epsilon))$, and the running time of the algorithms depends on ϵ. For δ_{sum}, the best approximation ratio is $(1 - 2/k)$ for Hamming distance (for Euclidean distance it is $(1 - 1/2)$) but the running time depends on k^3. Instead, the approximation ratio for volume-based diversity functions is $(1 - 1/e)$ and works *for every volume assignment* that can be computed in polynomial time (in particular, for most of the examples presented in Section 3). Furthermore, Algorithm 1 can be easily incorporated into the current query evaluation strategy of any database management system by finding all tuples in $[\![Q]\!](D)$ and then applying Algorithm 1.

We want to end this section by showing that, in general, $(1 - 1/e)$-approximation is the best one can get for volume-based diversity functions.

Theorem 6.3. *There exists a schema Σ, a volume assignment $\mathcal{V}$, and a CQ Q such that a $(1 - 1/e)$-approximation of CQEval$[\Sigma, \mathcal{V}, Q]$ is the best that one can get in polynomial time data complexity, unless P = NP.*

Proof sketch. The proof is by encoding the maximum coverage problem into a volume assignment $\mathcal{V}$. It is well-known that the maximum coverage problem is hard to approximate beyond $(1 - 1/e)$-approximation ratio, unless P = NP [10]. □

7 Approximating Volume-based Diverse Answers Under Combined Complexity

In the following, we aim to lift the results of Section 6 to the combined complexity case and provide a finer analysis. Note, in this section, we include the time required to compute $\delta_{\mathcal{V}}(S)$ in our analysis. We start by stating the main problem and recalling some standard notation for

efficient CQ evaluation. Then, we present the main approach for efficient CQ evaluation under volume-based diversity functions and apply it to some specific volume assignments. We conclude by demonstrating how to generalize the technique by connecting it to the ranked enumeration problem of CQ evaluation.

Problem statement and main definitions. In this section, we aim to solve the following problem:

Problem:	CQEval$[\Sigma, \mathcal{V}]$
Input:	A database D and a CQ Q over Σ, and $k \geq 1$
Output:	$\arg\max_{S \subseteq [\![Q]\!](D) : \lvert S\rvert = k} \delta_{\mathcal{V}}(S)$

where Σ and $\mathcal{V}$ are a fixed schema and a fixed volume assignment. Contrary to Section 6, we cannot afford to find a $(1 - 1/e)$-approximation by first computing $[\![Q]\!](D)$ (whose size is $O(|D|^{|Q|})$) and then applying Algorithm 1. In other words, the set $[\![Q]\!](D)$ is compactly represented by (Q, D), and the challenge is to find the most diverse k-subset or an approximation without computing $[\![Q]\!](D)$.

Recall that even determining the existence of answers to CQs is NP-hard in combined complexity [7]. Thus, we will restrict ourselves to CQs with bounded *fractional hypertree width* (fhw) [11]. To that end, we briefly recall the notions of tree decompositions and fhw.

Let $Q(\bar{x}) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$ be a CQ using variables in $\mathcal{X}$. For the sake of simplification, in the sequel, we assume that every sequence $\bar{x}_i$ does not repeat variables and, thus, by slight abuse of notation, we may treat $\bar{x}_i$ as a set (otherwise, one can remove duplicate variables by rewriting Q and preprocessing D in linear time w.r.t. $|D|$). A *tree decomposition* of Q is a tuple (T, χ) where $T = (V(T), E(T))$ is a rooted tree and $\chi : V(T) \mapsto 2^{\mathcal{X}}$ assigns to each $v \in V(T)$ a subset $\chi(v) \subseteq \mathcal{X}$ called a *bag*. Additionally, the following properties have to be satisfied:

(1) for every variable $x \in \mathcal{X}$, the set $\{v \in V(T) \mid x \in \chi(v)\}$ induces a connected subtree of T; and
(2) for every relation $R_i(\bar{x}_i)$, there exists $v \in V(T)$ that contains all of $\bar{x}_i$ in its bag $\chi(v)$.

The fractional hypertree width of a tree decomposition (T, χ) is $\max_{v \in V(T)} \rho^*(\chi(v))$ where $\rho^*(\chi(v))$ is the minimum fractional edge cover of the hypergraph induced by $\chi(v)$ over $Q(\bar{x})$. The *fractional hypertree width* $\mathrm{fhw}(Q)$ *of* Q is the minimum fractional hypertree width among all tree decompositions of Q. Finally, a conjunctive query is called an *acyclic CQ* (ACQ) iff $\mathrm{fhw}(Q) = 1$.

Approximation through maximizing the marginal diversity. Motivated by Theorem 6.2 and Algorithm 1, a reasonable strategy to find an approximation for CQEval$[\Sigma, \mathcal{V}]$ is to compute the next tuple t that maximizes the marginal diversity of $\delta_{\mathcal{V}}(S)$. In other words, we have to consider the problem of computing greedily the next best solution (see line 3 in Algorithm 1):

Problem:	CQNext$[\Sigma, \mathcal{V}]$
Input:	A database D and a CQ Q over Σ, and a subset $S \subseteq [\![Q]\!](D)$
Output:	$\arg\max_{t \in [\![Q]\!](D)} \delta_{\mathcal{V}}(S \cup \{t\})$

Similar to CQEval$[\Sigma, \mathcal{V}]$, the main challenge is to compute t from D, Q, and S, without necessarily computing $[\![Q]\!](D)$. Naturally, if we can solve CQNext$[\Sigma, \mathcal{V}]$ efficiently, then we can apply Algorithm 1 by calling CQNext$[\Sigma, \mathcal{V}]$ in line 3 and solve CQEval$[\Sigma, \mathcal{V}]$. In other words, we get the following result.

Theorem 7.1. *If CQNext$[\Sigma, \mathcal{V}]$ can be solved in time $O(f)$ for some function f, then the problem CQEval$[\Sigma, \mathcal{V}]$ can be $(1 - 1/e)$-approximated in time $O(k \cdot f)$.*

The converse of Theorem 7.1 does not necessarily hold. In particular, we do not know whether hardness of CQNext$[\Sigma, \mathcal{V}]$ implies that CQEval$[\Sigma, \mathcal{V}]$ cannot be approximated (see Section 8 for

further ideas). However, at least, if CQNext$[\Sigma, \mathcal{V}]$ is NP-hard for the singleton case (fixing $S = \varnothing$), also the problem (exact version) CQEval$[\Sigma, \mathcal{V}]$ must be NP-hard (for $k = 1$).

We now revisit the volume assignments from Section 3 and separate the hard and easy cases for solving CQNext$[\Sigma, \mathcal{V}]$. We start with the hard cases which, thus, do not translate to approximability results of CQEval$[\Sigma, \mathcal{V}]$:

Theorem 7.2. *Unless* P = NP, *the problem CQNext*$[\Sigma, \mathcal{V}]$ *cannot be solved in polynomial time for* $\mathcal{V} \in \{\mathcal{V}_{\text{elem}}, \mathcal{V}^w_{\text{elem}}, \mathcal{V}_{Q,D}\}$, *even if we only allow ACQs and subsets* $S = \varnothing$.

Proof sketch. We illustrate the basic idea by proving NP-hardness of the apparently simplest case $\mathcal{V} = \mathcal{V}_{\text{elem}}$. The proof is by reduction from (the directed version of) Hamiltonian path: Given an instance $G = (V(G), E(G))$ of Hamiltonian path, we define an instance (D, Q, S) of CQNext$[\Sigma, \mathcal{V}]$ as follows: database D consists of a single binary relation E storing the edges of G, we set $S = \varnothing$, and, for $n = |V(G)|$, we define the ACQ Q as follows:

$$Q(x_1, \ldots, x_n) \leftarrow E(x_1, x_2), \ldots, E(x_{n-1}, x_n).$$

Notice that a solution $Q(h(\bar{x})) \in [\![Q]\!](D)$ corresponds to a walk in G and $\delta_{\mathcal{V}_{\text{elem}}}$ applied to singletons (i.e., $\delta_{\mathcal{V}_{\text{elem}}}(\{Q(h(\bar{x}))\})$) counts the number of distinct vertices used in the corresponding walk. Hence, G is a positive instance of Hamiltonian path if, and only if, the solution to this instance of CQNext$[\Sigma, \mathcal{V}_{\text{elem}}]$ yields an answer $Q(h(\bar{x}))$ with $\delta_{\mathcal{V}_{\text{elem}}}(\{Q(h(\bar{x}))\}) = n$. □

Next, we show that even seemingly simple changes in the diversity function can affect the tractability of CQNext$[\Sigma, \mathcal{V}]$ and, thus, naturally lead to the approximability of CQEval$[\Sigma, \mathcal{V}]$ due to Theorem 7.1.

Theorem 7.3. *Restricted to ACQs, the problem CQNext*$[\Sigma, \mathcal{V}]$ *can be solved in time* $O(|Q| \cdot |D|)$ *for* $\mathcal{V} \in \{\mathcal{V}_{\text{pos}}, \mathcal{V}^w_{\text{pos}}\}$ *when only allowing ACQs. Hence, in this case, CQEval*$[\Sigma, \mathcal{V}_{\text{elem}}]$ *can be* $(1 - 1/e)$*-approximated in time* $O(k \cdot |Q| \cdot |D|)$.

Proof sketch. We explain why CQNext$[\Sigma, \mathcal{V}_{\text{pos}}]$ is tractable for ACQs $Q(\bar{x})$. To that end, let D be a database, and $h_1, \ldots h_k$ homomorphisms from Q to D, i.e., $S = \{Q(h_1(\bar{x})), \ldots, Q(h_k(\bar{x}))\} \subseteq [\![Q]\!](D)$.Consider the *marginal* diversity for a new solution $Q(h(\bar{x})) \in [\![Q]\!](D)$:

$$\delta_{\mathcal{V}_{\text{pos}}}(S \cup \{Q(h(\bar{x}))\}) - \delta_{\mathcal{V}_{\text{pos}}}(S) = \sum_{x \in \bar{x}} \alpha_x, \quad \text{with } \alpha_x = \begin{cases} 1 & \text{if } \forall i\colon h(x) \neq h_i(x), \\ 0 & \text{if } \exists i\colon h(x) = h_i(x). \end{cases}$$

That is, we count the number of *new* values. We can cast this then as a sum-product query over the tropical semi-ring $\mathbb{R}_{\max} := (\mathbb{R} \cup \{\infty\}, +, \max)$. Doing so shows that we can find the element that maximizes the marginal diversity in linear time. To do so, for every $x \in \bar{x}$ let us choose a *covering* relation $R^x := R_i(\bar{x}_i)$ used in Q where $x \in \bar{x}_i$. Then, we can define the $\mathbb{R}_{\max}$-relations $R^*_1, \ldots, R^*_m$. That is, for tuple $R_j(\bar{a})$ in the database, we add tuples $R^*_j(\bar{a})$ to the database and annotate it with the number of *new* values $\bar{a}$ adds at positions x such that R_j covers x. Then, as every variable $x \in \bar{x}$ is covered by exactly one relation, we have:

$$\delta_{\mathcal{V}_{\text{pos}}}(S \cup \{Q(h(\bar{x}))\}) - \delta_{\mathcal{V}_{\text{pos}}}(S) = \sum_i R^*_i(h(\bar{x}_i)) \tag{1}$$

for homomorphism h such that $Q(h(\bar{x})) \in [\![Q]\!](D)$. Then, due to results on sum-product queries [16, 24] we can find a $Q(h(\bar{x})) \in [\![Q]\!](D)$ maximizing Equation (1) in time $O(|Q| \cdot |D|)$ as this is then a scalar sum-product query. This solves CQNext$[\Sigma, \mathcal{V}_{\text{elem}}]$ for ACQs. Then, due to Theorem 7.1, we can compute a $(1 - 1/e)$-approximation of CQEval$[\Sigma, \mathcal{V}_{\text{elem}}]$ in time $O(k \cdot |Q| \cdot |D|)$. □

Diverse answers to CQs via ranked enumeration. Towards a more general criterion to ensure tractability of CQNext$[\Sigma, \mathcal{V}]$, we consider this problem as a top-k ranked enumeration problem, where the marginal diversity is the value by which we order the output and where we ask for the top-1 answer (we can ignore the additive constant $\delta_{\mathcal{V}}(S)$). Actually, top-k ranked enumeration has received considerable attention from the database community in the last years (see e.g., [9, 13, 18, 19]), where we consider [9] as the most general and most naturally extendable to our setting.

We briefly recall the setting and main result of [9] and then build on them. There, rank functions $\mathtt{rank}$ assign values $\mathtt{rank}(Q(h(\bar{x}))) \in \mathbb{R}$ to solutions of CQs $Q(h(\bar{x})) \in [\![Q]\!](D)$ and the goal is to enumerate $Q(h(\bar{x})) \in [\![Q]\!](D)$ in the order induced by $\mathtt{rank}$, i.e., $Q(h(\bar{x}))$ should be output before $Q(h'(\bar{x}))$ if $\mathtt{rank}(Q(h(\bar{x}))) > \mathtt{rank}(Q(h(\bar{x})))$. Informally speaking, the main result of [9] is that, with the help of a tree decomposition (T, χ) of the full CQ Q, enumeration is efficiently possible if $\mathtt{rank}$ is *compatible* with (T, χ).

Given a volume assignment $\mathcal{V} = (\mathcal{S}, \mu, \beta)$, we would like to apply the results of [9] to the functions $\mathtt{rank}_{\mathcal{V},S} := \delta_{\mathcal{V}}(S \cup \{\cdot\})$. Thus, naively, we would have to verify that $\mathtt{rank}_{\mathcal{V},S}$ is compatible with (T, χ) for every $S \subseteq [\![Q]\!](D)$. Inspired by their use of compatibility, in the remainder of this section, we develop a notion of compatibility (with a tree decomposition (T, χ)) of the ball function β. This will be a sufficient condition, such that $\mathtt{rank}_{\mathcal{V},S}$ is compatible with (T, χ) for every $S \subseteq [\![Q]\!](D)$. To that end, we start as in [9] by defining what it means (in our case for β) to be $\overline{y}$-decomposable.

Definition 7.4. Let $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ be a volume assignment, $R(\bar{x})$ be an atom over Σ with variables $\bar{x}$, and $\overline{y} \subseteq \overline{x}$. We say that β is *$\overline{y}$-decomposable* (w.r.t. R) if for every pair of homomorphisms h, h' over $\overline{y}$ and homomorphisms g, g' over $\overline{x} \setminus \overline{y}$ we have:

$$\beta(R((h \cup g)(\overline{x}))) \setminus \beta(R((h' \cup g)(\overline{x}))) = \beta(R((h \cup g')(\overline{x}))) \setminus \beta(R((h' \cup g')(\overline{x}))). \tag{2}$$

The intuition of $\overline{y}$-decompositions is the following: Whatever a partial homomorphism h on $\overline{y}$ contributes to the volume compared with another partial homomorphism h' should not depend on how h and h' are completed (i.e., either by g or g'). Let us denote the set in Equation (2) as $\beta(h, h')$.

For a set S of R-tuples, let us now consider the function $\mathtt{rank}_{\mathcal{V},S}$ defined for R-tuples. Then, to compare the function value of $\mathtt{rank}_{\mathcal{V},S}$ on two homomorphisms $\hat{h}$ and $\hat{h}$ that agree outside of $\bar{y}$, it suffices to compare $\mu(\beta(h, h') \setminus \bigcup_{s \in S} \beta(s))$ with $\mu(\beta(h', h) \setminus \bigcup_{s \in S} \beta(s))$. Consequently, the function $\mathtt{rank}_{\mathcal{V},S}$ is $\overline{y}$-decomposable in the sense of [9] for every set S

Thus, to extend the main result of [9] to our setting, we can extend our notion of decomposability to compatibility w.r.t. a tree decomposition analogously to how it is done there. We note that while Definition 7.4 significantly differs from the counterpart in [9], extending it to compatibility is rather immediate. Thus, we only give the following definitions for the sake of completeness.

Let $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ be a volume assignment, let $R(\bar{x})$ be an atom over Σ with variables $\bar{x}$, and let $\overline{y}, \overline{z} \subseteq \overline{x}$ be such that $\overline{y} \cap \overline{z} = \varnothing$. Further, let $R_{\overline{x} \setminus \overline{z}} \notin \Sigma$ be a new relation symbol of arity $|\overline{x} \setminus \overline{z}|$. We say that β is *$\overline{y}$-decomposable conditioned on $\overline{z}$* (w.r.t. R) if for every homomorphism f over $\overline{z}$, the ball function extended to $R_{\overline{x} \setminus \overline{z}}$-tuples via $\beta(R_{\overline{x} \setminus \overline{z}}(\hat{h}(\overline{x} \setminus \overline{z}))) := \beta(R((\hat{h} \cup f)(\overline{x})))$ for homomorphism $\hat{h}$ over $\overline{x} \setminus \overline{z}$ is $\overline{y}$-decomposable w.r.t $R_{\overline{x} \setminus \overline{z}}$.

Let (T, χ) be a rooted tree decomposition of a full CQ $Q(\bar{x})$. For $t \in V(T)$ we denote with $\chi(T_t)$ the union of the bags in the subtree rooted in t. Further, with $\mathtt{key}(t)$ we denote the variables $\chi(t) \cap \chi(p)$ where p is the parent of t and $\mathtt{key}(r) = \varnothing$ for the root r of T. We say that β is *compatible with* (T, χ) if for every node t it is $(\chi(T_t) \setminus \mathtt{key}(t))$-decomposable conditioned on $\mathtt{key}(t)$ w.r.t. Q.

As explained before, since $\overline{y}$-decomposability in our sense can be reduced to $\overline{y}$-decomposability for every set S in the sense of [9], we get the following by combining it with Theorem 7.1.

Theorem 7.5. *Let $\mathcal{V} = (\mathcal{S}, \mu, \beta)$ be a volume assignment over $\mathbb{T}_\Sigma$ such that β is compatible with a rooted tree decomposition (T, χ) of the full CQ $Q(\bar{x})$. Then, CQEval$[\Sigma, \mathcal{V}]$ can be $(1-1/e)$-approximated in time $O(|Q| \cdot |D|^{fhw(T,\chi)} \cdot k \cdot T_{\mathcal{V}})$ where $T_{\mathcal{V}}$ is the time to compute marginals of $\delta_{\mathcal{V}}$ for fixed sets.*

To showcase Theorem 7.5, we revisit the volume assignment $\mathcal{V}_{Q,D}$ from Example 3.4.

Theorem 7.6. *Let $Q(\bar{x})$ be a CQ such that every atom $R_i(\bar{x}_i)$ of Q uses a unique relation name and let (T, χ) be a tree decomposition of Q such that there is a subtree $T_{\bar{x}}$ of T containing the root of T and where $\bar{x} = \bigcup_{v \in V(T_{\bar{x}})} \chi(v)$. That is, the CQ is self-join-free and the tree decomposition is free-connex [5]. Then CQEval$[\Sigma, \mathcal{V}_{Q,D}]$ can be $(1 - 1/e)$-approximated in time $O(|Q| \cdot |D|^{fhw(T,\chi)+1} \cdot k)$.*

We juxtapose it with Theorem 7.2: In Theorem 7.2 we say that CQNext$[\Sigma, \mathcal{V}_{Q,D}]$ is intractable even for ACQs while we now state that computing a $(1 - 1/e)$-approximation of CQEval$[\Sigma, \mathcal{V}_{Q,D}]$ is tractable for CQs when $fhw(T, \chi)$ is small. The crucial restriction in Theorem 7.6 is *self-join-freeness*, which is in effect similar to keeping positions apart as $\mathcal{V}_{\text{pos}}$ does compared to $\mathcal{V}_{\text{elem}}$.

Proof Sketch of Theorem 7.6. Theorem 7.5 cannot directly be applied since Q is not necessarily a full CQ. To that end, let us consider the full CQ $Q^{\bar{x}}(\bar{x})$ defined as the subquery of Q where all body relations are projected onto $\bar{x}$. Then, $(T_{\bar{x}}, \chi|_{V(T_{\bar{x}})})$ is a tree decomposition of $Q^{\bar{x}}$ and $fhw(T_{\bar{x}}, \chi|_{V(T_{\bar{x}})}) \leq fhw(T, \chi)$. Now, we extend $\beta_{Q,D}$ to $Q^{\bar{x}}$-tuples via $\beta_{Q,D}(Q^{\bar{x}}(h(\bar{x}))) := \beta_{Q,D}(Q(h(\bar{x})))$. Defined as such, $\beta_{Q,D}$ is compatible with $(T_{\bar{x}}, \chi|_{V(T_{\bar{x}})})$ w.r.t. $Q^{\bar{x}}$ as Q and, hence, also $Q^{\bar{x}}$ are self-join-free.

Then, to compute $\mathsf{rank}_{\mathcal{V}_{Q,D},S}$, we have to keep track of the which-provenance [8] for each of the tuples in the bags of $v \in V(T_{\bar{x}})$ for what happens "outside" of $V(T_{\bar{x}})$. Thus, essentially, for each $v \in V(T_{\bar{x}})$, we have to look at its children in $T \setminus T_{\bar{x}}$, i.e., $C := child(v) \setminus V(T_{\bar{x}})$ and consider the sub-query $Q^v(\chi(v))$ that uses the variables $\chi(v)$, and the ones that appear in C and their descendants. Computing the provenance of these queries requires time $|D|^{fhw(T,\chi)+1}$ (where the $+1$ is to account for the semi-ring operations) [16, 24]. However, then, to compute marginals of $\delta_{\mathcal{V}}$ (essentially $\delta_{\mathcal{V}}(S \cup \{s\})$), it suffices to add together the provenance of every tuple $t \in S \cup \{s\}$. The provenance of a tuple t can be computed by looking-up and adding together the provenance of t projected to $\chi(v)$ in Q^v. Thus, as S can be considered fixed, this takes $O(|D|)$ time. □

In particular, this means that for self-join-free, free-connex, acyclic conjunctive queries, the problem CQEval$[\Sigma, \mathcal{V}_{Q,D}]$ can be $(1 - 1/e)$-approximated in quadratic time (for constant Q, k).

8 Conclusions

In this work, we have introduced the volume-based framework for diversity measures $\delta_{\mathcal{V}}$, providing several examples of them in relational databases, and we have studied their properties. Above all, given the intractability of query answering under diversity, we have shown an approximation algorithm that runs in polynomial time data complexity, and we have identified criteria for extending the tractability of the approximation to combined complexity. Arguably, all these results provide substantial evidence that volume-based diversity forms an alternative approach to distance-based diversity, which requires further consideration in both the theory and practice of database management systems.

For future work, we propose to take a closer look into the relationship between our framework of volume-based diversity measures and the distance-based approach. In Section 5, we have shown that Weitzman's (distance-based) diversity function δ_{W} essentially becomes a volume-based diversity function if the underlying distance function is an ultrametric. In [2], general criteria were presented that make the problem of computing the exact solution of CQEval tractable if

the distance underlying a diversity measure is an ultrametric. It would be interesting to explore restrictions under which this can be lifted to volume-based diversity functions.

Another interesting open problem is to find other strategies for approximating CQEval in combined complexity. According to Theorem 7.2, CQNext cannot be solved in polynomial time (under complexity assumptions) for many natural volume-based diversity functions. Nevertheless, even in the cases where CQNext is NP-hard, one might still be able to get a reasonable approximation algorithm. This approximation algorithm CQNext, combined with Algorithm 1, could then lead to an approximation of CQEval.

Acknowledgements

The work of Merkl and Pichler was supported by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT2201, 10.47379/VRG18013, 10.47379/NXT22018]. The work of Arenas and Riveros was supported by ANID – Millennium Science Initiative Program – Code ICN17_002. Riveros was also supported by ANID Fondecyt Regular project 1230935.

References

[1] P. K. Agarwal, A. Esmailpour, X. Hu, S. Sintos, and J. Yang. Computing A well-representative summary of conjunctive query results. *Proc. ACM Manag. Data*, 2(5):217:1–217:27, 2024.

[2] M. Arenas, T. C. Merkl, R. Pichler, and C. Riveros. Towards tractability of the diversity of query answers: Ultrametrics to the rescue. *Proc. ACM Manag. Data*, 2(5):215:1–215:26, 2024.

[3] M. Arenas, T. C. Merkl, R. Pichler, and C. Riveros. Query answering under volume-based diversity functions. *CoRR*, abs/2509.11929, 2025.

[4] S. Axler. *Measure, integration & real analysis*. Springer Nature, 2020.

[5] G. Bagan, A. Durand, and E. Grandjean. On acyclic conjunctive queries and constant delay enumeration. In J. Duparc and T. A. Henzinger, editors, *Computer Science Logic, 21st International Workshop, CSL 2007, 16th Annual Conference of the EACSL, Lausanne, Switzerland, September 11-15, 2007, Proceedings*, volume 4646 of *Lecture Notes in Computer Science*, pages 208–222. Springer, 2007.

[6] J. Baste, M. R. Fellows, L. Jaffke, T. Masarík, M. de Oliveira Oliveira, G. Philip, and F. A. Rosamond. Diversity of solutions: An exploration through the lens of fixed-parameter tractability theory. *Artif. Intell.*, 303:103644, 2022.

[7] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In J. E. Hopcroft, E. P. Friedman, and M. A. Harrison, editors, *Proceedings of the 9th Annual ACM Symposium on Theory of Computing, May 4-6, 1977, Boulder, Colorado, USA*, pages 77–90. ACM, 1977.

[8] Y. Cui and J. Widom. Lineage tracing for general data warehouse transformations. *VLDB J.*, 12(1):41–58, 2003.

[9] S. Deep and P. Koutris. Ranked enumeration of conjunctive query results. *Logical Methods in Computer Science*, 21, 2025.

[10] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[11] M. Grohe and D. Marx. Constraint solving via fractional edge covers. *ACM Trans. Algorithms*, 11(1):4:1–4:20, 2014.

[12] E. Hebrard, B. Hnich, B. O'Sullivan, and T. Walsh. Finding diverse and similar solutions in constraint programming. In M. M. Veloso and S. Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 372–377. AAAI Press / The MIT Press, 2005.

[13] I. F. Ilyas, R. Shah, W. G. Aref, J. S. Vitter, and A. K. Elmagarmid. Rank-aware query optimization. In G. Weikum, A. C. König, and S. Deßloch, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, pages 203–214. ACM, 2004.

[14] L. Ingmar, M. G. de la Banda, P. J. Stuckey, and G. Tack. Modelling diversity of solutions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1528–1535. AAAI Press, 2020.

[15] M. M. Islam, M. Asadi, S. Amer-Yahia, and S. B. Roy. A generic framework for efficient computation of top-k diverse results. *VLDB J.*, 32(4):737–761, 2023.

[16] M. A. Khamis, H. Q. Ngo, and A. Rudra. FAQ: questions asked frequently. In T. Milo and W. Tan, editors, *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 13–28. ACM, 2016.

[17] P. G. Kolaitis and M. Y. Vardi. Conjunctive-query containment and constraint satisfaction. *J. Comput. Syst. Sci.*, 61(2):302–332, 2000.
[18] C. Li, K. C. Chang, I. F. Ilyas, and S. Song. Ranksql: Query algebra and optimization for relational top-k queries. In F. Özcan, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, pages 131–142. ACM, 2005.
[19] C. Li, M. A. Soliman, K. C. Chang, and I. F. Ilyas. Ranksql: Supporting ranking queries in relational database management systems. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P. Larson, and B. C. Ooi, editors, *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005*, pages 1342–1345. ACM, 2005.
[20] T. C. Merkl, R. Pichler, and S. Skritek. Diversity of answers to conjunctive queries. In F. Geerts and B. Vandevoort, editors, *26th International Conference on Database Theory, ICDT 2023, March 28-31, 2023, Ioannina, Greece*, volume 255 of *LIPIcs*, pages 10:1–10:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.
[21] K. Nehring and C. Puppe. A theory of diversity. *Econometrica*, 70(3):1155–1198, 2002.
[22] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.
[23] S. Nikookar, M. Esfandiari, R. M. Borromeo, P. Sakharkar, S. Amer-Yahia, and S. B. Roy. Diversifying recommendations on sequences of sets. *VLDB J.*, 32(2):283–304, 2023.
[24] R. Pichler and S. Skritek. Tractable counting of the answers to conjunctive queries. *J. Comput. Syst. Sci.*, 79(6):984–1001, 2013.
[25] E. Vee, J. Shanmugasundaram, and S. Amer-Yahia. Efficient computation of diverse query results. *IEEE Data Eng. Bull.*, 32(4):57–64, 2009.
[26] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. Amer-Yahia. Efficient computation of diverse query results. In G. Alonso, J. A. Blakeley, and A. L. P. Chen, editors, *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, Mexico*, pages 228–236. IEEE Computer Society, 2008.
[27] M. L. Weitzman. On diversity. *The quarterly journal of economics*, 107(2):363–405, 1992.

Received June 2025; accepted August 2025