Parameterized and Fine-Grained Analysis of Query Evaluation Over Bag PDBs

Su Feng
Illinois Institute of Technology, Chicago, USA

Boris Glavic
Illinois Institute of Technology, USA

Aaron Huber
University at Buffalo, USA

Oliver Kennedy
University at Buffalo, USA

Atri Rudra
University at Buffalo, USA

Abstract

The problem of computing the marginal probability of a tuple in the result of a query over set-probabilistic databases (PDBs) is a fundamental problem in set-PDBs. In this work, we study the analog problem for bag semantics: computing a tuple's expected multiplicity exactly and approximately. We are specifically interested in the fine-grained complexity and how it compares to the complexity of deterministic query evaluation algorithms — if these complexities are comparable, it opens the door to practical deployment of probabilistic databases. Unfortunately, our results imply that computing expected multiplicities for Bag-PDBs based on the results produced by such query evaluation algorithms introduces super-linear overhead (under parameterized complexity hardness assumptions/conjectures). We proceed to study approximation of expected multiplicities of result tuples of positive relational algebra queries ($\mathcal{RA}^+$) over c-TIDBs and for a non-trivial subclass of block-independent databases (BIDBs). We develop a sampling algorithm that computes a $(1 \pm \epsilon)$-approximation of the expected multiplicity of an output tuple in time linear in the runtime of a comparable deterministic query for any $\mathcal{RA}^+$ query.

2012 ACM Subject Classification Information systems → Incomplete data

Keywords and phrases PDB, bags, polynomial, boolean formula, etc.

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction

This work explores the problem of computing the expectation of a tuple's multiplicity in an important special case of bag TIDB, which we call a c-TIDB. A c-TIDB, $\mathcal{D} = \left(\{0, \ldots, c\}^D, \mathcal{P}\right)$ encodes a bag of uncertain tuples such that each tuple in $\mathcal{D}$ has a multiplicity of at most c. D is the set of tuples appearing across all possible worlds, and the set of all worlds is encoded in $\{0, \ldots, c\}^D$, which is the set of all vectors of length $n = |D|$ such that each index corresponds to a distinct $t \in D$ storing its multiplicity. $\mathcal{P}$ is a product distribution over the set of all worlds. A given world $\mathbf{M} \in \{0, \ldots, c\}^D$ can be interpreted such that, for each $t \in D$, $\mathbf{M}[t]$ is the multiplicity of t in $\mathbf{M}$. The resulting product distribution can then be encoded as $p_t = Pr[W[t] = j]$ (for $j \in [c]$), where each t is an independent random event. Allowing for $\leq c$ multiplicities across all tuples gives rise to having $\leq (c+1)^n$ possible worlds instead of the usual 2^n possible worlds of a 1-TIDB, which (assuming set query semantics), is the same as the traditional set TIDB. In this work, since we are generally considering bag query input, we will only be considering bag query semantics. We denote by $Q(\mathbf{M})(t)$ the

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:62
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

multiplicity of t in query Q over possible world $\mathbf{M} \in \{0, \ldots, c\}^D$.

We can formally state our problem of computing the expected multiplicity of a result tuple as:

▶ **Problem 1.1.** *Given a* c-TIDB $\mathcal{D} = \left(\{0, \ldots, c\}^D, \mathcal{P}\right)$, $\mathcal{RA}^+$ *query* Q, *and result tuple* t, *compute the expected multiplicity of* t: $\mathbb{E}_{\mathbf{M} \sim \mathcal{P}}\left[Q\left(\mathbf{M}\right)(t)\right]$.

It is natural to explore computing the expected multiplicity of result tuple as this is the analog for computing the marginal probability of a tuple in a set PDB. In this work we will assume that $c = O(1)$ since this is what typically seen in practice. Allowing for unbounded c is an interesting open problem.

Hardness of Set Query Semantics and Bag Query Semantics. Set query evaluation semantics over 1-TIDBs have been studied extensively, and the data complexity of the problem in general has been shown by Dalvi and Suciu to be #P-hard [13]. For our setting, there exists a trivial polytime algorithm to compute Problem 1.1 for any query over a c-TIDB due to linearity of expection by simply computing the expectation over a 'sum-of-products' representation of the query operations of $Q(\mathcal{D})(t)$. Since we can compute Problem 1.1 in polynomial time, the interesting question that we explore deals with analyzing the hardness of computing expectation using fine-grained analysis and parameterized complexity, where we are interested in the exponent of polynomial runtime.

Specifically, in this work we ask if Problem 1.1 can be solved in time linear in the runtime of an equivalent deterministic query. If this is true, then this would open up the way for deployment of c-TIDBs in practice. To analyze this question we denote by $T^*(Q, \mathcal{D})$ the optimal runtime complexity of computing Problem 1.1 over c-TIDB $\mathcal{D}$.

Let $T_{det}\left(Q, \overline{D}, c\right) = Q\left(\overline{D}\right)$ for arbitrary query Q, deterministic database $\overline{D}$, and multiplicity bound c. Let $T^*_{det}(Q, \overline{D}, c) = \min_{Q' : Q' \equiv Q} T_{det}\left(Q, \overline{D}, c\right)$ be the optimal runtime (with some caveats; discussed in Sec. 2.3) of query Q on deterministic database D.

Lower bound on $T^*(Q, \mathcal{D})$	Num. $\mathcal{P}$s	Hardness Assumption
$\Omega\left(\left(T^*_{det}(Q, D, c)\right)^{1+\epsilon_0}\right)$ for *some* $\epsilon_0 > 0$	Single	Triangle Detection hypothesis
$\omega\left(\left(T^*_{det}(Q, D, c)\right)^{C_0}\right)$ for *all* $C_0 > 0$	Multiple	$\#\mathsf{W}[0] \neq \#\mathsf{W}[1]$
$\Omega\left(\left(T^*_{det}(Q, D, c)\right)^{c_0 \cdot k}\right)$ for *some* $c_0 > 0$	Multiple	Conjecture 3.2

Table 1 Our lower bounds for a specific hard query Q parameterized by k. For $\mathcal{D} = \left\{\{0, \ldots, c\}^D, \mathcal{P}\right\}$ those with 'Multiple' in the second column need the algorithm to be able to handle multiple $\mathcal{P}$s (for a given D). The last column states the hardness assumptions that imply the lower bounds in the first column (ϵ_o, C_0, c_0 are constants that are independent of k).

Our lower bound results. Our question is whether or not it is always true that $T^*(Q, \mathcal{D}) \leq T^*_{det}(Q, D, c)$. Unfortunately this is not the case. Table 1 shows our results.

Specifically, depending on what hardness result/conjecture we assume, we get various emphatic versions of *no* as an answer to our question. To make some sense of the other lower bounds in Table 1, we note that it is not too hard to show that $T^*(Q, \mathcal{D}) \leq O\left(\left(T^*_{det}(Q, D, c)\right)^k\right)$, where k is the join width (our notion of join width follows from Definition 2.2 and Fig. 1.) of the query Q over all result tuples t (and the parameter that defines our family of hard queries).

What our lower bound in the third row says is that one cannot get more than a polynomial improvement over essentially the trivial algorithm for Problem 1.1. However, this result

$$\Phi[\pi_A(Q), \overline{D}, t] = \sum_{t' : \pi_A(t') = t} \Phi[Q, \overline{D}, t'] \qquad \Phi[Q_1 \cup Q_2, \overline{D}, t] = \Phi[Q_1, \overline{D}, t] + \Phi[Q_2, \overline{D}, t]$$

$$\Phi[\sigma_\theta(Q), \overline{D}, t] = \begin{cases} \Phi[Q, \overline{D}, t] & \text{if } \theta(t) \\ 0 & \text{otherwise.} \end{cases} \qquad \Phi[Q_1 \bowtie Q_2, \overline{D}, t] = \Phi[Q_1, \overline{D}, \pi_{attr(Q_1)} t] \cdot \Phi[Q_2, \overline{D}, \pi_{attr(Q_2)} t]$$

$$\Phi[R, \overline{D}, t] = X_t$$

Figure 1 Construction of the lineage (polynomial) for an $\mathcal{RA}^+$ query Q over a arbitrary deterministic database $\overline{D}$, where $\mathbf{X}$ consists of all X_t over all R in $\overline{D}$ and t in R. Here $\overline{D}.R$ denotes the instance of relation R in $\overline{D}$. Please note, after we introduce the reduction to 1-BIDB, the base case will be expressed alternatively.

assumes a hardness conjecture that is not as well studied as those in the first two rows of the table (see Sec. 3 for more discussion on the hardness assumptions). Further, we note that existing results already imply the claimed lower bounds if we were to replace the $T^*_{det}(Q, D, c)$ by just n (indeed these results follow from known lower bound for deterministic query processing). Our contribution is to then identify a family of hard queries where deterministic query processing is 'easy' but computing the expected multiplicities is hard.

Our upper bound results. We introduce an $(1 \pm \epsilon)$-approximation algorithm that computes Problem 1.1 in time $O_\epsilon\left(T^*_{det}(Q, D, c)\right)$. This means, when we are okay with approximation, that we solve Problem 1.1 in time linear in the size of the deterministic query and bag PDBs are deployable in practice. In contrast, known approximation techniques ([38, 30]) in set-PDBs need time $\Omega(T^*_{det}(Q, D, c)^{2k})$ (see Appendix G). Further, our approximation algorithm works for a more general notion of bag PDBs beyond c-TIDBs (see Sec. 2.1.1).

1.1 Polynomial Equivalence

A common encoding of probabilistic databases (e.g., in [28, 27, 5, 2] and many others) relies on annotating tuples with lineages, propositional formulas that describe the set of possible worlds that the tuple appears in. The bag semantics analog is a provenance/lineage polynomial (see Fig. 1) $\Phi[Q, D, t]$ [25], a polynomial with non-zero integer coefficients and exponents, over integer variables $\mathbf{X}$ encoding input tuple multiplicities.

We drop Q, D, and t from $\Phi[Q, D, t]$ when they are clear from the context or irrelevant to the discussion. We now specify the problem of computing the expectation of tuple multiplicity in the language of lineage polynomials:

▶ **Problem 1.2** (Expected Multiplicity of Lineage Polynomials). *Given an $\mathcal{RA}^+$ query Q, c-TIDB $\mathcal{D}$ and result tuple t, compute the expected multiplicity of the polynomial $\Phi[Q, D, t]$ (i.e., $\mathbb{E}_{\mathbf{W} \sim \mathcal{P}}\left[\Phi[Q, D, t](\mathbf{W})\right]$, where $\mathbf{W} \in \{0, \ldots, c\}^D$).*

We note that computing Problem 1.1 is equivalent to computing Problem 1.2 (see Proposition 2.5).

All of our results rely on working with a *reduced* form of the lineage polynomial Φ. In fact, it turns out that for the 1-TIDB case, computing the expected multiplicity (over bag query semantics) is *exactly* the same as evaluating this reduced polynomial over the probabilities that define the 1-TIDB. This is also true when the query input(s) is a block independent disjoint probabilistice database (with tuple multiplicity of at most 1), which we refer to as a 1-BIDB. Next, we motivate this reduced polynomial. Consider the query Q_1 defined as follows over the bag relations of Fig. 2:

```
SELECT 1 FROM T t1, Route r, T t2
WHERE t1.city = r.city1 AND t2.city = r.city2
```

It can be verified that $\Phi(A,B,C,E,X,Y,Z)$ for the sole result tuple (i.e. the count) of Q is $AXB + BYE + BZC$. Now consider the product query $Q_1^2 = Q_1 \times Q_1$. The lineage polynomial for Q_1^2 is given by $\Phi_1^2(A,B,C,E,X,Y,Z)$

$$= A^2X^2B^2 + B^2Y^2E^2 + B^2Z^2C^2 + 2AXB^2YE + 2AXB^2ZC + 2B^2YEZC.$$

To compute $\mathbb{E}\left[\Phi_1^2\right]$ we can use linearity of expectation and push the expectation through each summand. To keep things simple, let us focus on the monomial $\Phi_1^{(ABX)^2} = A^2X^2B^2$ as the procedure is the same for all other monomials of Φ_1^2. Let W_X be the random variable corresponding to a lineage variable X. Because the distinct variables in the product are independent, we can push expectation through them yielding $\mathbb{E}\left[W_A^2W_X^2W_B^2\right] = \mathbb{E}\left[W_A^2\right]\mathbb{E}\left[W_X^2\right]\mathbb{E}\left[W_B^2\right]$. Since $W_A, W_B \in \{0,1\}$ we can further derive $\mathbb{E}[W_A]\mathbb{E}\left[W_X^2\right]\mathbb{E}[W_B]$ by the fact that for any $W \in \{0,1\}$, $W^2 = W$. However, we get stuck with $\mathbb{E}\left[W_X^2\right]$, since $W_X \in \{0,1,2\}$ and for $W_X \leftarrow 2$, $W_X^2 \neq W_X$.

Denote the variables of Φ to be $\textsc{Vars}(\Phi)$. In the c-TIDB setting, $\Phi(\mathbf{X})$ has an equivalent reformulation (Φ_R) that is of use to us. Given $X_t \in \textsc{Vars}(\Phi)$, by definition $X_t \in \{0,\ldots,c\}$. We can replace X_t by $\sum_{j\in[c]} X_{t,j}$ where each $X_{t,j} \in \{0,1\}$. Then for any $\mathbf{M} \in \{0,\ldots,c\}^D$, we set $X_{t,j} = 1$ for $\mathbf{M}_t = j$, while $X_{t,j'} = 0$ for all $j' \neq j \in [c]$. By construction then $\Phi(\mathbf{X}) \equiv \Phi_R(\mathbf{X})$ since for any $X_t \in \textsc{Vars}(\Phi)$ we have the equality $X_t = j = \sum_{j\in[c]} jX_j$.

Considering again our example,

$$\Phi_{1,R}^{(ABX)^2}(A,X,B) = \Phi^{(AXB)^2}\left(\sum_{j_1\in[c]} j_1A_{j_1}, \sum_{j_2\in[c]} j_2X_{j_2}, \sum_{j_3\in[c]} j_3B_{j_3}\right)$$
$$= \left(\sum_{j_1\in[c]} j_1A_{j_1}\right)^2\left(\sum_{j_2\in[c]} j_2X_{j_2}\right)^2\left(\sum_{j_3\in[c]} j_3B_{j_3}\right)^2.$$

Since the set of multiplicities for tuple t by nature are disjoint we can drop all cross terms and have $\Phi_{1,R}^2 = \sum_{j_1,j_2,j_3\in[c]} j_1^2A_{j_1}^2j_2^2X_{j_2}^2j_3^2B_{j_3}^2$. Computing expectation we get $\mathbb{E}\left[\Phi_{1,R}^2\right] = \sum_{j_1,j_2,j_3\in[c]} j_1^2j_2^2j_3^2\,\mathbb{E}\left[W_{A_{j_1}}\right]\mathbb{E}\left[W_{X_{j_2}}\right]\mathbb{E}\left[W_{B_{j_3}}\right]$, since we now have that all $W_{X_j} \in \{0,1\}$. This leads us to consider a structure related to the lineage polynomial.

▶ **Definition 1.3.** *For any polynomial $\Phi\left((X_t)_{t\in D}\right)$ define the reformulated polynomial $\Phi_R\left((X_{t,j})_{t\in D,j\in[c]}\right)$ to be the polynomial $\Phi_R = \Phi\left(\left(\sum_{j\in[c]} j\cdot X_{t,j}\right)_{t\in D}\right)$ and ii) define the* reduced polynomial $\widetilde{\Phi}\left((X_{t,j})_{t\in D,j\in[c]}\right)$ *to be the polynomial resulting from converting Φ_R into the standard monomial basis (SMB), [1] removing all monomials containing the term $X_{t,j}X_{t,j'}$ for $t \in D, j \neq j' \in [c]$, and setting all* variable *exponents $e > 1$ to 1.*

Continuing with the example $\Phi_1^2(A,B,C,E,X_1,X_2,Y,Z)$, to save clutter we i) do not show the full expansion for variables with greatest multiplicity $= 1$ since e.g. for variable A, the sum of products itself evaluates to $1^2 \cdot A^2 = A$, and ii) for $\sum_{j\in[c]} j^2 \cdot X_j$, we omit the

[1] This is the representation, typically used in set-PDBs, where the polynomial is reresented as sum of 'pure' products. See Definition 2.1 for a formal definition.

summands encoding multiplicities > 2, since the greatest multiplicity of the tuple annotated with X is 2, likewise those summands will always evaluated to 0 since the tuple will never have a multiplicity of > 2.

$$\widetilde{\Phi}_1^2(A,B,C,E,X_1,X_2,Y,Z) =$$

$$A\left(\sum_{j\in[c]} j^2X_j\right)B{+}BYE{+}BZC{+}2A\left(\sum_{j\in[c]} j^2X_j\right)BYE{+}2A\left(\sum_{j\in[c]} j^2X_j\right)BZC{+}2BYEZC =$$

$$ABX_1{+}AB\,(2)^2\,X_2{+}BYE{+}BZC{+}2AX_1BYE{+}2A\,(2)^2\,X_2BYE{+}2AX_1BZC{+}2A\,(2)^2\,X_2BZC{+}2BYEZC.$$

Note that we have argued that for our specific example the expectation that we want is $\widetilde{\Phi}_1^2(Pr\,(A=1)\,, Pr\,(B=1)\,, Pr\,(C=1)), Pr\,(E=1)\,, Pr\,(X_1=1)\,, Pr\,(X_2=1)\,, Pr\,(Y=1)\,, Pr\,(Z=1))$. Lemma 1.4 generalizes the equivalence to *all* $\mathcal{RA}^+$ queries on c-TIDBs (proof in Appendix B.5).

▶ **Lemma 1.4.** *For any c-TIDB $\mathcal{D}$, $\mathcal{RA}^+$ query Q, and lineage polynomial $\Phi\,(\mathbf{X}) = \Phi\,[Q,D,t]\,(\mathbf{X})$, it holds that $\mathbb{E}_{\mathbf{W}\sim\mathcal{P}}\,[\Phi_R\,(\mathbf{W})] = \widetilde{\Phi}\,(\mathbf{p})$, where $\mathbf{p} = \left((p_{t,j})_{t\in D, j\in[c]}\right)$.*

1.2 Our Techniques

Lower Bound Proof Techniques. Our main hardness result shows that computing Problem 1.1 is $\#\mathsf{W}[1]-hard$ for 1-TIDB. To prove this result we show that for the same Q_1 from the example above, for an arbitrary 'product width' k, the query Q^k is able to encode various hard graph-counting problems (assuming $O\,(n)$ tuples rather than the $O(1)$ tuples in Fig. 2). We do so by considering an arbitrary graph G (analogous to relation $\boldsymbol{R}$ of Q) and analyzing how the coefficients in the (univariate) polynomial $\widetilde{\Phi}\,(p,\ldots,p)$ relate to counts of subgraphs in G that are isomorphic to various graphs with k edges. E.g., we exploit the fact that the leading coefficient in Φ corresponding to Q^k is proportional to the number of k-matchings in G, a known hard problem in parameterized/fine-grained complexity literature.

Upper Bound Techniques. Our negative results (Table 1) indicate that c-TIDBs (even for $c = 1$) can not achieve comparable performance to deterministic databases for exact results (under complexity assumptions). In fact, under plausible hardness conjectures, one cannot (drastically) improve upon the trivial algorithm to exactly compute the expected multiplicities for 1-TIDBs. A natural followup is whether we can do better if we are willing to settle for an approximation to the expected multiplities.

We adopt the two-step intensional model of query evaluation used in set-PDBs, as illustrated in Fig. 2: (i) Lineage Computation (LC): Given input D and Q, output every tuple t that possibly satisfies Q, annotated with its lineage polynomial ($\Phi(\mathbf{X}) = \Phi[Q,D,t]\,(\mathbf{X})$); (ii) Expectation Computation (EC): Given $\Phi(\mathbf{X})$ for each tuple, compute $\mathbb{E}\,[\Phi(\mathbf{W})]$. Let $T_{LC}(Q,D,\mathtt{C})$ denote the runtime of LC when it outputs $\mathtt{C}$ (which is a representation of Φ as an arithmetic circuit — more on this representation shortly). Denote by $T_{EC}(\mathtt{C},\epsilon)$ (recall $\mathtt{C}$ is the output of LC) the runtime of EC, which we can leverage Definition 1.3 and Lemma 1.4 to address the next formal objective:

▶ **Problem 1.5** (c-TIDB linear time approximation)**.** *Given c-TIDB $\mathcal{D}$, $\mathcal{RA}^+$ query Q, is there a $(1\pm\epsilon)$-approximation of $\mathbb{E}_{\mathbf{W}\sim\mathcal{P}}\,[Q\,(\mathbf{W})\,(t)]$ for all result tuples t where $\exists\mathcal{C}$: $T_{LC}(Q,D,\mathcal{C}) + T_{EC}(\mathcal{C},\epsilon) \leq O_\epsilon(T^*_{det}(Q,D,c))$?*

We show in Appendix E.2.1 an $O(T^*_{det}(Q,D,c))$ algorithm for constructing the lineage polynomial for all result tuples of an $\mathcal{RA}^+$ query Q (or more more precisely, a single circuit

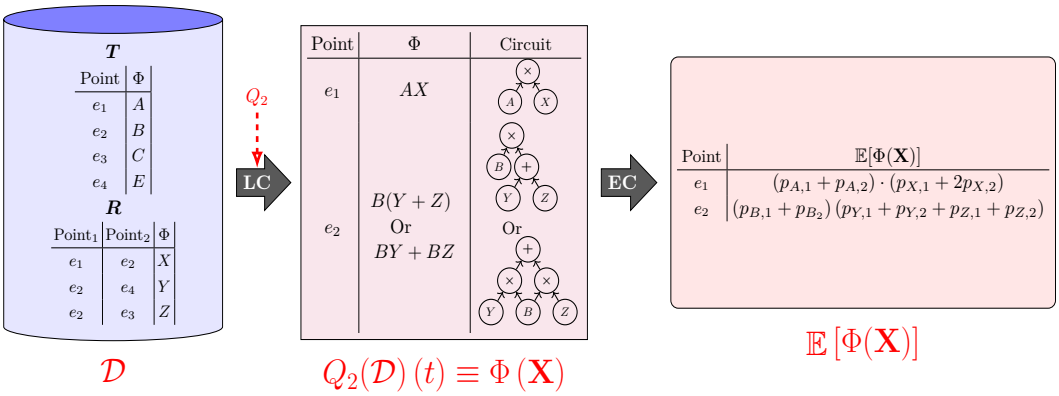

Figure 2 Intensional Query Evaluation Model ($Q_2 = \pi_{\text{Point}}\left(T \bowtie_{\text{Point}=\text{Point}_1} R\right)$ and $c = 2$).

C with one sink per tuple representing the tuple's lineage). A key insight of this paper is that the representation of C matters. For example, if we insist that C represent the lineage polynomial in SMB, the answer to the above question in general is no, since then we will need $|\mathtt{C}| \geq \Omega\left(\left(T^*_{det}(Q, D, c)\right)^k\right)$, and hence, just $T_{LC}(Q, D, \mathtt{C})$ will be too large.

However, systems can directly emit compact, factorized representations of $\Phi(\mathbf{X})$ (e.g., as a consequence of the standard projection push-down optimization [23]). For example, in Fig. 2, $B(Y + Z)$ is a factorized representation of the SMB-form $BY + BZ$. Accordingly, this work uses (arithmetic) circuits[2] as the representation system of $\Phi(\mathbf{X})$.

Given that there exists a representation $\mathtt{C}^*$ such that $T_{LC}(Q, D, \mathtt{C}^*) \leq O(T^*_{det}(Q, D, c))$, we can now focus on the complexity of EC. We can represent the factorized lineage polynomial by its correspoding arithmetic circuit C (whose size we denote by $|\mathtt{C}|$). As we also show in Appendix E.2.2, this size is also bounded by $T^*_{det}(Q, D, c)$ (i.e., $|\mathtt{C}^*| \leq O(T^*_{det}(Q, D, c))$). Thus, the question of approximation can be stated as the following stronger (since Problem 1.5 has access to *all* equivalent C representing $Q(\mathbf{W})(t)$), but sufficient condition:

▶ **Problem 1.6.** *Given one circuit $\mathcal{C}$ that encodes $\Phi[Q, D, t]$ for all result tuples t (one sink per t) for* bag-PDB $\mathcal{D}$ *and $\mathcal{RA}^+$ query Q, does there exist an algorithm that computes a $(1 \pm \epsilon)$-approximation of $\mathbb{E}_{\mathbf{W}\sim\mathcal{P}}[Q(\mathbf{W})(t)]$ (for all result tuples t) in $O(|\mathcal{C}|)$ time?*

For an upper bound on approximating the expected count, it is easy to check that if all the probabilties are constant then $\Phi(p_1, \ldots, p_n)$ (i.e. evaluating the original lineage polynomial over the probability values) is a constant factor approximation. For example, using Q^2 from above, using p_A to denote $Pr[A = 1]$ (and similarly for the other variables), we can see that

$$\Phi_1^2(\mathbf{p}) = p_A^2 p_X^2 p_B^2 + p_B^2 p_Y^2 p_E^2 + p_B^2 p_Z^2 p_C^2 + 2p_A p_X p_B^2 p_Y p_E + 2p_A p_X p_B^2 p_Z p_C + 2p_B^2 p_Y p_E p_Z p_C$$
$$\leq p_A p_X p_B + p_B p_Y p_E + p_B p_Z p_C + 2p_A p_X p_B p_Y p_E + 2p_A p_X p_B p_Z p_C + 2p_B p_Y p_E p_Z p_C = \widetilde{\Phi}_1^2(\mathbf{p})$$

If we assume that all seven probability values are at least $p_0 > 0$, we get that $\Phi_1^2(\mathbf{p})$ is in the range $[(p_0)^3 \cdot \widetilde{\Phi}_1^2(\mathbf{p}), \widetilde{\Phi}_1^2(\mathbf{p})]$. In sec. 4 we demonstrate that a $(1 \pm \epsilon)$ (multiplicative) approximation with competitive performance is achievable. To get an $(1 \pm \epsilon)$-multiplicative approximation and solve Problem 1.6, using C we uniformly sample monomials from the

[2] An arithmetic circuit is a DAG with variable and/or numeric source nodes and internal, each nodes representing either an addition or multiplication operator.

equivalent SMB representation of Φ (without materializing the SMB representation) and 'adjust' their contribution to $\widetilde{\Phi}(\cdot)$.

Applications. Recent work in heuristic data cleaning [49, 43, 40, 8, 43] emits a PDB when insufficient data exists to select the 'correct' data repair. Probabilistic data cleaning is a crucial innovation, as the alternative is to arbitrarily select one repair and 'hope' that queries receive meaningful results. Although PDB queries instead convey the trustworthiness of results [35], they are impractically slow [18, 17], even in approximation (see Appendix G). Bags, as we consider, are sufficient for production use, where bag-relational algebra is already the default for performance reasons. Our results show that bag-PDBs can be competitive, laying the groundwork for probabilistic functionality in production database engines.

Paper Organization. We present relevant background and notation in Sec. 2. We then prove our main hardness results in Sec. 3 and present our approximation algorithm in Sec. 4. Finally, we discuss related work in Sec. 5 and conclude in Sec. 6. All proofs are in the appendix.

2 Background and Notation

2.1 Polynomial Definition and Terminology

A polynomial over $\mathbf{X} = (X_1, \ldots, X_n)$ with individual degree $B < \infty$ is formally defined as (where $c_{\mathbf{d}} \in \mathbb{N}$):

$$\Phi(X_1, \ldots, X_n) = \sum_{\mathbf{d} \in \{0,\ldots,B\}^D} c_{\mathbf{d}} \cdot \prod_{t \in D} X_t^{d_t}. \tag{1}$$

▶ **Definition 2.1** (Standard Monomial Basis). *The term* $\prod_{t \in D} X_t^{d_t}$ *in Eq.* (1) *is a* monomial. *A polynomial* $\Phi(\mathbf{X})$ *is in standard monomial basis (SMB) when we keep only the terms with* $c_{\mathbf{d}} \neq 0$ *from Eq.* (1).

Unless othewise noted, we consider all polynomials to be in SMB representation. When it is unclear, we use $\mathrm{SMB}(\Phi)$ to denote the SMB form of a polynomial Φ.

▶ **Definition 2.2** (Degree). *The degree of polynomial* $\Phi(\mathbf{X})$ *is the largest* $\|\mathbf{d}\|_1$ *such that* $c_{(d_1,\ldots,d_n)} \neq 0$.

As an example, the degree of the polynomial $X^2 + 2XY^2 + Y^2$ is 3. Product terms in lineage arise only from join operations (Fig. 1), so intuitively, the degree of a lineage polynomial is analogous to the largest number of joins needed to produce a result tuple. We call a polynomial $\Phi(\mathbf{X})$ a c-TIDB-*lineage polynomial* (or simply lineage polynomial), if there exists a $\mathcal{RA}^+$ query Q, c-TIDB $\mathcal{D}$, and result tuple t such that $\Phi(\mathbf{X}) = \Phi[Q, D, t](\mathbf{X})$.

2.1.1 c-TIDBs and 1-BIDBs

An *incomplete database* Ω is a set of deterministic databases ω called possible worlds. A c-TIDB $\mathcal{D}$ is a pair $\left(\{0, \ldots, c\}^D, \mathcal{P}\right)$ such that $\{0, \ldots, c\}^D$ is an incomplete database whose set of possible worlds is the $c + 1^n$ tuple/multiplicity combinations across all $t \in D$, where $|D| = n$, $D = \bigcup_{\mathbf{M} \in \{0,\ldots,c\}^D,\ \mathbf{M}_t \geq 1} t$ is the set of possible tuples across possible worlds, and $\mathcal{P}$ is a probability distribution over $\{0, \ldots, c\}^D$.

A block independent database (BIDB) is a related probabilistic data model $\mathcal{D} = (\Omega, \mathcal{P})$ such that the base set of tuples $D = \bigcup_{\omega \in \Omega,\ t \in \omega} t$ is partitioned into a set of n independent

blocks $\left\{(b_t)_{t\in[n]}\right\}$ such that the set of tuples $\left\{(t_j)_{j\in[|b|]}\right\}$ in block b_t are disjoint from one another. This construction produces the set of possible worlds Ω that consists of all unique combinations of tuples in D with the constraint that for any $\omega \in \Omega$, no two tuples $t_j, t_{j'}, j \neq j'$ from the same block b_t exist together. A c-BIDB has the further requirement that each block has a multiplicity of at most c. We present a reduction that is useful in producing our results:

▶ **Definition 2.3** (c-TIDB reduction). *Given c-TIDB $\mathcal{D} = \left(\{0,\ldots,c\}^D, \mathcal{P}\right)$, let $\mathcal{D}' = (\Omega, \mathcal{P}')$ be the* 1-BIDB *obtained in the following manner: for each $t \in D$, create block $b_t = \left\{\langle t, jX_{t,j}\rangle_{j\in[c]}\right\}$, such that $X_{t,j} \in \{0,1\}$. The probability distribution $\mathcal{P}'$ is the one induced by $\mathbf{p} = \left((p_{t,j})_{t\in D, j\in[c]}\right)$ and the* BIDB *disjoint requirement.*

For the c-TIDB $\mathcal{D}$, each $X_t \in [c]$, while in the reduced 1-BIDB $\mathcal{D}'$, each $X_{t,j} \in \{0,1\}$. Hence, in the setting of 1-BIDB, the base case of Fig. 1 now becomes $\Phi[R, D, t] = \sum_{j\in[c]} jX_{t,j}$. Then given the disjoint requirement and the semantics for constructing the lineage polynomial over a 1-BIDB, $\Phi[R, D', t]$ is of the same structure as the reformulated polynomial Φ_R of step i) from Definition 1.3, which then implies that $\widetilde{\Phi}$ is the reduced polynomial that results from step ii) of Definition 1.3, and further that Lemma 1.4 immediately follows for 1-BIDB polynomials: $\mathbb{E}_{\mathbf{W}\sim\mathcal{P}'}[\Phi(\mathbf{W})] = \widetilde{\Phi}(\mathbf{p})$.

Aaron says: @atri, not sure if $\mathcal{P}'$ should be $\mathcal{P}''$ (in the above expectation) as discussed below. Since $\mathcal{P}' \equiv \mathcal{P}''$, then the proof still holds for Lemma 1.4, but maybe it is important to $\mathcal{P}''$ to drive the point home that we iterate over the all worlds set (as opposed to the set of possible worlds) when computing the expectation of a polynomial. Or maybe it suffices to note that $\mathcal{P}' \equiv \mathcal{P}''$.

Instead of looking only at the possible worlds of $\mathcal{D}$, one can consider all worlds, including those that cannot exist due to disjointness. The all worlds set can be modeled by $\mathbf{M} \in \{0,1\}^{cn}$,[3] such that $\mathbf{M}_{t,j} \in \mathbf{M}$ represents whether or not the multiplicity of t is j.We denote a probability distribution over all $\mathbf{M} \in \{0,1\}^n$ as $\mathcal{P}''$. When $\mathcal{P}''$ is the one induced from each $p_{t,j}$ while assigning $Pr[\mathbf{M}] = 0$ for any $\mathbf{M}$ with $\mathbf{M}_{t,j} = \mathbf{M}_{t,j'} = 1$ for $j \neq j'$, we end up with a bijective mapping from $\mathcal{P}'$ to $\mathcal{P}''$, such that each mapping is equivalent, implying the distributions are equivalent. Appendix B.2 has more details.

Let $|\Phi|$ be the number of operators in Φ.

▶ **Corollary 2.4.** *If Φ is a BIDB-lineage polynomial already in SMB, then the expectation of Φ, i.e., $\mathbb{E}[\Phi] = \widetilde{\Phi}(p_1,\ldots,p_n)$ can be computed in $O(|\Phi|)$ time.*

Queries over probabilistic databases are evaluated using the so-called possible world semantics. Under the possible world semantics, the result of a query Q over an incomplete database Ω is the set of query answers produced by evaluating Q over each possible world $\omega \in \Omega$: $\{Q(\omega) : \omega \in \Omega\}$.

The result of a query is the pair $(Q(\omega), \mathcal{P}')$ where $\mathcal{P}'$ is a probability distribution that assigns to each possible query result the sum of the probabilites of the worlds that produce this answer: $Pr[\omega \in \Omega] = \sum \omega' \in \Omega, Q(\omega') = Q(\omega) Pr[\omega']$.

Recalling Fig. 1 again, which defines the lineage polynomial $\Phi[Q, D, t]$ for any $\mathcal{RA}^+$ query. We now make a meaningful connection between possible world semantics and world assignments on the lineage polynomial.

[3] Here and later, especially in Sec. 4, we will rename the variables as $X_1,\ldots,X_n$, where $n = \sum_{i=1}^{\ell} |b_i|$.

▶ **Proposition 2.5** (Expectation of polynomials). *Given a* bag-PDB $\mathcal{D} = (\Omega, \mathcal{P})$, $\mathcal{RA}^+$ *query* Q, *and lineage polynomial* $\Phi[Q, D, t]$ *for arbitrary result tuple* t, *we have (denoting* $\mathbf{D}$ *as the random variable over* Ω*):* $\mathbb{E}_{\mathbf{D}\sim\mathcal{P}}[Q(\mathbf{D})(t)] = \mathbb{E}_{\mathbf{W}\sim\mathcal{P}}\left[\Phi[Q, D, t](\mathbf{W})\right]$.

A formal proof of Proposition 2.5 is given in Appendix B.3.[4] We focus on the problem of computing $\mathbb{E}_{\mathcal{P}}\left[\Phi[Q, D, t](\mathbf{W})\right]$ from now on, assume implicit Q, D, t, and drop them from $\Phi[Q, D, t]$ (i.e., $\Phi(\mathbf{X})$ will denote a polynomial).

2.2 Formalizing Problem 1.6

We represent lineage polynomials via *arithmetic circuits* [9], a standard way to represent polynomials over fields (particularly in the field of algebraic complexity) that we use for polynomials over $\mathbb{N}$ in the obvious way. Since we are particularly using circuits to model lineage polynomials, we can refer to these circuits as lineage circuits. However, when the meaning is clear, we will drop the term lineage and only refer to them as circuits.

▶ **Definition 2.6** (Circuit). *A circuit* $\mathtt{C}$ *is a Directed Acyclic Graph (DAG) whose source gates (in degree of* 0*) consist of elements in either* $\mathbb{N}$ *or* $\mathbf{X}$. *For each result tuple there exists one sink gate. The internal gates have binary input and are either sum* $(+)$ *or product* $(\times)$ *gates. Each gate has the following members:* `type`, `partial`, `input`, `degree`, `Lweight`, *and* `Rweight`, *where* `type` *is the value type* $\{+, \times, \textsc{var}, \textsc{num}\}$ *and* `input` *the list of inputs. Source gates have an extra member* `val` *storing the value.* $\mathtt{C}_L$ *(*$\mathtt{C}_R$*) denotes the left (right) input of* $\mathtt{C}$.

Aaron says: Does the following matter, i.e., does it point anything out special for our research?

When the underlying DAG is a tree (with edges pointing towards the root), the structure is an expression tree `T`. In such a case, the root of `T` is analogous to the sink of `C`. The fields `partial`, `degree`, `Lweight`, and `Rweight` are used in the proofs of Appendix D.

The circuits in Fig. 2 encode their respective polynomials in column Φ. Note that each circuit `C` encodes a tree, with edges pointing towards the root.

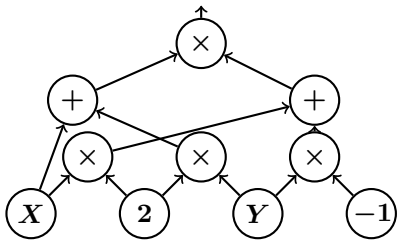

Figure 3 Circuit encoding of $(X + 2Y)(2X - Y)$

We next formally define the relationship of circuits with polynomials. While the definition assumes one sink for notational convenience, it easily generalizes to the multiple sinks case.

▶ **Definition 2.7** ($\textsc{poly}(\cdot)$). *Denote* $\textsc{poly}(\mathtt{C})$ *to be the function from the sink of circuit* $\mathtt{C}$ *to its corresponding polynomial (in SMB).* $\textsc{poly}(\cdot)$ *is recursively defined on* $\mathtt{C}$ *as follows, with addition and multiplication following the standard interpretation for polynomials:*

$$\textsc{poly}(\mathtt{C}) = \begin{cases} \textsc{poly}(\mathtt{C}_L) + \textsc{poly}(\mathtt{C}_R) & \text{if } \mathtt{C.type} = + \\ \textsc{poly}(\mathtt{C}_L) \cdot \textsc{poly}(\mathtt{C}_R) & \text{if } \mathtt{C.type} = \times \\ \mathtt{C.val} & \text{if } \mathtt{C.type} = \textsc{var} \ \textit{OR} \ \textsc{num}. \end{cases}$$

[4] Although Proposition 2.5 follows, e.g., as an obvious consequence of [28]'s Theorem 7.1, we are unaware of any formal proof for bag-probabilistic databases.

$\mathtt{C}$ need not encode $\Phi(\mathbf{X})$ in the same, default SMB representation. For instance, $\mathtt{C}$ could encode the factorized representation $(X+2Y)(2X-Y)$ of $\Phi(\mathbf{X}) = 2X^2+3XY-2Y^2$, as shown in Fig. 3, while $\text{POLY}(\mathtt{C}) = \Phi(\mathbf{X})$ is always the equivalent SMB representation.

▶ **Definition 2.8** (Circuit Set). *$\mathtt{CSet}(\Phi(\mathbf{X}))$ is the set of all possible circuits $\mathtt{C}$ such that $\text{POLY}(\mathtt{C}) = \Phi(\mathbf{X})$.*

The circuit of Fig. 3 is an element of $\mathtt{CSet}\left(2X^2+3XY-2Y^2\right)$. One can think of $\mathtt{CSet}(\Phi(\mathbf{X}))$ as the infinite set of circuits where for each element $\mathtt{C}$, $\text{POLY}(\mathtt{C}) = \Phi(\mathbf{X})$.

We are now ready to formally state the final version of Problem 1.6.

▶ **Definition 2.9** (The Expected Result Multiplicity Problem). *Let $\mathcal{D}$ be an arbitrary* BIDB-*PDB and $\mathbf{X}$ be the set of variables annotating tuples in D_Ω. Fix an $\mathcal{RA}^+$ query Q and a result tuple t. The Expected Result Multiplicity Problem is defined as follows:*

***Input**: $\mathtt{C} \in \mathtt{CSet}(\Phi(\mathbf{X}))$ for $\Phi(\mathbf{X}) = \Phi[Q,D,t]$* ***Output**: $\mathbb{E}_{\mathbf{W}\sim\mathcal{P}}[\Phi[Q,D,t](\mathbf{W})]$*

2.3 Relationship to Deterministic Query Runtimes

To decouple our results from specific join algorithms, we first abstract the cost of a join.

▶ **Definition 2.10** (Join Cost). *Denote by $T_{join}(R_1,\ldots,R_m)$ the runtime of an algorithm for computing the m-ary join $R_1 \bowtie \ldots \bowtie R_m$. We require only that the algorithm must enumerate its output, i.e., that $T_{join}(R_1,\ldots,R_m) \geq |R_1 \bowtie \ldots \bowtie R_m|$.*

Worst-case optimal join algorithms [37, 36] and query evaluation via factorized databases [39] (as well as work on FAQs [33]) can be modeled as $\mathcal{RA}^+$ queries (though the query size is data dependent). For these algorithms, $T_{join}(R_1,\ldots,R_n)$ is linear in the *AGM bound* [6]. Our cost model for general query evaluation follows from the join cost:

$$T_{det}(R,D) = |D.R| \quad T_{det}(\sigma Q, D) = T_{det}(Q,D) \quad T_{det}(\pi Q, D) = T_{det}(Q,D) + |Q(D)|$$
$$T_{det}(Q \cup Q', D) = T_{det}(Q,D) + T_{det}(Q',D) + |Q(D)| + |Q'(D)|$$
$$T_{det}(Q_1 \bowtie \ldots \bowtie Q_m, D) = T_{det}(Q_1,D) + \ldots + T_{det}(Q_m,D) + T_{join}(Q_1(D),\ldots,Q_m(D))$$

Under this model, an $\mathcal{RA}^+$ query Q evaluated over database D has runtime $O(T_{det}(Q,D))$. We assume that full table scans are used for every base relation access. We can model index scans by treating an index scan query $\sigma_\theta(R)$ as a base relation.

Finally, Lemma E.2 and Lemma E.3 show that for any $\mathcal{RA}^+$ query Q and D_Ω, there exists a circuit $\mathtt{C}^*$ such that $T_{LC}(Q, D_\Omega, \mathtt{C}^*)$ and $|\mathtt{C}^*|$ are both $O(T_{det}(Q, D_\Omega))$. Recall we assumed these two bounds when we moved from Problem 1.5 to Problem 1.6.

3 Hardness of Exact Computation

In this section, we will prove the hardness results claimed in Table 1 for a specific (family) of hard instance $(Q,\mathcal{D})$ for Problem 1.2 where $\mathcal{D}$ is a 1-TIDB. Note that this implies hardness for c-TIDBs ($c \geq 1$), BIDBs and general bag-PDB, showing Problem 1.2 cannot be done in $O(T^*_{det}(Q,D))$ runtime.

3.1 Preliminaries

Our hardness results are based on (exactly) counting the number of (not necessarily induced) subgraphs in G isomorphic to H. Let $\#(G,H)$ denote this quantity. We can think of H

as being of constant size and G as growing. In particular, we will consider the problems of computing the following counts (given G in its adjacency list representation): $\#\left(G, \text{△}\right)$ (the number of triangles), $\#\left(G, \text{⁞⁞⁞}\right)$ (the number of 3-matchings), and the latter's generalization $\#\left(G, \text{⁞}\cdots\text{⁞}^k\right)$ (the number of k-matchings). We use $T_{match}\left(k, G\right)$ to denote the optimal runtime of computing $\#\left(G, \text{⁞}\cdots\text{⁞}^k\right)$ exactly. Our hardness results in Sec. 3.2 are based on the following hardness results/conjectures:

▶ **Theorem 3.1** ([11]). *Given positive integer k and undirected graph $G = (V, E)$ with no self-loops or parallel edges, $T_{match}\left(k, G\right) \geq \omega\left(f(k) \cdot |E|^c\right)$ for any function f and fixed constant c independent of $|E|$ and k (assuming $\#\mathsf{W}[0] \neq \#\mathsf{W}[1]$).*

▶ **Conjecture 3.2.** *There exists an absolute constant $c_0 > 0$ such that for every $G = (V, E)$, we have $T_{match}\left(k, G\right) \geq \Omega\left(|E|^{c_0 \cdot k}\right)$ for large enough k.*

We note that the above conjecture is somewhat non-standard. In particular, the best known algorithm to compute $\#\left(G, \text{⁞}\cdots\text{⁞}^k\right)$ takes time $\Omega\left(|V|^{k/2}\right)$ (i.e. if this is the best algorithm then $c_0 = \frac{1}{4}$) [11]. What the above conjecture is saying is that one can only hope for a polynomial improvement over the state of the art algorithm to compute $\#\left(G, \text{⁞}\cdots\text{⁞}^k\right)$.

Our hardness result in Section 3.3 is based on the following conjectured hardness result:

▶ **Conjecture 3.3.** *There exists a constant $\epsilon_0 > 0$ such that given an undirected graph $G = (V, E)$, computing $\#\left(G, \text{△}\right)$ exactly cannot be done in time $o\left(|E|^{1+\epsilon_0}\right)$.*

The so called *Triangle detection hypothesis* (cf. [34]), which states that detecting the presence of triangles in G takes time $\Omega\left(|E|^{4/3}\right)$, implies that in Conjecture 3.3 we can take $\epsilon_0 \geq \frac{1}{3}$.

All of our hardness results rely on a simple lineage polynomial encoding of the edges of a graph. To prove our hardness result, consider a graph $G = (V, E)$, where $|E| = m$, $V = [n]$. Our lineage polynomial has a variable X_i for every i in $[n]$. Consider the polynomial $\Phi_G(\mathbf{X}) = \sum_{(i,j)\in E} X_i \cdot X_j$. The hard polynomial for our problem will be a suitable power $k \geq 3$ of the polynomial above:

▶ **Definition 3.4.** *For any graph $G = (V, E)$ and $k \geq 1$, define*

$$\Phi_G^k(X_1, \ldots, X_n) = \left(\sum_{(i,j)\in E} X_i \cdot X_j\right)^k.$$

Returning to Fig. 2, it is easy to see that $\Phi_G^k(\mathbf{X})$ is the lineage polynomial whose structure mirrors the query Q_2 from Sec. 1. Let us alias

```
SELECT 1 FROM T t1, R r, T t2
WHERE t1.city = r.city1 AND t2.city = r.city2
```

as R_i for each $i \in [k]$. The query Q^k then becomes

```
SELECT COUNT(*) FROM R1 JOIN R2 JOIN···JOIN Rk
```

Further, the c-TIDB instance of Fig. 2 generalizes to one compatible to Definition 3.4 as follows. Relation T has n tuples corresponding to each vertex for i in $[n]$, each with probability p_i and R has tuples corresponding to the edges E (each with probability of 1).[5] In other

[5] Technically, $\Phi_G^k(\mathbf{X})$ should have variables corresponding to tuples in R as well, but since they always are present with probability 1, we drop those. Our argument also works when all the tuples in R also are present with probability p but to simplify notation we assign probability 1 to edges.

words, for this instance D contains the set of n unary tuples in T (which corresponds to V) and m binary tuples in R (which corresponds to E). Note that this implies that Φ_G^k is indeed a c-TIDB-lineage polynomial.

Aaron says: Can the proofs generalize to 2-TIDB, as the new updated Fig. 2 now is?

Next, we note that the runtime for answering Q^k on deterministic database D, as defined above, is $O(m)$ (i.e. deterministic query processing is 'easy' for this query):

▶ **Lemma 3.5.** *Let Q^k and D be as defined above. Then $T_{det}\left(Q^k, D\right)$ is $O(km)$.*

Aaron says: Should the above be $T_{det}()$ or $T^*_{det}()$?

3.2 Multiple Distinct p Values

We are now ready to present our main hardness result.

Aaron says: Note that Definition 1.3 has been changed, where we compute $\widetilde{\Phi}$ from Φ_R.

▶ **Theorem 3.6.** *Let $p_0, \ldots, p_{2k}$ be $2k+1$ distinct values in $(0,1]$. Then computing $\widetilde{\Phi}_G^k(p_i, \ldots, p_i)$ (over all $i \in [2k+1]$ for arbitrary $G = (V, E)$ needs time $\Omega\left(T_{match}(k, G)\right)$, assuming $T_{match}(k, G) \geq \omega(|E|)$.*

Note that the second row of Table 1 follows from Proposition 2.5, Theorem 3.6, Lemma 3.5, and Theorem 3.1 while the third row is proved by Proposition 2.5, Theorem 3.6, Lemma 3.5, and Conjecture 3.2. Since Conjecture 3.2 is non-standard, the latter hardness result should be interpreted as follows. Any substantial polynomial improvement for Problem 1.2 (over the trivial algorithm that converts Φ into SMB and then uses Corollary 2.4 for EC) would lead to an improvement over the state of the art *upper* bounds on $T_{match}(k, G)$. Finally, note that Theorem 3.6 needs one to be able to compute the expected multiplicities over $(2k+1)$ distinct values of p_i, each of which corresponds to distinct $\mathcal{P}$ (for the same D), which explain the 'Multiple' entry in the second column in the second and third row in Table 1. Next, we argue how to get rid of this latter requirement.

3.3 Single p value

While Theorem 3.6 shows that computing $\widetilde{\Phi}(p, \ldots, p)$ for multiple values of p in general is hard it does not rule out the possibility that one can compute this value exactly for a *fixed* value of p. Indeed, it is easy to check that one can compute $\widetilde{\Phi}(p, \ldots, p)$ exactly in linear time for $p \in \{0, 1\}$. Next we show that these two are the only possibilities:

▶ **Theorem 3.7.** *Fix $p \in (0,1)$. Then assuming Conjecture 3.3 is true, any algorithm that computes $\widetilde{\Phi}_G^3(p, \ldots, p)$ for arbitrary $G = (V, E)$ exactly has to run in time $\Omega\left(|E|^{1+\epsilon_0}\right)$, where ϵ_0 is as defined in Conjecture 3.3.*

Note that Proposition 2.5 and Theorem 3.7 above imply the hardness result in the first row of Table 1. We note that Theorem 3.1 and Conjecture 3.2 (and the lower bounds in the second and third row of Table 1) need k to be large enough (in particular, we need a family of hard queries). But the above Theorem 3.7 (and the lower bound in first row of Table 1) holds for $k = 3$ (and hence for a fixed query).

4 $1 \pm \epsilon$ Approximation Algorithm

Aaron says: If we get rid of the problem statements, then we need to remember to get rid of references to particular problem statements, as in below.

In Sec. 3, we showed that the answer to Problem 1.6 is no. With this result, we now design an approximation algorithm for our problem that runs in $O(|\mathtt{C}|)$ for a very broad class of circuits (see the discussion after Lemma 4.8 for more). The following approximation algorithm applies to BIDB lineage polynomials (over $\mathcal{RA}^+$ queries), though our bounds are more meaningful for a non-trivial subclass of queries over BIDBs that contains all queries on TIDBs, as well as the queries of the PDBench benchmark [1]. All proofs and pseudocode can be found in Appendix D.

Aaron says: We are going to have to rework γ in this section, as well as the proof for our result.

4.1 Preliminaries and some more notation

We now introduce definitions and notation related to circuits and polynomials that we will need to state our upper bound results. First we introduce the expansion $\mathtt{E}(\mathtt{C})$ of circuit $\mathtt{C}$ which is used in our algorithm for sampling monomials (part of our approximation algorithm).

▶ **Definition 4.1** ($\mathtt{E}(\mathtt{C})$). *For a circuit C, we define $E(C)$ as a list of tuples (v, c), where v is a set of variables and $c \in \mathbb{N}$. $E(C)$ has the following recursive definition ($\circ$ is list concatenation).*

$$E(C) = \begin{cases} E(C_L) \circ E(C_R) & \textbf{if } C.type = + \\ \{(v_L \cup v_R, c_L \cdot c_R) \mid (v_L, c_L) \in E(C_L), (v_R, c_R) \in E(C_R)\} & \textbf{if } C.type = \times \\ List[(\emptyset, C.val)] & \textbf{if } C.type = NUM \\ List[(\{C.val\}, 1)] & \textbf{if } C.type = VAR. \end{cases}$$

Later on, we will denote the monomial composed of the variables in $\mathtt{v}$ as $\mathtt{v}_\mathtt{m}$. As an example of $\mathtt{E}(\mathtt{C})$, consider $\mathtt{C}$ illustrated in Fig. 3. $\mathtt{E}(\mathtt{C})$ is then $[(X, 2), (XY, -1), (XY, 4), (Y, -2)]$. This helps us redefine $\widetilde{\Phi}$ (see Eq. (2)) in a way that makes our algorithm more transparent.

▶ **Definition 4.2** ($|\mathtt{C}|$). *For any circuit C, the corresponding* positive circuit*, denoted $|C|$, is obtained from C as follows. For each leaf node ℓ of C where $\ell.type$ is NUM, update $\ell.value$ to $|\ell.value|$.*

We will overload notation and use $|\mathtt{C}|(\mathbf{X})$ to mean $\text{POLY}(|\mathtt{C}|)$. Conveniently, $|\mathtt{C}|(1, \ldots, 1)$ gives us $\sum_{(\mathtt{v},\mathtt{c}) \in \mathtt{E}(\mathtt{C})} |\mathtt{c}|$.

▶ **Definition 4.3** ($\text{SIZE}(\cdot)$, $\text{DEPTH}(\cdot)$). *The functions SIZE and DEPTH output the number of gates and levels respectively for input C.*

▶ **Definition 4.4** ($\text{DEG}(\cdot)$). [6] *$DEG(C)$ is defined recursively as follows:*

$$DEG(C) = \begin{cases} \max(DEG(C_L), DEG(C_R)) & \text{if } C.type = + \\ DEG(C_L) + DEG(C_R) + 1 & \text{if } C.type = \times \\ 1 & \text{if } C.type = VAR \\ 0 & \text{otherwise.} \end{cases}$$

[6] Note that the degree of $\text{POLY}(|\mathtt{C}|)$ is always upper bounded by $\text{DEG}(\mathtt{C})$ and the latter can be strictly larger (e.g. consider the case when $\mathtt{C}$ multiplies two copies of the constant 1– here we have $\deg(\mathtt{C}) = 1$ but degree of $\text{POLY}(|\mathtt{C}|)$ is 0).

Next, we use the following notation for the complexity of multiplying integers:

▶ **Definition 4.5** ($\overline{\mathcal{M}}(\cdot,\cdot)$)**.** [7] *In a RAM model of word size of W-bits, $\overline{\mathcal{M}}(M,W)$ denotes the complexity of multiplying two integers represented with M-bits. (We will assume that for input of size N, $W = O(\log N)$.)*

Finally, to get linear runtime results, we will need to define another parameter modeling the (weighted) number of monomials in $\mathtt{E}(\mathtt{C})$ that need to be 'canceled' when monomials with dependent variables are removed (**??**). Let $\textsc{isInd}(\cdot)$ be a boolean function returning true if monomial $\mathtt{v_m}$ is composed of independent variables and false otherwise; further, let $\mathbb{1}_\theta$ also be a boolean function returning true if θ evaluates to true.

▶ **Definition 4.6** (Parameter γ)**.** *Given a* BIDB *circuit $\mathcal{C}$ define*

Aaron says: *Technically, $\boldsymbol{v}$ is a set of variables rather than a monomial. Perhaps we don't need the $\textsc{var}(\cdot)$ function and can replace is with a function that returns the monomial represented by a set of variables. FIXED: need to propogate this to the appendix ($\boldsymbol{v_m}$)*

Aaron says: *To add, this is an issue on line 1073, 1117 of app C.*

$$\gamma(\mathcal{C}) = \frac{\sum_{(v,\mathtt{c})\in E(\mathcal{C})} |\mathtt{c}| \cdot \mathbb{1}_{\neg \textsc{isInd}(v_m)}}{|\mathcal{C}|\,(1,\ldots,1)}.$$

4.2 Our main result

Algorithm Idea. Our approximation algorithm ($\textsc{ApproximateP}\widetilde{\Phi}$ pseudo code in Appendix D.1) is based on the following observation. Given a lineage polynomial $\Phi(\mathbf{X}) = \textsc{poly}(\mathtt{C})$ for circuit $\mathtt{C}$ over $BIDB$, we have:

$$\widetilde{\Phi}\,(p_1,\ldots,p_n) = \sum_{(\mathtt{v},\mathtt{c})\in\mathtt{E}(\mathtt{C})} \mathbb{1}_{\textsc{isInd}(\mathtt{v_m})} \cdot \mathtt{c} \cdot \prod_{X_i\in\mathtt{v}} p_i. \tag{2}$$

Given the above, the algorithm is a sampling based algorithm for the above sum: we sample (via $\textsc{SampleMonomial}$) $(\mathtt{v},\mathtt{c}) \in \mathtt{E}(\mathtt{C})$ with probability proportional to $|\mathtt{c}|$ and compute $\mathtt{Y} = \mathbb{1}_{\textsc{isInd}(\mathtt{v_m})} \cdot \prod_{X_i\in\mathtt{v}} p_i$. Repeating the sampling appropriate number of times and computing the average of $\mathtt{Y}$ gives us our final estimate. $\textsc{OnePass}$ is used to compute the sampling probabilities needed in $\textsc{SampleMonomial}$ (details are in Appendix D).

Runtime analysis. We can argue the following runtime for the algorithm outlined above:

▶ **Theorem 4.7.** *Let $\mathcal{C}$ be an arbitrary* BIDB *circuit and define $\Phi(\mathbf{X}) = \textsc{poly}(\mathcal{C})$ and let $k = \textsc{deg}(\mathcal{C})$. Let $\gamma = \gamma(\mathcal{C})$. Further let it be the case that $p_i \geq p_0$ for all $i \in [n]$. Then an estimate $\mathcal{E}$ of $\widetilde{\Phi}(p_1,\ldots,p_n)$ satisfying*

$$Pr\left(\left|\mathcal{E} - \widetilde{\Phi}(p_1,\ldots,p_n)\right| > \epsilon' \cdot \widetilde{\Phi}(p_1,\ldots,p_n)\right) \leq \delta \tag{3}$$

[7] We note that when doing arithmetic operations on the RAM model for input of size N, we have that $\overline{\mathcal{M}}(O(\log N), O(\log N)) = O(1)$. More generally we have $\overline{\mathcal{M}}(N, O(\log N)) = O(N \log N \log\log N)$.

can be computed in time

$$O\left(\left(\textit{SIZE}(\mathcal{C}) + \frac{\log\frac{1}{\delta}\cdot k\cdot\log k\cdot \textit{DEPTH}(\mathcal{C}))}{(\epsilon')^2\cdot(1-\gamma)^2\cdot p_0^{2k}}\right)\cdot\overline{\mathcal{M}}\left(\log\left(|\mathcal{C}|\,(1,\ldots,1)\right),\log\left(\textit{SIZE}(\mathcal{C})\right)\right)\right). \quad (4)$$

In particular, if $p_0 > 0$ and $\gamma < 1$ are absolute constants then the above runtime simplifies to $O_k\left(\left(\frac{1}{(\epsilon')^2}\cdot\textit{SIZE}(\mathcal{C})\cdot\log\frac{1}{\delta}\right)\cdot\overline{\mathcal{M}}\left(\log\left(|\mathcal{C}|\,(1,\ldots,1)\right),\log\left(\textit{SIZE}(\mathcal{C})\right)\right)\right)$.

The restriction on γ is satisfied by any TIDB (where $\gamma = 0$) as well as for all three queries of the PDBench BIDB benchmark (see Appendix D.10 for experimental results).

We briefly connect the runtime in Eq. (4) to the algorithm outline earlier (where we ignore the dependence on $\overline{\mathcal{M}}(\cdot,\cdot)$, which is needed to handle the cost of arithmetic operations over integers). The $\textsc{size}(\mathtt{C})$ comes from the time take to run OnePass once (OnePass essentially computes $|\mathtt{C}|\,(1,\ldots,1)$ using the natural circuit evaluation algorithm on $\mathtt{C}$). We make $\frac{\log\frac{1}{\delta}}{(\epsilon')^2\cdot(1-\gamma)^2\cdot p_0^{2k}}$ many calls to SampleMonomial (each of which essentially traces $O(k)$ random sink to source paths in $\mathtt{C}$ all of which by definition have length at most $\textsc{depth}(\mathtt{C})$).

Finally, we address the $\overline{\mathcal{M}}\left(\log\left(|\mathtt{C}|\,(1,\ldots,1)\right),\log\left(\textsc{size}(\mathtt{C})\right)\right)$ term in the runtime.

▶ **Lemma 4.8.** *For any* BIDB *circuit $\mathcal{C}$ with $\textit{DEG}(\mathcal{C}) = k$, we have $|\mathcal{C}|\,(1,\ldots,1) \leq 2^{2^k\cdot\textit{DEPTH}(\mathcal{C})}$. Further, if $\mathcal{C}$ is a tree, then we have $|\mathcal{C}|\,(1,\ldots,1) \leq \textit{SIZE}(\mathcal{C})^{O(k)}$.*

Note that the above implies that with the assumption $p_0 > 0$ and $\gamma < 1$ are absolute constants from Theorem 4.7, then the runtime there simplifies to $O_k\left(\frac{1}{(\epsilon')^2}\cdot\textsc{size}(\mathtt{C})^2\cdot\log\frac{1}{\delta}\right)$ for general circuits $\mathtt{C}$. If $\mathtt{C}$ is a tree, then the runtime simplifies to $O_k\left(\frac{1}{(\epsilon')^2}\cdot\textsc{size}(\mathtt{C})\cdot\log\frac{1}{\delta}\right)$, which then answers Problem 1.6 is yes for such circuits.

Finally, note that by Proposition E.1 and Lemma E.2 for any $\mathcal{RA}^+$ query Q, there exists a circuit $\mathtt{C}^*$ for $\Phi[Q,D,t]$ such that $\textsc{depth}(\mathtt{C}^*) \leq O_{|Q|}(\log n)$ and $\textsc{size}(\mathtt{C}) \leq O_k\left(T^*_{det}(Q,D_\Omega)\right)$. Using this along with Lemma 4.8, Theorem 4.7 and the fact that $n \leq T^*_{det}(Q,D_\Omega)$, we answer Problem 1.5 in the affirmative as follows:

▶ **Corollary 4.9.** *Let Q be an $\mathcal{RA}^+$ query and $\mathcal{D}$ be an* BIDB *with $p_0 > 0$ and $\gamma < 1$ (where p_0, γ as in Theorem 4.7) are absolute constants. Let $\Phi(\mathbf{X}) = \Phi[Q,D,t]$ for any result tuple t with $\deg(\Phi) = k$. Then one can compute an approximation satisfying Eq. (3) in time $O_{k,|Q|,\epsilon',\delta}\left(T^*_{det}(Q,D_\Omega)\right)$ (given Q, D_Ω and p_i for each $i \in [n]$ that defines $\mathcal{P}$).*

If we want to approximate the expected multiplicities of all $Z = O(n^k)$ result tuples t simultaneously, we just need to run the above result with δ replaced by $\frac{\delta}{Z}$. Note this increases the runtime by only a logarithmic factor.

5 Related Work

Probabilistic Databases (PDBs) have been studied predominantly for set semantics. Approaches for probabilistic query processing (i.e., computing marginal probabilities of tuples), fall into two broad categories. *Intensional* (or *grounded*) query evaluation computes the *lineage* of a tuple and then the probability of the lineage formula. It has been shown that computing the marginal probability of a tuple is #P-hard [46] (by reduction from weighted model counting). The second category, *extensional* query evaluation, is in PTIME, but is limited to certain classes of queries. Dalvi et al. [14] and Olteanu et al. [21] proved dichotomies for UCQs and two classes of queries with negation, respectively. Amarilli et al. investigated tractable classes of databases for more complex queries [3]. Another line of work

studies which structural properties of lineage formulas lead to tractable cases [31, 41, 44]. In this paper we focus on intensional query evaluation with polynomials.

Many data models have been proposed for encoding PDBs more compactly than as sets of possible worlds. These include tuple-independent databases [47] (TIDBs), block-independent databases (BIDBs) [42], and *PC-tables* [26]. Fink et al. [19] study aggregate queries over a probabilistic version of the extension of K-relations for aggregate queries proposed in [4] (*pvc-tables*) that supports bags, and has runtime complexity linear in the size of the lineage. However, this lineage is encoded as a tree; the size (and thus the runtime) are still superlinear in $T^*_{det}(Q, D_\Omega)$. The runtime bound is also limited to a specific class of (hierarchical) queries, suggesting the possibility of a generalization of [14]'s dichotomy result to bag-PDBs.

Several techniques for approximating tuple probabilities have been proposed in related work [20, 15, 38, 12], relying on Monte Carlo sampling, e.g., [12], or a branch-and-bound paradigm [38]. Our approximation algorithm is also based on sampling.

Compressed Encodings are used for Boolean formulas (e.g, various types of circuits including OBDDs [29]) and polynomials (e.g., factorizations [39]) some of which have been utilized for probabilistic query processing, e.g., [29]. Compact representations for which probabilities can be computed in linear time include OBDDs, SDDs, d-DNNF, and FBDD. [16] studies circuits for absorptive semirings while [45] studies circuits that include negation (expressed as the monus operation). Algebraic Decision Diagrams [7] (ADDs) generalize BDDs to variables with more than two values. Chen et al. [10] introduced the generalized disjunctive normal form. Appendix H covers more related work on fine-grained complexity.

6 Conclusions and Future Work

We have studied the problem of calculating the expected multiplicity of a query result tuple, a problem that has a practical application in probabilistic databases over multisets. We show that under various parameterized complexity hardness results/conjectures computing the expected multiplicities exactly is not possible in time linear in the corresponding deterministic query processing time. We prove that it is possible to approximate the expectation of a lineage polynomial in linear time in the deterministic query processing over TIDBs and BIDBs (assuming that there are few cancellations). Interesting directions for future work include development of a dichotomy for bag PDBs. While we can handle higher moments (this follows fairly easily from our existing results– see Appendix F), more general approximations are an interesting area for exploration, including those for more general data models.

References

1 pdbench. `http://pdbench.sourceforge.net/`. Accessed: 2020-12-15.

2 Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha U. Nabar, Tomoe Sugihara, and Jennifer Widom. Trio: A system for data, uncertainty, and lineage. In *VLDB*, pages 1151–1154, 2006.

3 Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Probabilities and provenance via tree decompositions. *PODS*, 2015.

4 Yael Amsterdamer, Daniel Deutch, and Val Tannen. Provenance for aggregate queries. In *PODS*, pages 153–164, 2011.

5 Lyublena Antova, Thomas Jansen, Christoph Koch, and Dan Olteanu. Fast and simple relational processing of uncertain data.

6 Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. *SIAM J. Comput.*, 42(4):1737–1767, 2013. `doi:10.1137/110859440`.

7 R. Iris Bahar, Erica A. Frohm, Charles M. Gaona, Gary D. Hachtel, Enrico Macii, Abelardo Pardo, and Fabio Somenzi. Algebraic decision diagrams and their applications. In *IEEE CAD*, 1993.

8 George Beskales, Ihab F. Ilyas, and Lukasz Golab. Sampling the repairs of functional dependency violations under hard constraints. *Proc. VLDB Endow.*, 3(1):197–207, 2010.

9 Peter Bürgisser, Michael Clausen, and Mohammad Amin Shokrollahi. *Algebraic complexity theory*, volume 315. Springer, 1997.

10 Hubie Chen and Martin Grohe. Constraint satisfaction with succinctly specified relations. *J. Comput. Syst. Sci.*, 76(8):847–860, 2010.

11 Radu Curticapean. Counting matchings of size k is w[1]-hard. In *ICALP*, volume 7965, pages 352–363, 2013.

12 N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB*, 16(4):544, 2007.

13 Nilesh Dalvi and Dan Suciu. The dichotomy of conjunctive queries on probabilistic structures. In *PODS*, pages 293–302, 2007.

14 Nilesh Dalvi and Dan Suciu. The dichotomy of probabilistic inference for unions of conjunctive queries. *JACM*, 59(6):30, 2012.

15 Maarten Van den Heuvel, Peter Ivanov, Wolfgang Gatterbauer, Floris Geerts, and Martin Theobald. Anytime approximation in probabilistic databases via scaled dissociations. In *SIGMOD*, pages 1295–1312, 2019.

16 Daniel Deutch, Tova Milo, Sudeepa Roy, and Val Tannen. Circuits for datalog provenance. In *ICDT*, pages 201–212, 2014.

17 Su Feng, Boris Glavic, Aaron Huber, and Oliver Kennedy. Efficient uncertainty tracking for complex queries with attribute-level bounds. In *SIGMOD*, 2021.

18 Su Feng, Aaron Huber, Boris Glavic, and Oliver Kennedy. Uncertainty annotated databases - a lightweight approach for approximating certain answers. In *SIGMOD*, 2019.

19 Robert Fink, Larisa Han, and Dan Olteanu. Aggregation in probabilistic databases via knowledge compilation. *PVLDB*, 5(5):490–501, 2012.

20 Robert Fink, Jiewen Huang, and Dan Olteanu. Anytime approximation in probabilistic databases. *VLDBJ*, 22(6):823–848, 2013.

21 Robert Fink and Dan Olteanu. Dichotomies for queries with negation in probabilistic databases. *TODS*, 41(1):4:1–4:47, 2016.

22 Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2006. doi:10.1007/3-540-29953-X.

23 Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009.

24 George Grätzer. *Universal algebra*. Springer Science & Business Media, 2008.

25 Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.

26 Todd J Green and Val Tannen. Models for incomplete and probabilistic information. In *EDBT*, pages 278–296. 2006.

27 T. Imielinski and W. Lipski. Incomplete information in relational databases. 1989.

28 Tomasz Imieliński and Witold Lipski Jr. Incomplete information in relational databases. *JACM*, 31(4):761–791, 1984.

29 Abhay Kumar Jha and Dan Suciu. Probabilistic databases with markoviews. *PVLDB*, 5(11):1160–1171, 2012.

30 Richard M. Karp, Michael Luby, and Neal Madras. Monte-carlo approximation algorithms for enumeration problems. *J. Algorithms*, 10(3):429–448, 1989.

31 Batya Kenig, Avigdor Gal, and Ofer Strichman. A new class of lineage expressions over probabilistic databases computable in p-time. In *SUM*, volume 8078, pages 219–232, 2013.

32 Oliver Kennedy and Christoph Koch. Pip: A database system for great and small expectations. In *ICDE*, 2010.

33 Mahmoud Abo Khamis, Hung Q. Ngo, and Atri Rudra. Faq: Questions asked frequently. In *PODS*, pages 13–28, 2016.

34 Tsvi Kopelowitz and Virginia Vassilevska Williams. Towards optimal set-disjointness and set-intersection data structures. In *ICALP*, volume 168, pages 74:1–74:16, 2020.

35 Poonam Kumari, Said Achmiz, and Oliver Kennedy. Communicating data quality in on-demand curation. In *QDB*, 2016.

36 Hung Q. Ngo. Worst-case optimal join algorithms: Techniques, results, and open problems. In *PODS*, 2018.

37 Hung Q. Ngo, Christopher Ré, and Atri Rudra. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec.*, 42(4):5–16, 2013.

38 Dan Olteanu, Jiewen Huang, and Christoph Koch. Approximate confidence computation in probabilistic databases. In *ICDE*, pages 145–156, 2010.

39 Dan Olteanu and Maximilian Schleich. Factorized databases. *SIGMOD Rec.*, 45(2):5–16, 2016.

40 Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proc. VLDB Endow.*, 10(11):1190–1201, 2017.

41 Sudeepa Roy, Vittorio Perduca, and Val Tannen. Faster query answering in probabilistic databases using read-once functions. In *ICDT*, 2011.

42 C. Ré and D. Suciu. Materialized views in probabilistic databases: for information exchange and query optimization. In *VLDB*, pages 51–62, 2007.

43 Christopher De Sa, Alexander Ratner, Christopher Ré, Jaeho Shin, Feiran Wang, Sen Wu, and Ce Zhang. Incremental knowledge base construction using deepdive. *VLDB J.*, 26(1):81–105, 2017.

44 Prithviraj Sen, Amol Deshpande, and Lise Getoor. Read-once functions and query evaluation in probabilistic databases. *PVLDB*, 3(1):1068–1079, 2010.

45 Pierre Senellart. Provenance and probabilities in relational databases. *SIGMOD Record*, 46(4):5–15, 2018.

46 Leslie G. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8(3):410–421, 1979.

47 Guy Van den Broeck and Dan Suciu. Query processing on probabilistic data: A survey. 2017.

48 Virginia Vassilevska Williams. Some open problems in fine-grained complexity. *SIGACT News*, 49(4):29–35, 2018. doi:10.1145/3300150.3300158.

49 Ying Yang, Niccolò Meneghetti, Ronny Fehling, Zhen Hua Liu, Dieter Gawlick, and Oliver Kennedy. Lenses: An on-demand approach to etl. *PVLDB*, 8(12):1578–1589, 2015.

7 Acknowledgements

We thank Virginia Williams for showing us Eq. (20), which greatly simplified our earlier proof of Lemma 3.8, and for graciously allowing us to use it.

A Generalizing Beyond Set Inputs

A.1 TIDBs

In our definition of TIDBs (Sec. 2.1.1), we assumed a model of TIDBs where each input tuple is assigned a probability p of having multiplicity 1. That is, we assumed inputs to be sets, but interpret queries under bag semantics. Other sensible generalizations of TIDBs from set semantics to bag semantics also exist.

One very natural such generalization is to assign each input tuple t a multiplicity m_t and probability p: the tuple has probability p to exists with multiplicity m_t, and otherwise has multiplicity 0. If the maximal multiplicity of all input tuples in the TIDB is bounded by some constant, then a generalization of our hardness results and approximation algorithm can be achieved by changing the construction of lineage polynomials (in Fig. 1) as follows (all other cases remain the same as in fig. 1):

$$\Phi[R, D_\Omega, t] = \begin{cases} m_t X_t & \text{if } D_\Omega.R\,(t) = m_t \\ 0 & \text{otherwise.} \end{cases}$$

That is the variable representing a tuple is multiplied by m_t to encode the tuple's multiplicity m_t. We note that our lower bounds still hold for this model since we only need $m_t = 1$ for all tuples t. Further, it can be argued that our proofs (as is) for approximation algorithms also work for this model. The only change is that since we now allow $m_t > 1$ some of the constants in the runtime analysis of our algorithms change but the overall asymptotic runtime bound remains the same.

Yet another option would be to assign each tuple a probability distribution over multiplicities. It seems very unlikely that our results would extend to a model that allows arbitrary probability distributions over multiplicities (our current proof techniques definitely break down). However, we would like to note that the special case of a Poisson binomial distribution (sum of independent but not necessarily identical Bernoulli trials) over multiplicities can be handled as follows: we add an additional identifier attribute to each relation in the database. For a tuple t with maximal multiplicity m_t, we create m_t copies of t with different identifiers. To answer a query over this encoding, we first project away the identifier attribute (note that as per Fig. 1, in Φ this would add up all the variables corresponding to the same tuple t).

A.2 BIDBs

The approach described above works for BIDBs as well if we define the bag version of BIDBs to associate each tuple t a multiplicity m_t. Recall that we associate each tuple in a block with a unique variable. Thus, the modified lineage polynomial construction shown above can be applied for BIDBs too (and our approximation results also hold).

B Missing details from Section 2

B.1 $\mathcal{K}$-relations and $\mathbb{N}[\mathbf{X}]$-encoded PDBs

We can use $\mathcal{K}$-relations to model bags. A *$\mathcal{K}$-relation* [25] is a relation whose tuples are annotated with elements from a commutative semiring $\mathcal{K} = \{K, \oplus_\mathcal{K}, \otimes_\mathcal{K}, \mathbb{0}_\mathcal{K}, \mathbb{1}_\mathcal{K}\}$. A commutative semiring is a structure with a domain K and associative and commutative binary operations $\oplus_\mathcal{K}$ and $\otimes_\mathcal{K}$ such that $\otimes_\mathcal{K}$ distributes over $\oplus_\mathcal{K}$, $\mathbb{0}_\mathcal{K}$ is the identity of $\oplus_\mathcal{K}$,

$\mathbb{1}_\mathcal{K}$ is the identity of $\otimes_\mathcal{K}$, and $\mathbb{0}_\mathcal{K}$ annihilates all elements of K when combined by $\otimes_\mathcal{K}$. Let $\mathcal{U}$ be a countable domain of values. Formally, an n-ary $\mathcal{K}$-relation R over $\mathcal{U}$ is a function $R : \mathcal{U}^n \to K$ with finite support $supp(R) = \{t \mid R(t) \neq \mathbb{0}_\mathcal{K}\}$. A $\mathcal{K}$-database is defined similarly, where we view the $\mathcal{K}$-database (relation) as a function mapping tuples to their respective annotations. $\mathcal{RA}^+$ query semantics over $\mathcal{K}$-relations are analogous to the lineage construction semantics of Fig. 1, with the exception of replacing $+$ with $\oplus_\mathcal{K}$ and $\cdot$ with $\otimes_\mathcal{K}$.

Consider the semiring $\mathbb{N} = \{\mathbb{N}, +, \times, 0, 1\}$ of natural numbers. $\mathbb{N}$-databases model bag semantics by annotating each tuple with its multiplicity. A probabilistic $\mathbb{N}$-database ($\mathbb{N}$-PDB) is a PDB where each possible world is an $\mathbb{N}$-database. We study the problem of computing statistical moments for query results over such databases. Given an $\mathbb{N}$-PDB $\mathcal{D} = (\Omega, \mathcal{P})$, ($\mathcal{RA}^+$) query Q, and possible result tuple t, we sum $Q(D)(t) \cdot \mathcal{P}(D)$ for all $D \in \Omega$ to compute the expected multiplicity of t. Intuitively, the expectation of $Q(D)(t)$ is the number of duplicates of t we expect to find in result of query Q.

Let $\mathbb{N}[\mathbf{X}]$ denote the set of polynomials over variables $\mathbf{X} = (X_1, \ldots, X_n)$ with natural number coefficients and exponents. Consider now the semiring (abusing notation) $\mathbb{N}[\mathbf{X}] = \{\mathbb{N}[\mathbf{X}], +, \cdot, 0, 1\}$ whose domain is $\mathbb{N}[\mathbf{X}]$, with the standard addition and multiplication of polynomials. We define an $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$ as the tuple $(D_{\mathbb{N}[\mathbf{X}]}, \mathcal{P})$, where $\mathbb{N}[\mathbf{X}]$-database $D_{\mathbb{N}[\mathbf{X}]}$ is paired with the probability distribution $\mathcal{P}$ across the set of possible worlds *represented* by $D_{\mathbb{N}[\mathbf{X}]}$, i.e. the one induced from $\mathcal{P}_{\mathbb{N}[\mathbf{X}]}$, the probability distribution over $\mathbf{X}$. Note that the notation is slightly abused since the first element of the pair is an encoded set of possible worlds, i.e. $D_{\mathbb{N}[\mathbf{X}]}$ is the deterministic bounding database. We denote by $\Phi[Q, D_{\mathbb{N}[\mathbf{X}]}, t]$ the annotation of tuple t in the result of $Q(D_{\mathbb{N}[\mathbf{X}]})(t)$, and as before, interpret it as a function $\Phi[Q, D_{\mathbb{N}[\mathbf{X}]}, t] : \{0,1\}^{|\mathbf{X}|} \to \mathbb{N}$ from vectors of variable assignments to the corresponding value of the annotating polynomial. $\mathbb{N}[\mathbf{X}]$-encoded PDBs and a function Mod (which transforms an $\mathbb{N}[\mathbf{X}]$-encoded PDB to an equivalent $\mathbb{N}$-PDB) are both formalized next.

To justify the use of $\mathbb{N}[\mathbf{X}]$-databases, we need to show that we can encode any $\mathbb{N}$-PDB in this way and that the query semantics over this representation coincides with query semantics over its respective $\mathbb{N}$-PDB. For that it will be opportune to define representation systems for $\mathbb{N}$-PDBs.

Boris says: cite

▶ **Definition B.1** (Representation System). *A representation system for $\mathbb{N}$-PDBs is a tuple $(\mathcal{M}, Mod)$ where $\mathcal{M}$ is a set of representations and Mod associates with each $M \in \mathcal{M}$ an $\mathbb{N}$-PDB $\mathcal{D}$. We say that a representation system is* closed *under a class of queries $\mathcal{Q}$ if for any query $Q \in \mathcal{Q}$ and $M \in \mathcal{M}$ we have:*

$$Mod(Q(M)) = Q(Mod(M))$$

A representation system is complete *if for every $\mathbb{N}$-PDB $\mathcal{D}$ there exists $M \in \mathcal{M}$ such that:*

$$Mod(M) = \mathcal{D}$$

As mentioned above we will use $\mathbb{N}[\mathbf{X}]$-databases paired with a probability distribution as a representation system, referring to such databases as $\mathbb{N}[\mathbf{X}]$-encoded PDBs. Given $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$, one can think of the of $\mathcal{P}$ as the probability distribution across all worlds $\{0,1\}^n$. Denote a particular world to be $\mathbf{w}$. For convenience let $\psi_\mathbf{w} : \mathcal{D}_{\mathbb{N}[\mathbf{X}]} \to \mathcal{D}_\mathbb{N}$ be a function that computes the corresponding $\mathbb{N}$-PDB upon assigning all values $w_i \in \mathbf{w}$ to $X_i \in \mathbf{X}$ of $D_{\mathbb{N}[\mathbf{X}]}$. Note the one-to-one correspondence between elements $\mathbf{w} \in \{0,1\}^n$ to the worlds encoded by $D_{\mathbb{N}[\mathbf{X}]}$ when $\mathbf{w}$ is assigned to $\mathbf{X}$ (assuming a domain of $\{0,1\}$ for each X_i). We can think of $\psi_\mathbf{w}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})(t)$ as the semiring homomorphism $\mathbb{N}[\mathbf{X}] \to \mathbb{N}$ that applies the assignment $\mathbf{w}$ to all variables $\mathbf{X}$ of a polynomial and evaluates the resulting expression in $\mathbb{N}$.

Boris says: explain connection to homomorphism lifting in K-relations

▶ **Definition B.2** ($Mod\left(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right)$). *Given an* $\mathbb{N}[\mathbf{X}]$*-encoded PDB* $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$*, we compute its equivalent* $\mathbb{N}$-PDB $\mathcal{D}_{\mathbb{N}} = Mod\left(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right) = (\Omega, \mathcal{P}')$ *as:*

$$\Omega = \{\psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) \mid \mathbf{w} \in \{0,1\}^n\}$$

$$\forall D \in \Omega : Pr(D) = \sum_{\mathbf{w} \in \{0,1\}^n : \psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = D} Pr(\mathbf{w})$$

For instance, consider a $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$ consisting of a single tuple $t_1 = (1)$ annotated with $X_1 + X_2$ with probability distribution $Pr([0,0]) = 0$, $Pr([0,1]) = 0$, $Pr([1,0]) = 0.3$ and $Pr([1,1]) = 0.7$. This $\mathbb{N}[\mathbf{X}]$-encoded PDB encodes two possible worlds (with non-zero probability) that we denote using their world vectors.

$$D_{[0,1]}(t_1) = 1 \quad \textbf{and} \quad D_{[1,1]}(t_1) = 2$$

Importantly, as the following proposition shows, any finite $\mathbb{N}$-PDB can be encoded as an $\mathbb{N}[\mathbf{X}]$-encoded PDB and $\mathbb{N}[\mathbf{X}]$-encoded PDBs are closed under $\mathcal{RA}^+$[25].

▶ **Proposition B.3.** $\mathbb{N}[\mathbf{X}]$*-encoded PDBs are a complete representation system for* $\mathbb{N}$*-PDBs that is closed under* $\mathcal{RA}^+$ *queries.*

Proof. To prove that $\mathbb{N}[\mathbf{X}]$-encoded PDBs are complete consider the following construction that for any $\mathbb{N}$-PDB $\mathcal{D} = (\Omega, \mathcal{P})$ produces an $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]} = (D_{\mathbb{N}[\mathbf{X}]}, \mathcal{P}')$ such that $Mod(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = \mathcal{D}$. Let $\Omega = \{D_1, \ldots, D_{|\Omega|}\}$. For each world D_i we create a corresponding variable X_i. In $D_{\mathbb{N}[\mathbf{X}]}$ we assign each tuple t the polynomial:

$$D_{\mathbb{N}[\mathbf{X}]}(t) = \sum_{i=1}^{|\Omega|} D_i(t) \cdot X_i$$

The probability distribution $\mathcal{P}'$ assigns all world vectors zero probability except for $|\Omega|$ world vectors (representing the possible worlds) $\mathbf{w}_i$. All elements of $\mathbf{w}_i$ are zero except for the position corresponding to variables X_i which is set to 1. Unfolding definitions it is trivial to show that $Mod(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = \mathcal{D}$. Thus, $\mathbb{N}[\mathbf{X}]$-encoded PDBs are a complete representation system.

Since $\mathbb{N}[\mathbf{X}]$ is the free object in the variety of semirings, Birkhoff's HSP theorem implies that any assignment $\mathbf{X} \to \mathbb{N}$, which includes as a special case the assignments $\psi_{\mathbf{w}}$ used here, uniquely extends to the semiring homomorphism alluded to above, $\psi_{\mathbf{w}}\left(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right)(t) : \mathbb{N}[\mathbf{X}] \to \mathbb{N}$. For a polynomial $\psi_{\mathbf{w}}\left(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right)(t)$ substitutes variables based on $\mathbf{w}$ and then evaluates the resulting expression in $\mathbb{N}$. For instance, consider the polynomial $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}(t) = \Phi = X + Y$ and assignment $\mathbf{w} := X = 0, Y = 1$. We get $\psi_{\mathbf{w}}\left(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right)(t) = 0 + 1 = 1$. Closure under $\mathcal{RA}^+$ queries follows from this and from [25]'s Proposition 3.5, which states that semiring homomorphisms commute with queries over $\mathcal{K}$-relations. ◀

B.2 TIDBs and BIDBs in the $\mathbb{N}[\mathbf{X}]$-encoded PDB model

Two important subclasses of $\mathbb{N}[\mathbf{X}]$-encoded PDBs that are of interest to us are the bag versions of tuple-independent databases (TIDBs) and block-independent databases (BIDBs). Under set semantics, a TIDB is a deterministic database D where each tuple t is assigned a probability p_t. The set of possible worlds represented by a TIDB D is all subsets of D. The probability of each world is the product of the probabilities of all tuples that exist with one minus the probability of all tuples of D that are not part of this world, i.e., tuples are treated as independent random events. In a BIDB, we also assign each tuple a probability,

but additionally partition D into blocks. The possible worlds of a BIDB D are all subsets of D that contain at most one tuple from each block. Note then that the tuples sharing the same block are disjoint, and the sum of the probabilitites of all the tuples in the same block b is at most 1. The probability of such a world is the product of the probabilities of all tuples present in the world. For bag TIDBs and BIDBs, we define the probability of a tuple to be the probability that the tuple exists with multiplicity at least 1.

In this work, we define TIDBs and BIDBs as subclasses of $\mathbb{N}[\mathbf{X}]$-encoded PDBs defined over variables $\mathbf{X}$ (Definition B.2) where $\mathbf{X}$ can be partitioned into blocks that satisfy the conditions of a TIDB or BIDB (stated formally in Sec. 2.1.1). In this work, we consider one further deviation from the standard: We use bag semantics for queries. Even though tuples cannot occur more than once in the input TIDB or BIDB, they can occur with a multiplicity larger than one in the result of a query. Since in TIDBs and BIDBs, there is a one-to-one correspondence between tuples in the database and variables, we can interpret a vector $\mathbf{w} \in \{0,1\}^n$ as denoting which tuples exist in the possible world $\psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})$ (the ones where $w_i = 1$). For BIDBs specifically, note that at most one of the bits corresponding to tuples in each block will be set (i.e., for any pair of bits w_j, $w_{j'}$ that are part of the same block $b_i \supseteq \{t_{i,j}, t_{i,j'}\}$, at most one of them will be set). Denote the vector $\mathbf{p}$ to be a vector whose elements are the individual probabilities p_i of each tuple t_i. Given PDB $\mathcal{D}$t $\mathcal{P}$ is the distribution induced by $\mathbf{p}$, which we will denote $\mathcal{P}^{(\mathbf{p})}$.

$$\mathop{\mathbb{E}}_{\mathbf{W}\sim\mathcal{P}^{(\mathbf{p})}}[\Phi(\mathbf{W})] = \sum_{\substack{\mathbf{w}\in\{0,1\}^n \\ s.t. w_j, w_{j'}=1 \rightarrow \nexists b_i \supseteq \{t_{i,j}, t_{i',j}\}}} \Phi(\mathbf{w}) \prod_{\substack{j\in[n] \\ s.t. w_j=1}} p_j \prod_{\substack{j\in[n] \\ s.t. w_j=0}} (1-p_i) \tag{5}$$

Recall that tuple blocks in a TIDB always have size 1, so the outer summation of eq. (5) is over the full set of vectors.

Boris says: Oliver's conjecture: Bag-TIDBs + Q can express any finite bag-PDB: A well-known result for set semantics PDBs is that while not all finite PDBs can be encoded as TIDBs, any finite PDB can be encoded using a TIDB and a query. An analog result holds in our case: any finite $\mathbb{N}$-PDB can be encoded as a bag TIDB and a query (WHAT CLASS? ADD PROOF)

B.3 Proof of Proposition 2.5

Proof. We need to prove for $\mathbb{N}$-PDB $\mathcal{D} = (\Omega, \mathcal{P})$ and $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]} = (D'_{\mathbb{N}[\mathbf{X}]}, \mathcal{P}')$ where $Mod(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = \mathcal{D}$ that $\mathbb{E}_{\mathbf{D}\sim\mathcal{P}}[Q(D)(t)] = \mathbb{E}_{\mathbf{W}\sim\mathcal{P}'}\left[\Phi[Q, D_{\mathbb{N}[\mathbf{X}]}, t](\mathbf{W})\right]$ By expanding $\Phi[Q, D_{\mathbb{N}[\mathbf{X}]}, t]$ and the expectation we have:

$$\mathop{\mathbb{E}}_{\mathbf{W}\sim\mathcal{P}'}[\Phi(\mathbf{W})] = \sum_{\mathbf{w}\in\{0,1\}^n} Pr(\mathbf{w}) \cdot Q(D_{\mathbb{N}[\mathbf{X}]})(t)(\mathbf{w})$$

From $Mod(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = \mathcal{D}$, we have that the range of $\psi_{\mathbf{w}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})}$ is Ω, so

$$= \sum_{D\in\Omega} \sum_{\mathbf{w}\in\{0,1\}^n : \psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})=D} Pr(\mathbf{w}) \cdot Q(D_{\mathbb{N}[\mathbf{X}]})(t)(\mathbf{w})$$

The inner sum is only over $\mathbf{w}$ where $\psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = D$ (i.e., $Q(D_{\mathbb{N}[\mathbf{X}]})(t)(\mathbf{w}) = Q(D)(t)$)

$$= \sum_{D\in\Omega} \sum_{\mathbf{w}\in\{0,1\}^n : \psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})=D} Pr(\mathbf{w}) \cdot Q(D)(t)$$

By distributivity of $+$ over $\times$

$$= \sum_{D\in\Omega} Q(D)(t) \sum_{\mathbf{w}\in\{0,1\}^n : \psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})=D} Pr(\mathbf{w})$$

From the definition of $\mathcal{P}$ in definition B.2, given $Mod(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = \mathcal{D}$, we get

$$= \sum_{D\in\Omega} Q(D)(t) \cdot Pr(D) \quad = \mathop{\mathbb{E}}_{\mathbf{D}\sim\mathcal{P}}[Q(D)(t)]$$

◀

B.4 Proposition B.4

Note the following fact:

▶ **Proposition B.4.** *For any BIDB-lineage polynomial* $\Phi(X_1, \ldots, X_n)$ *and all* $\mathbf{w}$ *such that* $Pr[\mathbf{W} = \mathbf{w}] > 0$*, it holds that* $\Phi(\mathbf{w}) = \widetilde{\Phi}(\mathbf{w})$.

Proof. Note that any Φ in factorized form is equivalent to its SMB expansion. For each term in the expanded form, further note that for all $b \in \{0, 1\}$ and all $e \geq 1$, $b^e = b$. Finally, note that there are exactly three cases where the expectation of a monomial term $\mathbb{E}\left[c_{\mathbf{d}} \prod_{i=n\ s.t.\ \mathbf{d}_i \geq 1} X_i\right]$ is zero: (i) when $c_{\mathbf{d}} = 0$, (ii) when $p_i = 0$ for some i where $\mathbf{d}_i \geq 1$, and (iii) when X_i and X_j are in the same block for some i, j where $\mathbf{d}_i, \mathbf{d}_j \geq 1$. ◀

B.5 Proof for Lemma 1.4

Proof. Let Φ be a polynomial of n variables with highest degree $= B$, defined as follows:

$$\Phi(X_1, \ldots, X_n) = \sum_{\mathbf{d} \in \{0,\ldots,B\}^n} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} X_i^{d_i}.$$

Let the boolean function $\textsc{isInd}(\cdot)$ take $\mathbf{d}$ as input and return true if there does not exist any dependent variables in $\mathbf{d}$, i.e., $\nexists\ b, i \neq j \mid d_{b,i}, d_{b,j} \geq 1$.[8]. Then in expectation we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{W}}[\Phi(\mathbf{W})] &= \mathop{\mathbb{E}}_{\mathbf{W}}\left[\sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} W_i^{d_i} + \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \neg\textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} W_i^{d_i}\right] && (6) \\
&= \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} W_i^{d_i}\right] + \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \neg\textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} W_i^{d_i}\right] && (7) \\
&= \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} W_i^{d_i}\right] && (8) \\
&= \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} \mathop{\mathbb{E}}_{\mathbf{W}}\left[W_i^{d_i}\right] && (9) \\
&= \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} \mathop{\mathbb{E}}_{\mathbf{W}}[W_i] && (10) \\
&= \sum_{\substack{\mathbf{d} \in \{0,\ldots,B\}^n \\ \wedge\ \textsc{isInd}(\mathbf{d})}} c_{\mathbf{d}} \cdot \prod_{\substack{i=1 \\ s.t. d_i \geq 1}}^{n} p_i && (11) \\
&= \widetilde{\Phi}(p_1, \ldots, p_n). && (12)
\end{aligned}$$

[8] This BIDB notation is used and discussed in sec. 2.1.1

Eq. (6) is the result of substituting in the definition of Φ given above. Then we arrive at eq. (7) by linearity of expectation. Next, eq. (8) is the result of the independence constraint of BIDBs, specifically that any monomial composed of dependent variables, i.e., variables from the same block b, has a probability of 0. Eq. (9) is obtained by the fact that all variables in each monomial are independent, which allows for the expectation to be pushed through the product. In eq. (10), since $W_i \in \{0,1\}$ it is the case that for any exponent $e \geq 1$, $W_i^e = W_i$. Next, in eq. (11) the expectation of a tuple is indeed its probability.

Finally, it can be verified that Eq. (12) follows since eq. (11) satisfies the construction of $\widetilde{\Phi}(p_1, \ldots, p_n)$ in Definition 1.3. ◀

B.6 Proof For Corollary 2.4

Proof. Note that Lemma 1.4 shows that $\mathbb{E}[\Phi] = \widetilde{\Phi}(p_1, \ldots, p_n)$. Therefore, if Φ is already in SMB form, one only needs to compute $\Phi(p_1, \ldots, p_n)$ ignoring exponent terms (note that such a polynomial is $\widetilde{\Phi}(p_1, \ldots, p_n)$), which indeed has $O(|\Phi|)$ computations. ◀

C Missing details from Section 3

C.1 Lemma C.1

▶ **Lemma C.1.** *Assuming that each $v \in V$ has degree ≥ 1,*[9] *the* PDB *relations encoding the edges for Φ_G^k of Definition 3.4 can be computed in $O(m)$ time.*

Proof of Lemma C.1. Only two relations need be constructed, one for the set V and one for the set E. By a simple linear scan, each can be constructed in time $O(m+n)$. Given that the degree of each $v \in V$ is at least 1, we have that $m \geq \Omega(n)$, and this yields the claimed runtime. ◀

C.2 Proof of Lemma 3.5

Proof. By the recursive defintion of $T_{det}(\cdot,\cdot)$ (see Sec. 2.3), we have the following equation for our hard query Q when $k = 1$, (we denote this as Q^1).

$$T_{det}\left(Q^1, D\right) = |D.V| + |D.E| + |D.V| + T_{join}(D.V, D.E, D.V).$$

We argue that $T_{join}(D.V, D.E, D.V)$ is at most $O(m)$ by noting that there exists an algorithm that computes $D.V \bowtie D.E \bowtie D.V$ in the same runtime[10]. Then by the assumption of Lemma C.1 (each $v \in V$ has degree ≥ 1), the sum of the first three terms is $O(m)$. We then obtain that $T_{det}\left(Q^1, D\right) = O(m) + O(m) = O(m)$. For $Q^k = Q_1^1 \times \cdots \times Q_k^1$, we have the recurrence $T_{det}\left(Q^k, D\right) = T_{det}\left(Q_1^1, D\right) + \cdots + T_{det}\left(Q_k^1, D\right) + T_{join}(Q_1^1, \cdots, Q_k^1)$. Since Q^1 outputs a count, computing the join $Q_1^1 \bowtie \cdots \bowtie Q_k^1$ is just multiplying k numbers, which takes $O(k)$ time. Thus, we have

$$T_{det}\left(Q^k, D\right) \leq k \cdot O(m) + O(k) \leq O(km),$$

as desired. ◀

[9] This is WLOG, since any vertex with degree 0 can be dropped without affecting the result of our hard query.

[10] Indeed the trivial algorithm that computes the obvious pair-wise joins has the claimed runtime. That is, we first compute $D.V \bowtie D.E$, which takes $O(m)$ (assuming $D.V$ is stored in hash map) since tuples in $D.V$ can only filter tuples in $D.E$. The resulting subset of tuples in $D.E$ are then again joined (on the right) with $D.V$, which by the same argument as before also takes $O(m)$ time, as desried.

C.3 Lemma C.2

The following lemma reduces the problem of counting k-matchings in a graph to our problem (and proves Theorem 3.6):

▶ **Lemma C.2.** *Let $p_0, \ldots, p_{2k}$ be distinct values in $(0, 1]$. Then given the values $\widetilde{\Phi}^k_G(p_i, \ldots, p_i)$ for $0 \leq i \leq 2k$, the number of k-matchings in G can be computed in $O\left(k^3\right)$ time.*

C.4 Proof of Lemma C.2

Proof. We first argue that $\widetilde{\Phi}^k_G(p, \ldots, p) = \sum_{i=0}^{2k} c_i \cdot p^i$. First, since $\Phi_G(\mathbf{X})$ has degree 2, it follows that $\Phi^k_G(\mathbf{X})$ has degree $2k$. By definition, $\widetilde{\Phi}^k_G(\mathbf{X})$ sets every exponent $e > 1$ to $e = 1$, which means that $\text{DEG}(\widetilde{\Phi}^k_G) \leq \text{DEG}(\Phi^k_G) = 2k$. Thus, if we think of p as a variable, then $\widetilde{\Phi}^k_G(p, \ldots, p)$ is a univariate polynomial of degree at most $\text{DEG}(\widetilde{\Phi}^k_G) \leq 2k$. Thus, we can write

$$\widetilde{\Phi}^k_G(p, \ldots, p) = \sum_{i=0}^{2k} c_i p^i$$

We note that c_i is *exactly* the number of monomials in the SMB expansion of $\Phi^k_G(\mathbf{X})$ composed of i distinct variables.[11]

Given that we then have $2k + 1$ distinct values of $\widetilde{\Phi}^k_G(p, \ldots, p)$ for $0 \leq i \leq 2k$, it follows that we have a linear system of the form $\mathbf{M} \cdot \mathbf{c} = \mathbf{b}$ where the ith row of $\mathbf{M}$ is $\left(p_i^0 \ldots p_i^{2k}\right)$, $\mathbf{c}$ is the coefficient vector $(c_0, \ldots, c_{2k})$, and $\mathbf{b}$ is the vector such that $\mathbf{b}[i] = \widetilde{\Phi}^k_G(p_i, \ldots, p_i)$. In other words, matrix $\mathbf{M}$ is the Vandermonde matrix, from which it follows that we have a matrix with full rank (the p_i's are distinct), and we can solve the linear system in $O(k^3)$ time (e.g., using Gaussian Elimination) to determine $\mathbf{c}$ exactly. Thus, after $O(k^3)$ work, we know $\mathbf{c}$ and in particular, c_{2k} exactly.

Next, we show why we can compute $\#\left(G, \wr \cdots \wr^k\right)$ from c_{2k} in $O(1)$ additional time. We claim that c_{2k} is $k! \cdot \#\left(G, \wr \cdots \wr^k\right)$. This can be seen intuitively by looking at the expansion of the original factorized representation

$$\Phi^k_G(\mathbf{X}) = \sum_{(i_1, j_1), \cdots, (i_k, j_k) \in E} X_{i_1} X_{j_1} \cdots X_{i_k} X_{j_k},$$

where a unique k-matching in the multi-set of product terms can be selected $\prod_{i=1}^{k} i = k!$ times. Indeed, note that each k-matching $(i_1, j_1) \ldots (i_k, j_k)$ in G corresponds to the monomial $\prod_{\ell=1}^{k} X_{i_\ell} X_{j_\ell}$ in $\Phi^k_G(\mathbf{X})$, with distinct indexes, and this implies that each distinct k-matching appears the exact number of permutations that exist for its particular set of k edges, or $k!$.

Since, as noted earlier, c_{2k} represents the number of monomials with $2k$ distinct variables, then it must be that c_{2k} is the overall number of k-matchings. And since we have $k!$ copies of each distinct k-matching, it follows that $c_{2k} = k! \cdot \#\left(G, \wr \cdots \wr^k\right)$. Thus, simply dividing c_{2k} by $k!$ gives us $\#\left(G, \wr \cdots \wr^k\right)$, as needed. ◀

C.5 Proof of Theorem 3.6

Proof. For the sake of contradiction, assume we can solve our problem in $o\left(T_{match}\left(k, G\right)\right)$ time. Given a graph G by Lemma C.1 we can compute the PDB encoding in $O\left(m\right)$ time. Then

[11] Since $\widetilde{\Phi}^k_G(\mathbf{X})$ does not have any monomial with degree < 2, it is the case that $c_0 = c_1 = 0$ but for the sake of simplcity we will ignore this observation.

after we run our algorithm on $\widetilde{\Phi}^k_G$, we get $\widetilde{\Phi}^k_G(p_i, \ldots, p_i)$ for every $0 \leq i \leq 2k$ in additional $O(k) \cdot o(T_{match}(k, G))$ time. Lemma C.2 then computes the number of k-matchings in G in $O(k^3)$ time. Adding the runtime of all of these steps, we have an algorithm for computing the number of k-matchings that runs in time

$$O(m) + O(k) \cdot o(T_{match}(k, G)) + O(k^3) \tag{13}$$

$$\leq o(T_{match}(k, G)). \tag{14}$$

We obtain Eq. (14) from the facts that k is fixed (related to m) and the assumption that $T_{match}(k, G) \geq \omega(m)$. Thus we obtain the contradiction that we can achieve a runtime $o(T_{match}(k, G))$ that is better than the optimal time $T_{match}(k, G)$ required to compute k-matchings. ◀

C.6 Subgraph Notation and $O(1)$ Closed Formulas

We need all the possible edge patterns in an arbitrary G with at most three distinct edges. We have already seen , and  , so we define the remaining patterns:

- Single Edge ()
- 2-path ()
- 2-matching ()
- 3-star ()–this is the graph that results when all three edges share exactly one common endpoint. The remaining endpoint for each edge is disconnected from any endpoint of the remaining two edges.
- Disjoint Two-Path ()–this subgraph consists of a two-path and a remaining disjoint edge.

For any graph G, the following formulas for $\#(G, H)$ compute their respective patterns exactly in $O(m)$ time, with d_i representing the degree of vertex i (proofs are in Appendix C.7):

$$\#(G, \text{}) = m, \tag{15}$$

$$\#(G, \text{}) = \sum_{i \in V} \binom{d_i}{2} \tag{16}$$

$$\#(G, \text{}) = \sum_{(i,j) \in E} \frac{m - d_i - d_j + 1}{2} \tag{17}$$

$$\#(G, \text{}) = \sum_{i \in V} \binom{d_i}{3} \tag{18}$$

$$\#(G, \text{}) + 3\#(G, \text{}) = \sum_{(i,j) \in E} \binom{m - d_i - d_j + 1}{2} \tag{19}$$

$$\#(G, \text{}) + 3\#(G, \text{}) = \sum_{(i,j) \in E} (d_i - 1) \cdot (d_j - 1) \tag{20}$$

C.7 Proofs of Eq. (15)-Eq. (20)

The proofs for Eq. (15), Eq. (16) and Eq. (18) are immediate.

Proof of Eq. (17). For edge (i, j) connecting arbitrary vertices i and j, finding all other edges in G disjoint to (i, j) is equivalent to finding all edges that are not connected to either

vertex i or j. The number of such edges is $m - d_i - d_j + 1$, where we add 1 since edge (i, j) is removed twice when subtracting both d_i and d_j. Since the summation is iterating over all edges such that a pair $((i, j), (k, \ell))$ will also be counted as $((k, \ell), (i, j))$, division by 2 then eliminates this double counting. Note that m and d_i for all $i \in V$ can be computed in one pass over the set of edges by simply maintaining counts for each quantity. Finally, the summation is also one traversal through the set of edges where each operation is either a lookup ($O(1)$ time) or an addition operation (also $O(1)$) time. ◀

Proof of Eq. (19). Eq. (19) is true for similar reasons. For edge (i, j), it is necessary to find two additional edges, disjoint or connected. As in our argument for Eq. (17), once the number of edges disjoint to (i, j) have been computed, then we only need to consider all possible combinations of two edges from the set of disjoint edges, since it doesn't matter if the two edges are connected or not. Note, the factor 3 of is necessary to account for the triple counting of 3-matchings, since it is indistinguishable to the closed form expression which of the remaining edges are either disjoint or connected to each of the edges in the *initial* set of edges disjoint to the edge under consideration. Observe that the disjoint case will be counted 3 times since each edge of a 3-path is visited once, and the same 3-path counted in each visitation. For the latter case however, it is true that since the two path in  is connected, there will be no multiple counting by the fact that the summation automatically disconnects the current edge, meaning that a two matching at the current vertex under consideration will not be counted. Thus,  will only be counted once, precisely when the single disjoint edge is visited in the summation. The sum over all such edge combinations is precisely then $\#\left(G, \right.$ $\left.\right) + 3\#\left(G, \right.$ $\left.\right)$. Note that all factorials can be computed in $O(m)$ time, and then each combination $\binom{n}{2}$ can be performed with constant time operations, yielding the claimed $O(m)$ run time. ◀

Proof of Eq. (20). To compute $\#\left(G, \right.$  $\left.\right)$, note that for an arbitrary edge (i, j), a 3-path exists for edge pair (i, ℓ) and (j, k) where i, j, k, ℓ are distinct. Further, the quantity $(d_i - 1) \cdot (d_j - 1)$ represents the number of 3-edge subgraphs with middle edge (i, j) and outer edges $(i, \ell), (j, k)$ such that $\ell \neq j$ and $k \neq i$. When $k = \ell$, the resulting subgraph is a triangle, and when $k \neq \ell$, the subgraph is a 3-path. Summing over all edges (i, j) gives Eq. (20) by observing that each triangle is counted thrice, while each 3-path is counted just once. For reasons similar to Eq. (17), all d_i can be computed in $O(m)$ time and each summand can then be computed in $O(1)$ time, yielding an overall $O(m)$ run time. ◀

C.8 Tools to prove Theorem 3.7

Note that $\widetilde{\Phi}^3_G(p, \ldots, p)$ as a polynomial in p has degree at most six. Next, we figure out the exact coefficients since this would be useful in our arguments:

▶ **Lemma C.3.** *For any p, we have:*

$$\widetilde{\Phi}^3_G(p, \ldots, p) = \#\left(G, \right. \text{  } \left.\right) p^2 + 6\#\left(G, \right. \text{  } \left.\right) p^3 + 6\#\left(G, \right. \text{  } \left.\right) p^4 + 6\#\left(G, \right. \text{  } \left.\right) p^3$$
$$+ 6\#\left(G, \right. \text{} \left.\right) p^4 + 6\#\left(G, \right. \text{} \left.\right) p^4 + 6\#\left(G, \right. \text{  } \left.\right) p^5 + 6\#\left(G, \right. \text{  } \left.\right) p^6. \quad (21)$$

C.8.1 Proof for Lemma C.3

Proof. By definition we have that

$$\Phi^3_G(\mathbf{X}) = \sum_{(i_1,j_1),(i_2,j_2),(i_3,j_3) \in E} \prod_{\ell=1}^{3} X_{i_\ell} X_{j_\ell}.$$

Hence $\widetilde{\Phi}^3_G(\mathbf{X})$ has degree six. Note that the monomial $\prod_{\ell=1}^3 X_{i_\ell} X_{j_\ell}$ will contribute to the coefficient of p^ν in $\widetilde{\Phi}^3_G(\mathbf{X})$, where ν is the number of distinct variables in the monomial. Let $e_1 = (i_1, j_1), e_2 = (i_2, j_2)$, and $e_3 = (i_3, j_3)$. We compute $\widetilde{\Phi}^3_G(\mathbf{X})$ by considering each of the three forms that the triple (e_1, e_2, e_3) can take.

CASE 1: $e_1 = e_2 = e_3$ (all edges are the same). When we have that $e_1 = e_2 = e_3$, then the monomial corresponds to $\#\left(G, \right)$. There are exactly m such triples, each with a p^2 factor in $\widetilde{\Phi}^3_G(p, \ldots, p)$.

CASE 2: This case occurs when there are two distinct edges of the three, call them e and e'. When there are two distinct edges, there is then the occurence when 2 variables in the triple (e_1, e_2, e_3) are bound to e. There are three combinations for this occurrence in $\Phi^3_G(\mathbf{X})$. Analogusly, there are three such occurrences in $\Phi^3_G(\mathbf{X})$ when there is only one occurrence of e, i.e. 2 of the variables in (e_1, e_2, e_3) are e'. This implies that all $3 + 3 = 6$ combinations of two distinct edges e and e' contribute to the same monomial in $\widetilde{\Phi}^3_G$. Since $e \neq e'$, this case produces the following edge patterns: , , which contribute $6p^3$ and $6p^4$ respectively to $\widetilde{\Phi}^3_G(p, \ldots, p)$.

CASE 3: All e_1, e_2 and e_3 are distinct. For this case, we have $3! = 6$ permutations of (e_1, e_2, e_3), each of which contribute to the same monomial in the SMB representation of $\Phi^3_G(\mathbf{X})$. This case consists of the following edge patterns: , , , , , which contribute $6p^3, 6p^4, 6p^4, 6p^5$ and $6p^6$ respectively to $\widetilde{\Phi}^3_G(p, \ldots, p)$. ◀

Since p is fixed, Lemma C.3 gives us one linear equation in $\#\left(G, \right)$ and $\#\left(G, \right)$ (we can handle the other counts due to equations (15)-(20)). However, we need to generate one more independent linear equation in these two variables. Towards this end we generate another graph related to G:

▶ **Definition C.4.** *For $\ell \geq 1$, let graph $G^{(\ell)}$ be a graph generated from an arbitrary graph G, by replacing every edge e of G with an ℓ-path, such that all inner vertexes of an ℓ-path replacement edge are disjoint from all other vertexes.*[12]

We will prove Theorem 3.7 by the following reduction:

▶ **Theorem C.5.** *Fix $p \in (0, 1)$. Let G be a graph on m edges. If we can compute $\widetilde{\Phi}^3_G(p, \ldots, p)$ exactly in $T(m)$ time, then we can exactly compute $\#\left(G, \right)$ in $O\left(T(m) + m\right)$ time.*

For clarity, we repeat the notion of $\#\left(G, H\right)$ to mean the count of subgraphs in G isomorphic to H. The following lemmas relate these counts in $G^{(2)}$ to $G^{(1)}$ (G). The lemmas are used to prove Lemma C.8.

▶ **Lemma C.6.** *The 3-matchings in graph $G^{(2)}$ satisfy the identity:*

$$\#\left(G^{(2)}, \right) = 8 \cdot \#\left(G^{(1)}, \right) + 6 \cdot \#\left(G^{(1)}, \right) + 4 \cdot \#\left(G^{(1)}, \right) + 4 \cdot \#\left(G^{(1)}, \right) + 2 \cdot \#\left(G^{(1)}, \right).$$

▶ **Lemma C.7.** *For $\ell > 1$ and any graph $G^{(\ell)}$, $\#\left(G^{(\ell)}, \right) = 0$.*

Finally, the following result immediately implies Theorem C.5:

▶ **Lemma C.8.** *Fix $p \in (0, 1)$. Given $\widetilde{\Phi}^3_{G^{(\ell)}}(p, \ldots, p)$ for $\ell \in [2]$, we can compute in $O(m)$ time a vector $\mathbf{b} \in \mathbb{R}^3$ such that*

$$\begin{pmatrix} 1 - 3p & -(3p^2 - p^3) \\ 10(3p^2 - p^3) & 10(3p^2 - p^3) \end{pmatrix} \cdot \begin{pmatrix} \#\left(G, \right)] \\ \#\left(G, \right) \end{pmatrix} = \mathbf{b},$$

[12] Note that $G \equiv G^{(1)}$.

allowing us to compute $\#(G, \text{⁂})$ *and* $\#(G, \text{⁞⁞⁞})$ *in* $O(1)$ *time.*

C.9 Proofs for Lemma C.6, Lemma C.7, and Lemma C.8

Before proceeding, let us introduce a few more helpful definitions.

▶ **Definition C.9** ($E^{(\ell)}$). *For* $\ell > 1$*, we use* $E^{(\ell)}$ *to denote the set of edges in* $G^{(\ell)}$*. For any graph* $G^{(\ell)}$*, its edges are denoted by the a pair* (e, b)*, such that* $b \in \{0, \ldots, \ell - 1\}$ *where* $(e, 0), \ldots, (e, \ell - 1)$ *is the* ℓ*-path that replaces the edge* e *for* $e \in E^{(1)}$*.*

▶ **Definition C.10** ($E_S^{(\ell)}$). *Given an arbitrary subgraph* $S^{(1)}$ *of* $G^{(1)}$*, let* $E_S^{(1)}$ *denote the set of edges in* $S^{(1)}$*. Define then* $E_S^{(\ell)}$ *for* $\ell > 1$ *as the set of edges in the generated subgraph* $S^{(\ell)}$ *(i.e. when we apply Definition C.4 to* S *to generate* $S^{(\ell)}$*).*

For example, consider $S^{(1)}$ with edges $E_S^{(1)} = \{e_1\}$. Then the edge set of $S^{(2)}$ is defined as $E_S^{(2)} = \{(e_1, 0), (e_1, 1)\}$.

▶ **Definition C.11** ($\binom{E}{t}$ and $\binom{E}{\leq t}$). *Let* $\binom{E}{t}$ *denote the set of subsets in* E *with exactly* t *edges. In a similar manner,* $\binom{E}{\leq t}$ *is used to mean the subsets of* E *with* t *or fewer edges.*

The following function f_ℓ is a mapping from every 3-edge shape in $G^{(\ell)}$ to its 'projection' in $G^{(1)}$.

▶ **Definition C.12.** *Let* $f_\ell : \binom{E^{(\ell)}}{3} \to \binom{E^{(1)}}{\leq 3}$ *be defined as follows. For any element* $s \in \binom{E^{(\ell)}}{3}$ *such that* $s = \{(e_1, b_1), (e_2, b_2), (e_3, b_3)\}$*, define:*

$$f_\ell\left(\{(e_1, b_1), (e_2, b_2), (e_3, b_3)\}\right) = \{e_1, e_2, e_3\}\,.$$

▶ **Definition C.13** (f_ℓ^{-1}). *For an arbitrary subgraph* $S^{(1)}$ *of* $G^{(1)}$ *with at most* $m \leq 3$ *edges, the inverse function* $f_\ell^{-1} : \binom{E^{(1)}}{\leq 3} \to 2^{\binom{E^{(\ell)}}{3}}$ *takes* $E_S^{(1)}$ *and outputs the set of all elements* $s \in \binom{E_S^{(\ell)}}{3}$ *such that* $f_\ell(s) = E_S^{(1)}$*.*

Note, importantly, that when we discuss f_ℓ^{-1}, that each *edge* present in $E_S^{(1)}$ must have an edge in $s \in f_\ell^{-1}(E_S^{(1)})$ that projects down to it. In particular, if $|E_S^{(1)}| = 3$, then it must be the case that each $s \in f_\ell^{-1}(E_S^{(1)})$ consists of the following set of edges: $\{(e_i, b), (e_j, b'), (e_m, b'')\}$, where i, j and m are distinct.

We are now ready to prove the structural lemmas. To prove the structural lemmas, we will count the number of occurrences of ⁂ and ⁞⁞⁞ in $G^{(\ell)}$ we count for each $S \in \binom{E_1}{\leq 3}$, how many ⁞⁞⁞ and ⁂ subgraphs appear in $f_\ell^{-1}(E_S^{(1)})$.

C.9.1 Proof of Lemma C.6

Proof. For each subset $E_S^{(1)} \in \binom{E_1}{\leq 3}$, we count the number of 3-*matchings* in the 3-edge subgraphs of $G^{(2)}$ in $f_2^{-1}(E_S^{(1)})$. We first consider the case of $E_S^{(1)} \in \binom{E_1}{3}$, where $E_S^{(1)}$ is composed of the edges e_1, e_2, e_3 and $f_2^{-1}(E_S^{(1)})$ is the set of all 3-edge subsets $s \in \{(e_1, 0), (e_1, 1), (e_2, 0), (e_2, 1), (e_3, 0), (e_3, 1)\}$ such that $f_\ell(s) = \{e_1, e_2, e_3\}$. The size of the output is denoted $\left|f_2^{-1}(E^{(1)})\right|$. For the case where each set of edges of the form $\{(e_1, b_1), (e_2, b_2), (e_3, b_3)\}$ for $b_i \in [2], i \in [3]$ is present, we have $\left|f_2^{-1}(E^{(1)})\right| = 8$. We count the number of 3-matchings from the set $f_2^{-1}(E_S^{(1)})$.

We do a case analysis based on the subgraph $S^{(1)}$ induced by $E_S^{(1)}$.

- 3-matching ( )

When $S^{(1)}$ is isomorphic to , it is the case that edges in $E_S^{(2)}$ are *not* disjoint only for the pairs $(e_i, 0), (e_i, 1)$ for $i \in \{1, 2, 3\}$. By definition, each set of edges in $f_2^{-1}\left(E_S^{(1)}\right)$ is a three matching and $\left|f_2^{-1}\left(E_S^{(1)}\right)\right| = 8$ possible 3-matchings.

- Disjoint Two-Path ( )

For $S^{(1)}$ isomorphic to edges e_2, e_3 form a 2-path with e_1 being disjoint. This means that in $S^{(2)}$ edges $(e_2, 0), (e_2, 1), (e_3, 0), (e_3, 1)$ form a 4-path while $(e_1, 0), (e_1, 1)$ is its own disjoint 2-path. We can pick either $(e_1, 0)$ or $(e_1, 1)$ for the first edge in the 3-matching, while it is necessary to have a 2-matching from path $(e_2, 0), \ldots (e_3, 1)$. Note that the 4-path allows for three possible 2-matchings, specifically,

$$\{(e_2, 0), (e_3, 0)\}, \{(e_2, 0), (e_3, 1)\}, \{(e_2, 1), (e_3, 1)\}.$$

Since these two selections can be made independently, $\left|f_2^{-1}\left(E_S^{(1)}\right)\right| = 2 \cdot 3 = 6$ *distinct* 3-matchings in $f_2^{-1}(E_S^{(1)})$.

- 3-star ( )

When $S^{(1)}$ is isomorphic to , the inner edges $(e_i, 1)$ of $S^{(2)}$ are all connected, and the outer edges $(e_i, 0)$ are all disjoint. Note that for a valid 3-matching it must be the case that at most one inner edge can be part of the set of disjoint edges. For the case of when exactly one inner edge is chosen, there exist 3 possiblities, based on which inner edge is chosen. Note that if $(e_i, 1)$ is chosen, the matching has to choose $(e_j, 0)$ for $j \neq i$ and $(e_{j'}, 0)$ for $j' \neq i, j' \neq j$. The remaining possible 3-matching occurs when all 3 outer edges are chosen, and $\left|f_2^{-1}\left(E_S^{(1)}\right)\right| = 4$.

- 3-path ( )

When $S^{(1)}$ is isomorphic to it is the case that all edges beginning with e_1 and ending with e_3 are successively connected. This means that the edges of $E_S^{(2)}$ form a 6-path. For a 3-matching to exist in $f_2^{-1}(E_S^{(1)})$, we cannot pick both $(e_i, 0)$ and $(e_i, 1)$ or both $(e_i, 1)$ and $(e_j, 0)$ where $j = i + 1$. There are four such possibilities: $\{(e_1, 0), (e_2, 0), (e_3, 0)\}$, $\{(e_1, 0), (e_2, 0), (e_3, 1)\}$, $\{(e_1, 0), (e_2, 1), (e_3, 1)\}$, $\{(e_1, 1), (e_2, 1), (e_3, 1)\}$ and $\left|f_2^{-1}\left(E_S^{(1)}\right)\right| = 4$.

- Triangle ( )

For $S^{(1)}$ isomorphic to , note that it is the case that the edges in $E_S^{(2)}$ are connected in a successive manner, but this time in a cycle, such that $(e_1, 0)$ and $(e_3, 1)$ are also connected. While this is similar to the discussion of the three path above, the first and last edges are not disjoint.This rules out both subsets of $(e_1, 0), (e_2, 0), (e_3, 1)$ and $(e_1, 0), (e_2, 1), (e_3, 1)$, so that $\left|f_2^{-1}\left(E_S^{(1)}\right)\right| = 2$.

Let us now consider when $E_S^{(1)} \in \binom{E_1}{\leq 2}$, i.e. fixed subgraphs among

- 2-matching ( ), 2-path ( ), 1 edge ( )

When $|E_S^{(1)}| = 2$, we can only pick one from each of two pairs, $\{(e_1, 0), (e_1, 1)\}$ and $\{(e_2, 0), (e_2, 1)\}$. The third edge choice in $E_S^{(2)}$ will break the disjoint property of a 3-matching. Thus, a 3-matching cannot exist in $f_2^{-1}(E_S^{(1)})$. A similar argument holds for $|E_S^{(1)}| = 1$, where the output of f_2^{-1} is $\{\emptyset\}$ since there are not enough edges in the input to produce any other output.

Observe that all of the arguments above focused solely on the property of subgraph $S^{(1)}$ being isomorphmic. In other words, all $E_S^{(1)}$ of a given "shape" yield the same number of 3-matchings in $f_2^{-1}(E_S^{(1)})$, and this is why we get the required identity using the above case analysis. ◀

C.9.2 Proof of Lemma C.7

Proof. The number of triangles in $G^{(\ell)}$ for $\ell \geq 2$ will always be 0 for the simple fact that all cycles in $G^{(\ell)}$ will have at least six edges. ◀

C.9.3 Proof of Lemma C.8

Proof. The proof consists of two parts. First we need to show that a vector $\mathbf{b}$ satisfying the linear system exists and further can be computed in $O(m)$ time. Second we need to show that $\#(G, \text{triangle}), \#(G, \text{3-matching})$ can indeed be computed in time $O(1)$.

The lemma claims that for $\mathbf{M} = \begin{pmatrix} 1-3p & -(3p^2-p^3) \\ 10(3p^2-p^3) & 10(3p^2-p^3) \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} \#(G, \text{triangle})] \\ \#(G, \text{3-matching}) \end{pmatrix}$ satisfies the linear system $\mathbf{M} \cdot \mathbf{x} = \mathbf{b}$.

To prove the first step, we use Lemma C.3 to derive the following equality (dropping the superscript and referring to $G^{(1)}$ as G):

$$\begin{aligned} &\#(G, \text{edge})\, p^2 + 6\#(G, \text{2-path})\, p^3 + 6\#(G, \text{2-matching})\, p^4 + 6\#(G, \text{triangle})\, p^3 + 6\#(G, \text{3-star})\, p^4 \\ &\quad + 6\#(G, \text{3-path})\, p^4 + 6\#(G, \text{edge + 2-path})\, p^5 + 6\#(G, \text{3-matching})\, p^6 = \widetilde{\Phi}^3_G(p, \ldots, p) \end{aligned} \tag{22}$$

$$\begin{aligned} &\#(G, \text{triangle}) + \#(G, \text{3-path})\, p + \#(G, \text{edge + 2-path})\, p^2 + \#(G, \text{3-matching})\, p^3 \\ &\quad = \frac{\widetilde{\Phi}^3_G(p, \ldots, p)}{6p^3} - \frac{\#(G, \text{edge})}{6p} - \#(G, \text{2-path}) - \#(G, \text{2-matching})\, p - \#(G, \text{3-star})\, p \end{aligned} \tag{23}$$

$$\begin{aligned} &\#(G, \text{triangle})\,(1-3p) - \#(G, \text{3-matching})\,(3p^2 - p^3) = \\ &\quad \frac{\widetilde{\Phi}^3_G(p, \ldots, p)}{6p^3} - \frac{\#(G, \text{edge})}{6p} - \#(G, \text{2-path}) - \#(G, \text{2-matching})\, p - \#(G, \text{3-star})\, p \\ &\quad - [\#(G, \text{3-path})\, p + 3\#(G, \text{triangle})\, p] - \left[\#(G, \text{edge + 2-path})\, p^2 + 3\#(G, \text{3-matching})\, p^2\right] \end{aligned} \tag{24}$$

Eq. (22) is the result of Lemma C.3. We obtain the remaining equations through standard algebraic manipulations.

Note that the LHS of Eq. (24) is obtained using eq. (19) and eq. (20) and is indeed the product $\mathbf{M}[1] \cdot \mathbf{x}[1]$. Further note that this product is equal to the RHS of Eq. (24), where every term is computable in $O(m)$ time (by equations (15)-(20)). We set $\mathbf{b}[1]$ to the RHS of Eq. (24).

We follow the same process in deriving an equality for $G^{(2)}$. Replacing occurrences of G with $G^{(2)}$, we obtain an equation (below) of the form of eq. (24) for $G^{(2)}$. Substituting

identities from lemma C.6 and Lemma C.7 we obtain

$$0-(8\#\left(G,\ \right)+6\#\left(G,\ \right)+4\#\left(G,\ \right)+4\#\left(G,\ \right)+2\#\left(G,\ \right))(3p^2-p^3)=$$
$$\frac{\widetilde{\Phi}^3_{G^{(2)}}(p,\ldots,p)}{6p^3}-\frac{\#\left(G^{(2)},\ \right)}{6p}-\#\left(G^{(2)},\ \right)-\#\left(G^{(2)},\ \right)p-\#\left(G^{(2)},\ \right)p$$
$$-\left[\#\left(G^{(2)},\ \right)p^2+3\#\left(G^{(2)},\ \right)p^2\right]-\left[\#\left(G^{(2)},\ \right)p+3\#\left(G^{(2)},\ \right)p\right] \tag{25}$$

$$(10\#\left(G,\ \right)+10G\ )(3p^2-p^3)=$$
$$\frac{\widetilde{\Phi}^3_{G^{(2)}}(p,\ldots,p)}{6p^3}-\frac{\#\left(G^{(2)},\ \right)}{6p}-\#\left(G^{(2)},\ \right)-\#\left(G^{(2)},\ \right)p-\#\left(G^{(2)},\ \right)p$$
$$-\left[\#\left(G^{(2)},\ \right)p+3\#\left(G^{(2)},\ \right)p\right]-\left[\#\left(G^{(2)},\ \right)p^2-3\#\left(G^{(2)},\ \right)p^2\right]$$
$$+(4\#\left(G,\ \right)+[6\#\left(G,\ \right)+18\#\left(G,\ \right)]+[4\#\left(G,\ \right)+12\#\left(G,\ \right)])(3p^2-p^3) \tag{26}$$

The steps to obtaining eq. (26) are analogous to the derivation immediately preceding. As in the previous derivation, note that the LHS of Eq. (26) is the same as $\mathbf{M}[2]\cdot\mathbf{x}[2]$. The RHS of Eq. (26) has terms all computable (by equations (15)-(20)) in $O(m)$ time. Setting $\mathbf{b}[2]$ to the RHS then completes the proof of step 1.

Note that if $\mathbf{M}$ has full rank then one can compute $\#\left(G,\ \right)$ and $\#\left(G,\ \right)$ in $O(1)$ using Gaussian elimination.

To show that $\mathbf{M}$ indeed has full rank, we show in what follows that $Det\left(\mathbf{M}\right)\neq 0$ for every $p\in(0,1)$. $Det\left(\mathbf{M}\right)=$

$$\begin{vmatrix}1-3p & -(3p^2-p^3)\\ 10(3p^2-p^3) & 10(3p^2-p^3)\end{vmatrix}=(1-3p)\cdot 10(3p^2-p^3)+10(3p^2-p^3)\cdot(3p^2-p^3)$$
$$=10(3p^2-p^3)\cdot(1-3p+3p^2-p^3)=10(3p^2-p^3)\cdot(-p^3+3p^2-3p+1)$$
$$=10p^2(3-p)\cdot(1-p)^3 \tag{27}$$

From Eq. (27) it can easily be seen that the roots of $Det\left(\mathbf{M}\right)$ are $0,1$, and 3. Hence there are no roots in $(0,1)$ and Lemma C.8 follows. ◀

C.10 Proof of Theorem C.5

Proof. We can compute $G^{(2)}$ from $G^{(1)}$ in $O(m)$ time. Additionally, if in time $O(T(m))$, we have $\widetilde{\Phi}^3_{G^{(\ell)}}(p,\ldots,p)$ for $\ell\in[2]$, then the theorem follows by Lemma C.8. ◀

In other words, if Theorem C.5 holds, then so must Theorem 3.7.

C.11 Proof of Theorem 3.7

Proof. For the sake of contradiction, assume that for any G, we can compute $\widetilde{\Phi}^3_G(p,\ldots,p)$ in $o\left(m^{1+\epsilon_0}\right)$ time. Let G be the input graph. Then by Theorem C.5 we can compute $\#\left(G,\ \right)$ in further time $o\left(m^{1+\epsilon_0}\right)+O(m)$. Thus, the overall, reduction takes $o\left(m^{1+\epsilon_0}\right)+O(m)=o\left(m^{1+\epsilon_0}\right)$ time, which violates Conjecture 3.3. ◀

D Missing Details from Section 4

In the following definitions and examples, we use the following polynomial as an example:

$$\Phi(X,Y)=2X^2+3XY-2Y^2. \tag{28}$$

▶ **Definition D.1** (Pure Expansion). *The pure expansion of a polynomial Φ is formed by computing all product of sums occurring in Φ, without combining like monomials. The pure expansion of Φ generalizes Definition 2.1 by allowing monomials $m_i = m_j$ for $i \neq j$.*

Note that similar in spirit to **??**, $\mathtt{E}(\mathtt{C})$ Definition 4.1 reduces all variable exponents $e > 1$ to $e = 1$. Further, it is true that $\mathtt{E}(\mathtt{C})$ is the pure expansion of $\mathtt{C}$.

▶ **Example D.2** (Example of Pure Expansion). *Consider the factorized representation $(X + 2Y)(2X - Y)$ of the polynomial in Eq. (28). Its circuit $\mathtt{C}$ is illustrated in Fig. 3. The pure expansion of the product is $2X^2 - XY + 4XY - 2Y^2$. As an additional example of Definition 4.1, $\mathtt{E}(\mathtt{C}) = [(X, 2), (XY, -1), (XY, 4), (Y, -2)]$.*

$\mathtt{E}(\mathtt{C})$ effectively[13] encodes the *reduced* form of $\textsc{poly}(\mathtt{C})$, decoupling each monomial into a set of variables $\mathtt{v}$ and a real coefficient $\mathtt{c}$. However, unlike the constraint on the input Φ to compute $\widetilde{\Phi}$, the input circuit $\mathtt{C}$ does not need to be in SMB/SOP form.

▶ **Example D.3** (Example for Definition 4.2). *Using the same factorization from Example D.2, $\textsc{poly}(|\mathcal{C}|) = (X + 2Y)(2X + Y) = 2X^2 + XY + 4XY + 2Y^2 = 2X^2 + 5XY + 2Y^2$. Note that this is not the same as the polynomial from Eq. (28). As an example of the slight abuse of notation we alluded to, $\textsc{poly}(|\mathcal{C}|(1, \ldots, 1)) = 2(1)^2 + 5(1)(1) + 2(1)^2 = 9$.*

Aaron says: Verify whether we need pure expansion or not.

▶ **Definition D.4** (Subcircuit). *A subcircuit of a circuit $\mathtt{C}$ is a circuit $\mathtt{S}$ such that $\mathtt{S}$ is a DAG subgraph of the DAG representing $\mathtt{C}$. The sink of $\mathtt{S}$ has exactly one gate $\mathtt{g}$.*

The following results assume input circuit $\mathtt{C}$ computed from an arbitrary $\mathcal{RA}^+$ query Q and arbitrary BIDB $\mathcal{D}$. We refer to $\mathtt{C}$ as a BIDB circuit.

Aaron says: Verify that the proof for Theorem D.5 doesn't rely on properties of $\mathcal{RA}^+$ or BIDB.

▶ **Theorem D.5.** *Let $\mathcal{C}$ be an arbitrary* BIDB *circuit and define $\Phi(\mathbf{X}) = \textsc{poly}(\mathcal{C})$ and let $k = \textsc{deg}(\mathcal{C})$. Then an estimate $\mathcal{E}$ of $\widetilde{\Phi}(p_1, \ldots, p_n)$ can be computed in time*

$$O\left(\left(\textsc{size}(\mathcal{C}) + \frac{\log \frac{1}{\delta} \cdot |\mathcal{C}|^2 (1, \ldots, 1) \cdot k \cdot \log k \cdot \textsc{depth}(\mathcal{C}))}{(\epsilon)^2 \cdot \widetilde{\Phi}^2(p_1, \ldots, p_n)}\right) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathcal{C}|(1, \ldots, 1)\right), \log\left(\textsc{size}(\mathcal{C})\right)\right)\right)$$

such that

$$Pr\left(\left|\mathcal{E} - \widetilde{\Phi}(p_1, \ldots, p_n)\right| > \epsilon \cdot \widetilde{\Phi}(p_1, \ldots, p_n)\right) \leq \delta. \tag{29}$$

Atri says: Aaron: Just copied over from S4. The above text might need smoothening into the appendix.

The slight abuse of notation seen in $|\mathtt{C}|(1, \ldots, 1)$ is explained after Definition 4.2 and an example is given in Example D.3. The only difference in the use of this notation in Theorem D.5 is that we include an additional exponent to square the quantity.

D.1 Proof of Theorem D.5

We prove Theorem D.5 constructively by presenting an algorithm $\textsc{Approximate}\widetilde{\Phi}$ (Algorithm 1) which has the desired runtime and computes an approximation with the desired approximation

[13] The minor difference here is that $\mathtt{E}(\mathtt{C})$ encodes the *reduced* form over the SOP pure expansion of the compressed representation, as opposed to the SMB representation

Algorithm 1 ApproximateΦ̃(C, **p**, δ, ϵ)

Input: C: Circuit
Input: $\mathbf{p} = (p_1, \ldots, p_n) \in [0,1]^N$
Input: $\delta \in [0,1]$
Input: $\epsilon \in [0,1]$
Output: acc $\in \mathbb{R}$

```
acc ← 0
N ← ⌈2 log(2/δ) / ϵ²⌉
(C_mod, size) ← OnePass (C)                ▷ OnePass is Algorithm 2
for i ∈ 1 to N do                          ▷ Perform the required number of samples
    (M, sgn_i) ← SampleMonomial (C_mod)    ▷ SampleMonomial is Algorithm 3. Note that sgn_i is the sign of the monomial's coefficient and not the coefficient itself
    if M has at most one variable from each block then
        Y_i ← ∏_{X_j ∈ M} p_j              ▷ M is the sampled monomial's set of variables (cref. appendix D.8)
        Y_i ← Y_i × sgn_i
        acc ← acc + Y_i                    ▷ Store the sum over all samples
    end if
end for
acc ← acc × size / N
return acc
```

guarantee. Algorithm ApproximateΦ̃ uses Algorithm OnePass to compute weights on the edges of a circuits. These weights are then used to sample a set of monomials of Φ(C) from the circuit C by traversing the circuit using the weights to ensure that monomials are sampled with an appropriate probability. The correctness of ApproximateΦ̃ relies on the correctness (and runtime behavior) of auxiliary algorithms OnePass and SampleMonomial that we state in the following lemmas (and prove later in this part of the appendix).

▶ **Lemma D.6.** *The OnePass function completes in time:*

$$O\left(\mathit{size}(\mathcal{C}) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathcal{C}(1\ldots,1)|\right), \log \mathit{size}(\mathcal{C})\right)\right)$$

OnePass guarantees two post-conditions: First, for each subcircuit $\mathcal{S}$ of $\mathcal{C}$, we have that $\mathcal{S}.\mathtt{partial}$ is set to $|\mathcal{S}|(1,\ldots,1)$. Second, when $\mathcal{S}.\mathtt{type} = +$, $\mathcal{S}.\mathtt{Lweight} = \frac{|\mathcal{S}_L|(1,\ldots,1)}{|\mathcal{S}|(1,\ldots,1)}$ and likewise for $\mathcal{S}.\mathtt{Rweight}$.

To prove correctness of Algorithm 1, we only use the following fact that follows from the above lemma: for the modified circuit ($\mathtt{C}_{\mathtt{mod}}$) output by OnePass, $\mathtt{C}_{\mathtt{mod}}.\mathtt{partial} = |\mathtt{C}|(1,\ldots,1)$.

▶ **Lemma D.7.** *The function SampleMonomial completes in time*

$$O(\log k \cdot k \cdot \mathit{depth}(\mathcal{C}) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathcal{C}|(1,\ldots,1)\right), \log \mathit{size}(\mathcal{C})\right))$$

where $k = \mathit{deg}(\mathcal{C})$. The function returns every $(v, \mathit{sign}(c))$ for $(v, c) \in E(\mathcal{C})$ with probability $\frac{|c|}{|\mathcal{C}|(1,\ldots,1)}$.

With the above two lemmas, we are ready to argue the following result:

▶ **Theorem D.8.** *For any $\mathcal{C}$ with*

Aaron says: *Pretty sure this is $\mathit{deg}(|\mathcal{C}|)$. Have to read on to be sure.*

DEG$(poly(|\mathcal{C}|)) = k$, *algorithm 1 outputs an estimate* $\texttt{acc}$ *of* $\widetilde{\Phi}(p_1, \ldots, p_n)$ *such that*

$$Pr\left(\left|\texttt{acc} - \widetilde{\Phi}(p_1, \ldots, p_n)\right| > \epsilon \cdot |\mathcal{C}|(1, \ldots, 1)\right) \leq \delta,$$

in $O\left(\left(\mathit{SIZE}(\mathcal{C}) + \frac{\log \frac{1}{\delta}}{\epsilon^2} \cdot k \cdot \log k \cdot \mathit{DEPTH}(\mathcal{C})\right) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathcal{C}|(1, \ldots, 1)\right), \log \mathit{SIZE}(\mathcal{C})\right)\right)$ *time.*

Before proving Theorem D.8, we use it to argue the claimed runtime of our main result, Theorem D.5.

Proof of Theorem D.5. Set $\mathcal{E} = \textsc{Approximate}\widetilde{\Phi}(\mathtt{C}, (p_1, \ldots, p_n), \delta, \epsilon')$, where

$$\epsilon' = \epsilon \cdot \frac{\widetilde{\Phi}(p_1, \ldots, p_n)}{|\mathtt{C}|(1, \ldots, 1)},$$

which achieves the claimed error bound on $\mathcal{E}$ ($\texttt{acc}$) trivially due to the assignment to ϵ' and theorem D.8, since $\epsilon' \cdot |\mathtt{C}|(1, \ldots, 1) = \epsilon \cdot \frac{\widetilde{\Phi}(1, \ldots, 1)}{|\mathtt{C}|(1, \ldots, 1)} \cdot |\mathtt{C}|(1, \ldots, 1) = \epsilon \cdot \widetilde{\Phi}(1, \ldots, 1)$.

The claim on the runtime follows from Theorem D.8 since

$$\begin{aligned}\frac{1}{(\epsilon')^2} \cdot \log\left(\frac{1}{\delta}\right) &= \frac{\log \frac{1}{\delta}}{\epsilon^2 \left(\frac{\widetilde{\Phi}(p_1, \ldots, p_N)}{|\mathtt{C}|(1, \ldots, 1)}\right)^2} \\ &= \frac{\log \frac{1}{\delta} \cdot |\mathtt{C}|^2(1, \ldots, 1)}{\epsilon^2 \cdot \widetilde{\Phi}^2(p_1, \ldots, p_n)}.\end{aligned}$$

◀

Let us now prove Theorem D.8:

D.2 Proof of Theorem D.8

Proof. Consider now the random variables $\mathtt{Y}_1, \ldots, \mathtt{Y}_\mathtt{N}$, where each $\mathtt{Y}_\mathtt{i}$ is the value of $\mathtt{Y}_\mathtt{i}$ in algorithm 1 after line 8 is executed. Overloading $\textsc{isInd}(\cdot)$ to receive monomial input (recall $\mathtt{v}_\mathtt{m}$ is the monomial composed of the variables in the set $\mathtt{v}$), we have

$$\mathtt{Y}_\mathtt{i} = \mathbb{1}_{(\textsc{isInd}(\mathtt{v}_\mathtt{m}))} \cdot \prod_{X_i \in \textsc{var}(v)} p_i,$$

where the indicator variable handles the check in Line 6 Then for random variable $\mathtt{Y}_i$, it is the case that

$$\begin{aligned}\mathbb{E}\left[\mathtt{Y}_\mathtt{i}\right] &= \sum_{(\mathtt{v},\mathtt{c}) \in \mathtt{E}(\mathtt{C})} \frac{\mathbb{1}_{(\textsc{isInd}(\mathtt{v}_\mathtt{m}))} \cdot c \cdot \prod_{X_i \in \textsc{var}(v)} p_i}{|\mathtt{C}|(1, \ldots, 1)} \\ &= \frac{\widetilde{\Phi}(p_1, \ldots, p_n)}{|\mathtt{C}|(1, \ldots, 1)},\end{aligned}$$

where in the first equality we use the fact that $\mathtt{sgn}_\mathtt{i} \cdot |\mathtt{c}| = \mathtt{c}$ and the second equality follows from Eq. (2) with X_i substituted by p_i.

Let $\overline{\mathtt{Y}} = \frac{1}{N} \sum_{i=1}^{N} \mathtt{Y}_\mathtt{i}$. It is also true that

$$\mathbb{E}\left[\overline{\mathtt{Y}}\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathtt{Y}_\mathtt{i}\right] = \frac{\widetilde{\Phi}(p_1, \ldots, p_n)}{|\mathtt{C}|(1, \ldots, 1)}.$$

Hoeffding's inequality states that if we know that each $\mathtt{Y}_i$ (which are all independent) always lie in the intervals $[a_i, b_i]$, then it is true that

$$Pr\left(|\overline{\mathtt{Y}} - \mathbb{E}\left[\overline{\mathtt{Y}}\right]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right).$$

Line 5 shows that $\mathtt{sgn}_\mathtt{i}$ has a value in $\{-1, 1\}$ that is multiplied with $O(k)$ $p_i \in [0, 1]$, which implies the range for each $\mathtt{Y}_i$ is $[-1, 1]$. Using Hoeffding's inequality, we then get:

$$Pr\left(\ |\overline{\mathtt{Y}} - \mathbb{E}\left[\overline{\mathtt{Y}}\right]\ | \geq \epsilon\right) \leq 2\exp\left(-\frac{2N^2\epsilon^2}{2^2 N}\right) = 2\exp\left(-\frac{N\epsilon^2}{2}\right) \leq \delta,$$

where the last inequality dictates our choice of N in Line 2.

For the claimed probability bound of $Pr\left(\left|\mathtt{acc} - \widetilde{\Phi}(p_1, \ldots, p_n)\right| > \epsilon \cdot |\mathtt{C}|\,(1, \ldots, 1)\right) \leq \delta$, note that in the algorithm, $\mathtt{acc}$ is exactly $\overline{\mathtt{Y}} \cdot |\mathtt{C}|\,(1, \ldots, 1)$. Multiplying the rest of the terms by the additional factor $|\mathtt{C}|\,(1, \ldots, 1)$ yields the said bound.

This concludes the proof for the first claim of theorem D.8. Next, we prove the claim on the runtime.

Run-time Analysis

The runtime of the algorithm is dominated first by Line 3 (which by Lemma D.6 takes time $O\left(\textsc{size}(\mathtt{C}) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathtt{C}|\,(1, \ldots, 1)\right), \log\left(\textsc{size}(\mathtt{C})\right)\right)\right)$ and then by N iterations of the loop in Line 4. Each iteration's run time is dominated by the call to SampleMonomial in Line 5 (which by Lemma D.7 takes $O\left(\log k \cdot k \cdot \textsc{depth}(\mathtt{C}) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathtt{C}|\,(1, \ldots, 1)\right), \log\left(\textsc{size}(\mathtt{C})\right)\right)\right)$) and the check Line 6, which by the subsequent argument takes $O(k \log k)$ time. We sort the $O(k)$ variables by their block IDs and then check if there is a duplicate block ID or not. Combining all the times discussed here gives us the desired overall runtime. ◀

D.3 Proof of Theorem 4.7

Proof. The result follows by first noting that by definition of γ, we have

$$\widetilde{\Phi}(1, \ldots, 1) = (1 - \gamma) \cdot |\mathtt{C}|\,(1, \ldots, 1).$$

Further, since each $p_i \geq p_0$ and $\Phi(\mathbf{X})$ (and hence $\widetilde{\Phi}(\mathbf{X})$) has degree at most k, we have that

$$\widetilde{\Phi}(1, \ldots, 1) \geq p_0^k \cdot \widetilde{\Phi}(1, \ldots, 1).$$

The above two inequalities implies $\widetilde{\Phi}(1, \ldots, 1) \geq p_0^k \cdot (1 - \gamma) \cdot |\mathtt{C}|\,(1, \ldots, 1)$. Applying this bound in the runtime bound in Theorem D.5 gives the first claimed runtime. The final runtime of $O_k\left(\frac{1}{\epsilon^2} \cdot \textsc{size}(\mathtt{C}) \cdot \log\frac{1}{\delta} \cdot \overline{\mathcal{M}}\left(\log\left(|\mathtt{C}|\,(1, \ldots, 1)\right), \log\left(\textsc{size}(\mathtt{C})\right)\right)\right)$ follows by noting that $\textsc{depth}(\mathtt{C}) \leq \textsc{size}(\mathtt{C})$ and absorbing all factors that just depend on k. ◀

D.4 Proof of Lemma 4.8

We will prove Lemma 4.8 by considering the two cases separately. We start by considering the case when $\mathtt{C}$ is a tree:

▶ **Lemma D.9.** *Let $\mathtt{C}$ be a tree (i.e. the sub-circuits corresponding to two children of a node in $\mathtt{C}$ are completely disjoint). Then we have*

$$|\mathtt{C}|\,(1, \ldots, 1) \leq \left(\textsc{size}(\mathtt{C})\right)^{\textsc{deg}(\mathtt{C})+1}.$$

Proof of Lemma D.9. For notational simplicity define $N = \text{SIZE}(\mathtt{C})$ and $k = \text{DEG}(\mathtt{C})$. We use induction on $\text{DEPTH}(\mathtt{C})$ to show that $|\mathtt{C}|(1,\ldots,1) \leq N^{k+1}$. For the base case, we have that $\text{DEPTH}(\mathtt{C}) = 0$, and there can only be one node which must contain a coefficient or constant. In this case, $|\mathtt{C}|(1,\ldots,1) = 1$, and $\text{SIZE}(\mathtt{C}) = 1$, and by Definition 4.4 it is the case that $0 \leq k = \text{DEG}(\mathtt{C}) \leq 1$, and it is true that $|\mathtt{C}|(1,\ldots,1) = 1 \leq N^{k+1} = 1^{k+1} = 1$ for $k \in \{0,1\}$.

Assume for $\ell > 0$ an arbitrary circuit $\mathtt{C}$ of $\text{DEPTH}(\mathtt{C}) \leq \ell$ that it is true that $|\mathtt{C}|(1,\ldots,1) \leq N^{k+1}$.

For the inductive step we consider a circuit $\mathtt{C}$ such that $\text{DEPTH}(\mathtt{C}) = \ell + 1$. The sink can only be either a $\times$ or $+$ gate. Let $k_{\mathtt{L}}, k_{\mathtt{R}}$ denote $\text{DEG}(\mathtt{C}_{\mathtt{L}})$ and $\text{DEG}(\mathtt{C}_{\mathtt{R}})$ respectively. Consider when sink node is $\times$. Then note that

$$\begin{aligned} |\mathtt{C}|(1,\ldots,1) &= |\mathtt{C}_{\mathtt{L}}|(1,\ldots,1) \cdot |\mathtt{C}_{\mathtt{R}}|(1,\ldots,1) \\ &\leq (N-1)^{k_{\mathtt{L}}+1} \cdot (N-1)^{k_{\mathtt{R}}+1} \\ &= (N-1)^{k+1} \qquad (30) \\ &\leq N^{k+1}. \end{aligned}$$

In the above the first inequality follows from the inductive hypothesis (and the fact that the size of either subtree is at most $N-1$) and Eq. (30) follows by definition 4.4 which states that for $k = \text{DEG}(\mathtt{C})$ we have $k = k_{\mathtt{L}} + k_{\mathtt{R}} + 1$.

For the case when the sink gate is a $+$ gate, then for $N_{\mathtt{L}} = \text{SIZE}(\mathtt{C}_{\mathtt{L}})$ and $N_{\mathtt{R}} = \text{SIZE}(\mathtt{C}_{\mathtt{R}})$ we have

$$\begin{aligned} |\mathtt{C}|(1,\ldots,1) &= |\mathtt{C}_{\mathtt{L}}|(1,\ldots,1) + |\mathtt{C}_{\mathtt{R}}|(1,\ldots,1) \\ &\leq N_{\mathtt{L}}^{k+1} + N_{\mathtt{R}}^{k+1} \\ &\leq (N-1)^{k+1} \qquad (31) \\ &\leq N^{k+1}. \end{aligned}$$

In the above, the first inequality follows from the inductive hypothes and definition 4.4 (which implies the fact that $k_{\mathtt{L}}, k_{\mathtt{R}} \leq k$). Note that the RHS of this inequality is maximized when the base and exponent of one of the terms is maximized. The second inequality follows from this fact as well as the fact that since $\mathtt{C}$ is a tree we have $N_{\mathtt{L}} + N_{\mathtt{R}} = N - 1$ and, lastly, the fact that $k \geq 0$. This completes the proof.

The upper bound in Lemma 4.8 for the general case is a simple variant of the above proof (but we present a proof sketch of the bound below for completeness):

▶ **Lemma D.10.** *Let $\mathcal{C}$ be a (general) circuit. Then we have*

$$|\mathcal{C}|(1,\ldots,1) \leq 2^{2^{DEG(\mathcal{C})} \cdot DEPTH(\mathcal{C})}.$$

Proof Sketch of Lemma D.10. We use the same notation as in the proof of Lemma D.9 and further define $d = \text{DEPTH}(\mathtt{C})$. We will prove by induction on $\text{DEPTH}(\mathtt{C})$ that $|\mathtt{C}|(1,\ldots,1) \leq 2^{2^k \cdot d}$. The base case argument is similar to that in the proof of Lemma D.9. In the inductive case we have that $d_{\mathtt{L}}, d_{\mathtt{R}} \leq d - 1$.

For the case when the sink node is $\times$, we get that

$$\begin{aligned} |\mathtt{C}|(1,\ldots,1) &= |\mathtt{C}_{\mathtt{L}}|(1,\ldots,1) \times |\mathtt{C}_{\mathtt{R}}|(1,\ldots,1) \\ &\leq 2^{2^{k_{\mathtt{L}}} \cdot d_{\mathtt{L}}} \times 2^{2^{k_{\mathtt{R}}} \cdot d_{\mathtt{R}}} \\ &\leq 2^{2 \cdot 2^{k-1} \cdot (d-1)} \end{aligned}$$

$$\leq 2^{2^k d}.$$

In the above the first inequality follows from inductive hypothesis while the second inequality follows from the fact that $k_{\mathtt{L}}, k_{\mathtt{R}} \leq k-1$ and $d_{\mathtt{L}}, d_{\mathtt{R}} \leq d-1$, where we substitute the upperbound into every respective term.

Now consider the case when the sink node is $+$, we get that

$$\begin{aligned}
|\mathtt{C}|\,(1,\ldots,1) &= |\mathtt{C_L}|\,(1,\ldots,1) + |\mathtt{C_R}|\,(1,\ldots,1) \\
&\leq 2^{2^{k_{\mathtt{L}}} \cdot d_{\mathtt{L}}} + 2^{2^{k_{\mathtt{R}}} \cdot d_{\mathtt{R}}} \\
&\leq 2 \cdot 2^{2^k (d-1)} \\
&\leq 2^{2^k d}.
\end{aligned}$$

In the above the first inequality follows from the inductive hypothesis while the second inequality follows from the facts that $k_{\mathtt{L}}, k_{\mathtt{R}} \leq k$ and $d_{\mathtt{L}}, d_{\mathtt{R}} \leq d-1$. The final inequality follows from the fact that $k \geq 0$. ◀

D.5 OnePass Remarks

Please note that it is *assumed* that the original call to OnePass consists of a call on an input circuit `C` such that the values of members `partial`, `Lweight` and `Rweight` have been initialized to Null across all gates.

The evaluation of $|\mathtt{C}|\,(1,\ldots,1)$ can be defined recursively, as follows (where $\mathtt{C_L}$ and $\mathtt{C_R}$ are the 'left' and 'right' inputs of `C` if they exist):

$$|\mathtt{C}|\,(1,\ldots,1) = \begin{cases} |\mathtt{C_L}|\,(1,\ldots,1) \cdot |\mathtt{C_R}|\,(1,\ldots,1) & \textbf{if } \mathtt{C.type} = \times \\ |\mathtt{C_L}|\,(1,\ldots,1) + |\mathtt{C_R}|\,(1,\ldots,1) & \textbf{if } \mathtt{C.type} = + \\ |\mathtt{C.val}| & \textbf{if } \mathtt{C.type} = \text{NUM} \\ 1 & \textbf{if } \mathtt{C.type} = \text{VAR}. \end{cases} \tag{32}$$

It turns out that for proof of Lemma D.7, we need to argue that when $\mathtt{C.type} = +$, we indeed have

$$\mathtt{C.Lweight} \leftarrow \frac{|\mathtt{C_L}|\,(1,\ldots,1)}{|\mathtt{C_L}|\,(1,\ldots,1) + |\mathtt{C_R}|\,(1,\ldots,1)}; \tag{33}$$

$$\mathtt{C.Rweight} \leftarrow \frac{|\mathtt{C_R}|\,(1,\ldots,1)}{|\mathtt{C_L}|\,(1,\ldots,1) + |\mathtt{C_R}|\,(1,\ldots,1)} \tag{34}$$

D.6 OnePass Example

▶ **Example D.11.** *Let $\mathtt{T}$ encode the expression $(X+Y)(X-Y)+Y^2$. After one pass, Algorithm 2 would have computed the following weight distribution. For the two inputs of the sink gate $\mathtt{C}$, $\mathtt{C.Lweight} = \frac{4}{5}$ and $\mathtt{C.Rweight} = \frac{1}{5}$. Similarly, for $\mathtt{S}$ denoting the left input of $\mathtt{C_L}$, $\mathtt{S.Lweight} = \mathtt{S.Rweight} = \frac{1}{2}$. This is depicted in Fig. 4.*

D.7 Proof of OnePass (Lemma D.6)

Proof. We prove the correct computation of `partial`, `Lweight`, `Rweight` values on `C` by induction over the number of iterations in the topological order TopOrd (line 1) of the input circuit `C`. TopOrd follows the standard definition of a topological ordering over the DAG structure of `C`.

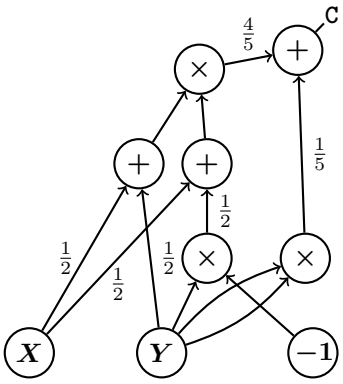

Figure 4 Weights computed by OnePass in Example D.11.

Algorithm 2 OnePass (C)

Input: C: Circuit
Output: C: Annotated Circuit
Output: sum $\in \mathbb{N}$

1: **for** g in TopOrd (C) **do** ▷ TopOrd $(\cdot)$ is the topological order of C
2: **if** g.type = VAR **then**
3: g.partial $\leftarrow 1$
4: **else if** g.type = NUM **then**
5: g.partial $\leftarrow |$g.val$|$
6: **else if** g.type = $\times$ **then**
7: g.partial $\leftarrow$ g_L.partial $\times$ g_R.partial
8: **else**
9: g.partial $\leftarrow$ g_L.partial + g_R.partial
10: g.Lweight $\leftarrow \frac{g_L.\text{partial}}{\text{g.partial}}$
11: g.Rweight $\leftarrow \frac{g_R.\text{partial}}{\text{g.partial}}$
12: **end if**
13: sum $\leftarrow$ g.partial
14: **end for**
15: **return** (sum, C)

For the base case, we have only one gate, which by definition is a source gate and must be either VAR or NUM. In this case, as per eq. (32), lines 3 and 5 correctly compute C.partial as 1.

For the inductive hypothesis, assume that OnePass correctly computes S.partial, S.Lweight, and S.Rweight for all gates g in C with $k \geq 0$ iterations over TopOrd.

Aaron says: Notes above: Algo uses Reduce, but we don't use that anymore. The figure needs to change to a circuit.

We now prove for $k+1$ iterations that OnePass correctly computes the partial, Lweight, and Rweight values for each gate g_i in C for $i \in [k+1]$. The g_{k+1} must be in the last ordering of all gates g_i. When SIZE (C) > 1, if g_{k+1} is a leaf node, we are back to the base case. Otherwise g_{k+1} is an internal node which requires binary input.

When g_{k+1}.type = +, then by line 9 g_{k+1}.partial = g_{k+1_L}.partial + g_{k+1_R}.partial, a correct computation, as per eq. (32). Further, lines 10 and 11 compute g_{k+1}.Lweight =

Algorithm 3 SampleMonomial (C)

Input: C: Circuit
Output: vars: TreeSet
Output: sgn $\in \{-1, 1\}$ ▷ Algorithm 2 should have been run before this one

```
1:  vars ← ∅
2:  if C.type = + then                                   ▷ Sample at every + node
3:      C_samp ← Sample from left input (C_L) and right input (C_R) w.p. C.Lweight and
    C.Rweight.                  ▷ Each call to SampleMonomial uses fresh randomness
4:      (v, s) ← SampleMonomial(C_samp)
5:      return (v, s)
6:  else if C.type = × then                  ▷ Multiply the sampled values of all inputs
7:      sgn ← 1
8:      for input in C.input do
9:          (v, s) ← SampleMonomial(input)
10:         vars ← vars ∪ {v}
11:         sgn ← sgn × s
12:     end for
13:     return (vars, sgn)
14: else if C.type = NUM then                          ▷ The leaf is a coefficient
15:     return ({}, sgn(C.val)) ▷ sgn(·) outputs −1 for C.val ≥ 1 and −1 for C.val ≤ −1
16: else if C.type = var then
17:     return ({C.val}, 1)
18: end if
```

$\frac{g_{k+1_L}.\texttt{partial}}{g_{k+1}.\texttt{partial}}$ and analogously for $g_{k+1}.\texttt{Rweight}$. All values needed for each computation have been correctly computed by the inductive hypothesis.

When $g_{k+1}.\texttt{type} = \times$, then line 7 computes $g_{k+1}.\texttt{partial} = g_{k+1_L.\texttt{partial}} \times g_{k+1_R}.\texttt{partial}$, which indeed by eq. (32) is correct. This concludes the proof of correctness.

Runtime Analysis

It is known that TopOrd(G) is computable in linear time. There are size(C) iterations, each of which takes $O\left(\overline{\mathcal{M}}\left(\log\left(|\texttt{C}(1\ldots,1)|\right), \log\left(\text{size}(\texttt{C})\right)\right)\right)$ time. This can be seen since each of all the numbers which the algorithm computes is at most $|\texttt{C}|\,(1,\ldots,1)$. Hence, by definition each such operation takes $\overline{\mathcal{M}}\left(\log\left(|\texttt{C}(1\ldots,1)|\right), \log \text{size}(\texttt{C})\right)$ time, which proves the claimed runtime. ◀

D.8 SampleMonomial Remarks

We briefly describe the top-down traversal of SampleMonomial. When C.type = +, the input to be visited is sampled from the weighted distribution precomputed by OnePass. When a C.type = × node is visited, both inputs are visited. The algorithm computes two properties: the set of all variable leaf nodes visited, and the product of the signs of visited coefficient leaf nodes. We will assume the TreeSet data structure to maintain sets with logarithmic time insertion and linear time traversal of its elements. While we would like to take advantage of the space efficiency gained in using a circuit C instead an expression tree T, we do not know that such a method exists when computing a sample of the input polynomial representation.

The efficiency gains of circuits over trees is found in the capability of circuits to only require space for each *distinct* term in the compressed representation. This saves space in such polynomials containing non-distinct terms multiplied or added to each other, e.g., x^4. However, to avoid biased sampling, it is imperative to sample from both inputs of a multiplication gate, independently, which is indeed the approach of SampleMonomial.

D.9 Proof of SampleMonomial (Lemma D.7)

Proof. We first need to show that SampleMonomial samples a valid monomial $\mathtt{v_m}$ by sampling and returning a set of variables $\mathtt{v}$, such that $(\mathtt{v}, \mathtt{c})$ is in $\mathtt{E}(\mathtt{C})$ and $\mathtt{v_m}$ is indeed a monomial of the $\widetilde{\Phi}(\mathbf{X})$ encoded in $\mathtt{C}$. We show this via induction over the depth of $\mathtt{C}$.

For the base case, let the depth d of $\mathtt{C}$ be 0. We have that the single gate is either a constant $\mathtt{c}$ for which by line 15 we return $\{\ \}$, or we have that $\mathtt{C.type} = \text{VAR}$ and $\mathtt{C.val} = x$, and by line 17 we return $\{x\}$. By definition 4.1, both cases return a valid $\mathtt{v}$ for some $(\mathtt{v}, \mathtt{c})$ from $\mathtt{E}(\mathtt{C})$, and the base case is proven.

For the inductive hypothesis, assume that for $d \leq k$ for some $k \geq 0$, that it is indeed the case that SampleMonomial returns a valid monomial.

For the inductive step, let us take a circuit $\mathtt{C}$ with $d = k + 1$. Note that each input has depth $d - 1 \leq k$, and by inductive hypothesis both of them sample a valid monomial. Then the sink can be either a $+$ or $\times$ gate. For the case when $\mathtt{C.type} = +$, line 3 of SampleMonomial will choose one of the inputs of the source. By inductive hypothesis it is the case that some valid monomial is being randomly sampled from each of the inputs. Then it follows when $\mathtt{C.type} = +$ that a valid monomial is sampled by SampleMonomial. When the $\mathtt{C.type} = \times$, line 10 computes the set union of the monomials returned by the two inputs of the sink, and it is trivial to see by definition 4.1 that $\mathtt{v_m}$ is a valid monomial encoded by some $(\mathtt{v}, \mathtt{c})$ of $\mathtt{E}(\mathtt{C})$.

We will next prove by induction on the depth d of $\mathtt{C}$ that for $(\mathtt{v}, \mathtt{c}) \in \mathtt{E}(\mathtt{C})$, $\mathtt{v}$ is sampled with a probability $\frac{|\mathtt{c}|}{|\mathtt{C}|(1,\ldots,1)}$.

For the base case $d = 0$, by definition 2.6 we know that the $\text{SIZE}(\mathtt{C}) = 1$ and $\mathtt{C.type} = \text{NUM}$ or VAR. For either case, the probability of the value returned is 1 since there is only one value to sample from. When $\mathtt{C.val} = x$, the algorithm always return the variable set $\{x\}$. When $\mathtt{C.type} = \text{NUM}$, SampleMonomial will always return $\emptyset$.

For the inductive hypothesis, assume that for $d \leq k$ and $k \geq 0$ SampleMonomial indeed returns $\mathtt{v}$ in $(\mathtt{v}, \mathtt{c})$ of $\mathtt{E}(\mathtt{C})$ with probability $\frac{|\mathtt{c}|}{|\mathtt{C}|(1,\ldots,1)}$.

We prove now for $d = k + 1$ the inductive step holds. It is the case that the sink of $\mathtt{C}$ has two inputs $\mathtt{C_L}$ and $\mathtt{C_R}$. Since $\mathtt{C_L}$ and $\mathtt{C_R}$ are both depth $d - 1 \leq k$, by inductive hypothesis, SampleMonomial will return $\mathtt{v_L}$ in $(\mathtt{v_L}, \mathtt{c_L})$ of $\mathtt{E}(\mathtt{C_L})$ and $\mathtt{v_R}$ in $(\mathtt{v_R}, \mathtt{c_R})$ of $\mathtt{E}(\mathtt{C_R})$, from $\mathtt{C_L}$ and $\mathtt{C_R}$ with probability $\frac{|\mathtt{c_L}|}{|\mathtt{C_L}|(1,\ldots,1)}$ and $\frac{|\mathtt{c_R}|}{|\mathtt{C_R}|(1,\ldots,1)}$.

Consider the case when $\mathtt{C.type} = \times$. For the term $(\mathtt{v}, \mathtt{c})$ from $\mathtt{E}(\mathtt{C})$ that is being sampled it is the case that $\mathtt{v} = \mathtt{v_L} \cup \mathtt{v_R}$, where $\mathtt{v_L}$ is coming from $\mathtt{C_L}$ and $\mathtt{v_R}$ from $\mathtt{C_R}$. The probability that SampleMonomial $(\mathtt{C_L})$ returns $\mathtt{v_L}$ is $\frac{|\mathtt{c}_{\mathtt{v_L}}|}{|\mathtt{C_L}|(1,\ldots,1)}$ and $\frac{|\mathtt{c}_{\mathtt{v_R}}|}{|\mathtt{C_R}|(1,\ldots,1)}$ for $\mathtt{v_R}$. Since both $\mathtt{v_L}$ and $\mathtt{v_R}$ are sampled with independent randomness, the final probability for sample $\mathtt{v}$ is then $\frac{|\mathtt{c}_{\mathtt{v_L}}| \cdot |\mathtt{c}_{\mathtt{v_R}}|}{|\mathtt{C_L}|(1,\ldots,1) \cdot |\mathtt{C_R}|(1,\ldots,1)}$. For $(\mathtt{v}, \mathtt{c})$ in $\mathtt{E}(\mathtt{C})$, by definition 4.1 it is indeed the case that $|\mathtt{c}| = |\mathtt{c}_{\mathtt{v_L}}| \cdot |\mathtt{c}_{\mathtt{v_R}}|$ and that (as shown in eq. (32)) $|\mathtt{C}|(1, \ldots, 1) = |\mathtt{C_L}|(1, \ldots, 1) \cdot |\mathtt{C_R}|(1, \ldots, 1)$, and therefore $\mathtt{v}$ is sampled with correct probability $\frac{|\mathtt{c}|}{|\mathtt{C}|(1,\ldots,1)}$.

For the case when $\mathtt{C.type} = +$, SampleMonomial will sample $\mathtt{v}$ from one of its inputs. By inductive hypothesis we know that any $\mathtt{v_L}$ in $\mathtt{E}(\mathtt{C_L})$ and any $\mathtt{v_R}$ in $\mathtt{E}(\mathtt{C_R})$ will both be sampled with correct probability $\frac{|\mathtt{c}_{\mathtt{v_L}}|}{|\mathtt{C_L}|(1,\ldots,1)}$ and $\frac{|\mathtt{c}_{\mathtt{v_R}}|}{|\mathtt{C_R}|(1,\ldots,1)}$, where either $\mathtt{v_L}$ or $\mathtt{v_R}$ will equal $\mathtt{v}$, depending on whether $\mathtt{C_L}$ or $\mathtt{C_R}$ is sampled. Assume that $\mathtt{v}$ is sampled from $\mathtt{C_L}$, and note that

a symmetric argument holds for the case when $\mathtt{v}$ is sampled from $\mathtt{C_R}$. Notice also that the probability of choosing $\mathtt{C_L}$ from $\mathtt{C}$ is $\frac{|\mathtt{C_L}|(1,\ldots,1)}{|\mathtt{C_L}|(1,\ldots,1)+|\mathtt{C_R}|(1,\ldots,1)}$ as computed by OnePass. Then, since SampleMonomial goes top-down, and each sampling choice is independent (which follows from the randomness in the root of $\mathtt{C}$ being independent from the randomness used in its subtrees), the probability for $\mathtt{v}$ to be sampled from $\mathtt{C}$ is equal to the product of the probability that $\mathtt{C_L}$ is sampled from $\mathtt{C}$ and $\mathtt{v}$ is sampled in $\mathtt{C_L}$, and

$$
\begin{aligned}
&Pr(\textsc{SampleMonomial}(\mathtt{C}) = \mathtt{v}) = \\
&Pr(\textsc{SampleMonomial}(\mathtt{C_L}) = \mathtt{v}) \cdot Pr(\mathit{SampledChild}(\mathtt{C}) = \mathtt{C_L}) \\
&= \frac{|\mathtt{c_v}|}{|\mathtt{C_L}|(1,\ldots,1)} \cdot \frac{|\mathtt{C_L}|\,(1,\ldots,1)}{|\mathtt{C_L}|(1,\ldots,1) + |\mathtt{C_R}|(1,\ldots,1)} \\
&= \frac{|\mathtt{c_v}|}{|\mathtt{C}|\,(1,\ldots,1)},
\end{aligned}
$$

and we obtain the desired result.

Lastly, we show by simple induction of the depth d of $\mathtt{C}$ that SampleMonomial indeed returns the correct sign value of $\mathtt{c}$ in $(\mathtt{v}, \mathtt{c})$.

In the base case, $\mathtt{C.type}$ = num or var. For the former, SampleMonomial correctly returns the sign value of the gate. For the latter, SampleMonomial returns the correct sign of 1, since a variable is a neutral element, and 1 is the multiplicative identity, whose product with another sign element will not change that sign element.

For the inductive hypothesis, we assume for a circuit of depth $d \leq k$ and $k \geq 0$ that the algorithm correctly returns the sign value of $\mathtt{c}$.

Similar to before, for a depth $d \leq k+1$, it is true that $\mathtt{C_L}$ and $\mathtt{C_R}$ both return the correct sign of $\mathtt{c}$. For the case that $\mathtt{C.type} = \times$, the sign value of both inputs are multiplied, which is the correct behavior by definition 4.1. When $\mathtt{C.type} = +$, only one input of $\mathtt{C}$ is sampled, and the algorithm returns the correct sign value of $\mathtt{c}$ by inductive hyptothesis.

Run-time Analysis

It is easy to check that except for lines 3 and 10, all lines take $O(1)$ time. Consider an execution of line 10. We note that we will be adding a given set of variables to some set at most once: since the sum of the sizes of the sets at a given level is at most $\textsc{deg}(\mathtt{C})$, each gate visited takes $O(\log \textsc{deg}(\mathtt{C}))$. For Line 3, note that we pick $\mathtt{C_L}$ with probability $\frac{a}{a+b}$ where $a = \mathtt{C.Lweight}$ and $b = \mathtt{C.Rweight}$. We can implement this step by picking a random number $r \in [a+b]$ and then checking if $r \leq a$. It is easy to check that $a + b \leq |\mathtt{C}|\,(1,\ldots,1)$. This means we need to add and compare $\log |\mathtt{C}|\,(1,\ldots,1)$-bit numbers, which can certainly be done in time $\overline{\mathcal{M}}\left(\log\left(|\mathtt{C}(1\ldots,1)|\right), \log \textsc{size}(\mathtt{C})\right)$ (note that this is an over-estimate). Denote $\textsc{Cost}(\mathtt{C})$ (Eq. (35)) to be an upper bound of the number of gates visited by SampleMonomial. Then the runtime is $O\left(\textsc{Cost}(\mathtt{C}) \cdot \log \textsc{deg}(\mathtt{C}) \cdot \overline{\mathcal{M}}\left(\log\left(|\mathtt{C}(1\ldots,1)|\right), \log \textsc{size}(\mathtt{C})\right)\right)$.

We now bound the number of recursive calls in SampleMonomial by $O\left((\textsc{deg}(\mathtt{C}) + 1) \cdot \textsc{depth}(\mathtt{C})\right)$, which by the above will prove the claimed runtime.

Let $\textsc{Cost}(\cdot)$ be a function that models an upper bound on the number of gates that can be visited in the run of SampleMonomial. We define $\textsc{Cost}(\cdot)$ recursively as follows.

$$
\textsc{Cost}(\mathtt{C}) = \begin{cases} 1 + \textsc{Cost}(\mathtt{C_L}) + \textsc{Cost}(\mathtt{C_R}) & \textbf{if } \mathtt{C.type} = \times \\ 1 + \max\left(\textsc{Cost}(\mathtt{C_L}), \textsc{Cost}(\mathtt{C_R})\right) & \textbf{if } \mathtt{C.type} = + \\ 1 & \textbf{otherwise} \end{cases} \tag{35}
$$

First note that the number of gates visited in SampleMonomial is $\leq$ Cost(C). To show that eq. (35) upper bounds the number of nodes visited by SampleMonomial, note that when SampleMonomial visits a gate such that C.type $= \times$, line 8 visits each input of C, as defined in (35). For the case when C.type $= +$, line 3 visits exactly one of the input gates, which may or may not be the subcircuit with the maximum number of gates traversed, which makes Cost $(\cdot)$ an upperbound. Finally, it is trivial to see that when C.type $\in \{\text{VAR}, \text{NUM}\}$, i.e., a source gate, that only one gate is visited.

We prove the following inequality holds.

$$2\left(\text{DEG}(\texttt{C})+1\right)\cdot\text{DEPTH}(\texttt{C})+1 \geq \text{Cost}(\texttt{C}) \tag{36}$$

Note that eq. (36) implies the claimed runtime. We prove eq. (36) for the number of gates traversed in SampleMonomial using induction over DEPTH(C). Recall how degree is defined in definition 4.4.

For the base case $\text{DEG}(\texttt{C}) = \{0,1\}, \text{DEPTH}(\texttt{C}) = 0$, $\text{Cost}(\texttt{C}) = 1$, and it is trivial to see that the inequality $2\text{DEG}(\texttt{C})\cdot\text{DEPTH}(\texttt{C})+1 \geq \text{Cost}(\texttt{C})$ holds.

For the inductive hypothesis, we assume the bound holds for any circuit where $\ell \geq \text{DEPTH}(\texttt{C}) \geq 0$. Now consider the case when SampleMonomial has an arbitrary circuit C input with $\text{DEPTH}(\texttt{C}) = \ell+1$. By definition C.type $\in \{+,\times\}$. Note that since $\text{DEPTH}(\texttt{C}) \geq 1$, C must have input(s). Further we know that by the inductive hypothesis the inputs $\texttt{C}_i$ for $i \in \{\texttt{L},\texttt{R}\}$ of the sink gate C uphold the bound

$$2\left(\text{DEG}(\texttt{C}_i)+1\right)\cdot\text{DEPTH}(\texttt{C}_i)+1 \geq \text{Cost}(\texttt{C}_i). \tag{37}$$

In particular, since for any i, eq. (37) holds, then it immediately follows that an inequality whose operands consist of a sum of the aforementioned inequalities must also hold. This is readily seen in the inequality of eq. (39) and eq. (40), where $2\left(\text{DEG}(\texttt{C}_\texttt{L})+1\right)\cdot\text{DEPTH}(\texttt{C}_\texttt{L}) \geq \text{Cost}(\texttt{C}_\texttt{L})$, likewise for $\texttt{C}_\texttt{R}$, and $1 \geq 1$. It is also true that $\text{DEPTH}(\texttt{C}_\texttt{L}) \leq \text{DEPTH}(\texttt{C}) - 1$ and $\text{DEPTH}(\texttt{C}_\texttt{R}) \leq \text{DEPTH}(\texttt{C}) - 1$.

If C.type $= +$, then $\text{DEG}(\texttt{C}) = \max\left(\text{DEG}(\texttt{C}_\texttt{L}), \text{DEG}(\texttt{C}_\texttt{R})\right)$. Otherwise C.type $= \times$ and $\text{DEG}(\texttt{C}) = \text{DEG}(\texttt{C}_\texttt{L})+\text{DEG}(\texttt{C}_\texttt{R})+1$. In either case it is true that $\text{DEPTH}(\texttt{C}) = \max\left(\text{DEPTH}(\texttt{C}_\texttt{L}), \text{DEPTH}(\texttt{C}_\texttt{R})\right)+1$.

If C.type $= \times$, then, by eq. (35), substituting values, the following should hold,

$$\begin{aligned} &2\left(\text{DEG}(\texttt{C}_\texttt{L})+\text{DEG}(\texttt{C}_\texttt{R})+2\right)\cdot\left(\max(\text{DEPTH}(\texttt{C}_\texttt{L}), \text{DEPTH}(\texttt{C}_\texttt{R}))+1\right)+1 && (38)\\ &\quad\geq 2\left(\text{DEG}(\texttt{C}_\texttt{L})+1\right)\cdot\text{DEPTH}(\texttt{C}_\texttt{L})+2\left(\text{DEG}(\texttt{C}_\texttt{R})+1\right)\cdot\text{DEPTH}(\texttt{C}_\texttt{R})+3 && (39)\\ &\quad\geq 1+\text{Cost}(\texttt{C}_\texttt{L})+\text{Cost}(\texttt{C}_\texttt{R}) = \text{Cost}(\texttt{C}). && (40)\end{aligned}$$

To prove (39), first, eq. (38) expands to,

$$2\text{DEG}(\texttt{C}_\texttt{L})\cdot\text{DEPTH}_{\max}+2\text{DEG}(\texttt{C}_\texttt{R})\cdot\text{DEPTH}_{\max}+4\text{DEPTH}_{\max}+2\text{DEG}(\texttt{C}_\texttt{L})+2\text{DEG}(\texttt{C}_\texttt{R})+4+1 \tag{41}$$

where $\text{DEPTH}_{\max}$ is used to denote the maximum depth of the two input subcircuits. Eq. (39) expands to

$$2\text{DEG}(\texttt{C}_\texttt{L})\cdot\text{DEPTH}(\texttt{C}_\texttt{L})+2\text{DEPTH}(\texttt{C}_\texttt{L})+2\text{DEG}(\texttt{C}_\texttt{R})\cdot\text{DEPTH}(\texttt{C}_\texttt{R})+2\text{DEPTH}(\texttt{C}_\texttt{R})+3 \tag{42}$$

Putting Eq. (41) and Eq. (42) together we get

$$\begin{aligned} &2\text{DEG}(\texttt{C}_\texttt{L})\cdot\text{DEPTH}_{\max}+2\text{DEG}(\texttt{C}_\texttt{R})\cdot\text{DEPTH}_{\max}+4\text{DEPTH}_{\max}+2\text{DEG}(\texttt{C}_\texttt{L})+2\text{DEG}(\texttt{C}_\texttt{R})+5\\ &\quad\geq 2\text{DEG}(\texttt{C}_\texttt{L})\cdot\text{DEPTH}(\texttt{C}_\texttt{L})+2\text{DEG}(\texttt{C}_\texttt{R})\cdot\text{DEPTH}(\texttt{C}_\texttt{R})+2\text{DEPTH}(\texttt{C}_\texttt{L})+2\text{DEPTH}(\texttt{C}_\texttt{R})+3 \end{aligned} \tag{43}$$

Since the following is always true,

$$2\textsc{deg}(\mathtt{C_L}) \cdot \textsc{depth}_{\max} + 2\textsc{deg}(\mathtt{C_R}) \cdot \textsc{depth}_{\max} + 4\textsc{depth}_{\max} + 5$$
$$\geq 2\textsc{deg}(\mathtt{C_L}) \cdot \textsc{depth}(\mathtt{C_L}) + 2\textsc{deg}(\mathtt{C_R}) \cdot \textsc{depth}(\mathtt{C_R}) + 2\textsc{depth}(\mathtt{C_L}) + 2\textsc{depth}(\mathtt{C_R}) + 3,$$

then it is the case that Eq. (43) is *always* true.

Now to justify (40) which holds for the following reasons. First, eq. (40) is the result of Eq. (35) when $\mathtt{C.type} = \times$. Eq. (39) is then produced by substituting the upperbound of (37) for each $\textsc{Cost}(\mathtt{C}_i)$, trivially establishing the upper bound of (40). This proves eq. (36) for the $\times$ case.

For the case when $\mathtt{C.type} = +$, substituting values yields

$$2\left(\max(\textsc{deg}(\mathtt{C_L}), \textsc{deg}(\mathtt{C_R})) + 1\right) \cdot \left(\max(\textsc{depth}(\mathtt{C_L}), \textsc{depth}(\mathtt{C_R})) + 1\right) + 1 \tag{44}$$
$$\geq \max\left(2\left(\textsc{deg}(\mathtt{C_L}) + 1\right) \cdot \textsc{depth}(\mathtt{C_L}) + 1, 2\left(\textsc{deg}(\mathtt{C_R}) + 1\right) \cdot \textsc{depth}(\mathtt{C_R}) + 1\right) + 1 \tag{45}$$
$$\geq 1 + \max(\textsc{Cost}(\mathtt{C_L}), \textsc{Cost}(\mathtt{C_R})) = \textsc{Cost}(\mathtt{C}) \tag{46}$$

To prove (45), eq. (44) expands to

$$2\textsc{deg}_{\max}\textsc{depth}_{\max} + 2\textsc{deg}_{\max} + 2\textsc{depth}_{\max} + 2 + 1. \tag{47}$$

Since $\textsc{deg}_{\max} \cdot \textsc{depth}_{\max} \geq \textsc{deg}(\mathtt{C}_i) \cdot \textsc{depth}(\mathtt{C}_i)$, the following upper bound holds for the expansion of eq. (45):

$$2\textsc{deg}_{\max}\textsc{depth}_{\max} + 2\textsc{depth}_{\max} + 2 \tag{48}$$

Putting it together we obtain the following for (45):

$$2\textsc{deg}_{\max}\textsc{depth}_{\max} + 2\textsc{deg}_{\max} + 2\textsc{depth}_{\max} + 3$$
$$\geq 2\textsc{deg}_{\max}\textsc{depth}_{\max} + 2\textsc{depth}_{\max} + 2, \tag{49}$$

where it can be readily seen that the inequality stands and (49) follows. This proves (45).

Similar to the case of $\mathtt{C.type} = \times$, (46) follows by equations (35) and (37).

This proves (36) as desired. ◀

D.10 Experimental Results

Recall that by definition of BIDB, a query result cannot be derived by a self-join between non-identical tuples belonging to the same block. Note, that by Theorem 4.7, γ must be a constant in order for Algorithm 1 to acheive linear time. We would like to determine experimentally whether queries over BIDB instances in practice generate a constant number of cancellations or not. Such an experiment would ideally use a database instance with queries both considered to be typical representations of what is seen in practice.

We ran our experiments using Windows 10 WSL Operating System with an Intel Core i7 2.40GHz processor and 16GB RAM. All experiments used the PostgreSQL 13.0 database system.

For the data we used the MayBMS data generator [1] tool to randomly generate uncertain versions of TPCH tables. The queries computed over the database instance are Q_1, Q_2, and Q_3 from [5], all of which are modified versions of TPC-H queries Q_3, Q_6, and Q_7 where all aggregations have been dropped.

As written, the queries disallow BIDB cross terms. We first ran all queries, noting the result size for each. Next the queries were rewritten so as not to filter out the cross terms. The comparison of the sizes of both result sets should then suggest in one way or another whether or not there exist many cross terms in practice. As seen, the experimental query results contain little to no cancelling terms. Fig. 5 shows the result sizes of the queries, where column CF is the result size when all cross terms are filtered out, column CI shows the number of output tuples when the cancelled tuples are included in the result, and the last column is the value of γ. The experiments show γ to be in a range between $[0, 0.1]\%$, indicating that only a negligible or constant (compare the result sizes of $Q_1 < Q_2$ and their respective γ values) amount of tuples are cancelled in practice when running queries over a typical BIDB instance. Interestingly, only one of the three queries had tuples that violated the BIDB constraint.

To conclude, the results in Fig. 5 show experimentally that γ is negligible in practice for BIDB queries. We also observe that (i) tuple presence is independent across blocks, so the corresponding probabilities (and hence p_0) are independent of the number of blocks, and (ii) BIDBs model uncertain attributes, so block size (and hence γ) is a function of the "messiness" of a dataset, rather than its size. Thus, we expect Theorem 4.7 to hold in general.

Query	CF	CI	γ
Q_1	$46,714$	$46,768$	0.1%
Q_2	179.917	$179,917$	0%
Q_3	$11,535$	$11,535$	0%

Figure 5 Number of Cancellations for Queries Over BIDB.

E Circuits

E.1 Representing Polynomials with Circuits

E.1.1 Circuits for query plans

Atri says: Since this comment is not showing up below, I do not follow why the last sentence of this para is true.

We now formalize circuits and the construction of circuits for $\mathcal{RA}^+$ queries. As mentioned earlier, we represent lineage polynomials as arithmetic circuits over $\mathbb{N}$-valued variables with $+$, $\times$. A circuit for query Q and $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$ is a directed acyclic graph $\langle V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, E_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \phi_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \ell_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \rangle$ with vertices $V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}$ and directed edges $E_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \subset V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}^2$. The sink function $\phi_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} : \mathcal{U}^n \to V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}$ is a partial function that maps the tuples of the n-ary

Atri says: In the main paper we have used n to denote the number of input tuples so we need to use some other notation n but since I do not know where all this change will need to be propagated so am not changing it for now.

relation $Q(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})$ to vertices. We require that $\phi_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}$'s range be limited to sink vertices (i.e., vertices with out-degree 0). A function $\ell_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} : V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \to \{ +, \times \} \cup \mathbb{N} \cup \mathbf{X}$ assigns a label to each node: Source nodes (i.e., vertices with in-degree 0) are labeled with constants or variables (i.e., $\mathbb{N} \cup \mathbf{X}$), while the remaining nodes are labeled with the symbol $+$ or $\times$. We require that vertices have an in-degree of at most two. Note that we can construct circuits for

BIDBs in time linear in the time required for deterministic query processing over a possible world of the BIDB under the aforementioned assumption that $\left|\mathcal{D}_{\mathbb{N}[\mathbf{X}]}\right| \leq c \cdot |D|$.

Atri says: I do not follow the last sentence.

E.2 Modeling Circuit Construction

We now connect the size of a circuit (where the size of a circuit is the number of vertices in the corresponding DAG) for a given $\mathcal{RA}^+$ query Q and $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$ to the runtime $T^*_{det}(Q, D_\Omega)$ of the PDB's deterministic bounding database D_Ω. We do this formally by showing that the size of the circuit is asymptotically no worse than the corresponding runtime of a large class of deterministic query processing algorithms.

Each vertex $v \in V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}$ in the arithmetic circuit for

$$\left\langle V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, E_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \phi_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \ell_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \right\rangle$$

encodes a polynomial, realized as

$$\mathbf{lin}\,(v) = \begin{cases} \sum_{v':(v',v)\in E_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}} \mathbf{lin}\,(v') & \textbf{if } \ell(v) = + \\ \prod_{v':(v',v)\in E_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}} \mathbf{lin}\,(v') & \textbf{if } \ell(v) = \times \\ \ell(v) & \textbf{otherwise} \end{cases}$$

We define the circuit for a $\mathcal{RA}^+$ query Q recursively by cases as follows. In each case, let $\left\langle V_{Q_i,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, E_{Q_i,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \phi_{Q_i,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}, \ell_{Q_i,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \right\rangle$ denote the circuit for subquery Q_i. We implicitly include in all circuits a global zero node v_0 s.t., $\ell_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}(v_0) = 0$ for any $Q, \mathcal{D}_{\mathbb{N}[\mathbf{X}]}$.

Algorithm 4 defines how the circuit for a query result is constructed. We quickly review the number of vertices emitted in each case.

Base Relation. This circuit has $|D_\Omega.R|$ vertices.

Selection. If we assume dead sinks are iteratively garbage collected, this circuit has at most $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}|$ vertices.

Projection. This formulation will produce vertices with an in-degree greater than two, a problem that we correct by replacing every vertex with an in-degree over two by an equivalent fan-in two tree. The resulting structure has at most $|Q_1| - 1$ new vertices. The corrected circuit thus has at most $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1|$ vertices.

Union. This circuit has $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |V_{Q_2,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1 \cap Q_2|$ vertices.

k-ary Join. As in projection, newly created vertices will have an in-degree of k, and a fan-in two tree is required. There are $|Q_1 \bowtie \ldots \bowtie Q_k|$ such vertices, so the corrected circuit has $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + \ldots + |V_{Q_k,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + (k-1)|Q_1 \bowtie \ldots \bowtie Q_k|$ vertices.

E.2.1 Bounding circuit depth

We first show that the depth of the circuit (DEPTH; Definition 4.3) is bounded by the size of the query. Denote by $|Q|$ the number of relational operators in query Q, which recall we assume is a constant.

▶ **Proposition E.1** (Circuit depth is bounded). *Let Q be a relational query and D_Ω be a deterministic bounding database with n tuples. There exists a (lineage) circuit $\mathcal{C}^*$ encoding the lineage of all tuples $t \in Q(D_\Omega)$ for which $\text{DEPTH}(\mathcal{C}^*) \leq O(k|Q|\log(n))$.*

Algorithm 4 LC $(Q, D_\Omega, E, V, \ell)$

Input: Q: query
Input: D_Ω: a deterministic bounding database
Input: E, V, ℓ: accumulators for the edge list, vertex list, and vertex label list.
Output: $\mathtt{C} = \langle E, V, \phi, \ell \rangle$: a circuit encoding the lineage of each tuple in $Q(D_\Omega)$

if Q is R **then** ▷ **Case 1**: Q is a relation atom
  for $t \in D_\Omega.R$ **do**
    $V \leftarrow V \cup \{v_t\}$; $\ell \leftarrow \ell \cup \{(v_t, R(t))\}$ ▷ Allocate a fresh node v_t
    $\phi(t) \leftarrow v_t$
  end for
else if Q is $\sigma_\theta(Q')$ **then** ▷ **Case 2**: Q is a Selection
  $\langle V, E, \phi', \ell \rangle \leftarrow LC(Q', D_\Omega, V, E, \ell)$
  for $t \in \textsc{Dom}(\phi')$ **do**
    if $\theta(t)$ **then** $\phi(t) \leftarrow \phi'(t)$ **else** $\phi(t) \leftarrow v_0$
  end for
else if Q is $\pi_{\vec{A}}(Q')$ **then** ▷ **Case 3**: Q is a Projection
  $\langle V, E, \phi', \ell \rangle \leftarrow LC(Q', D_\Omega, V, E, \ell)$
  for $t \in \pi_{\vec{A}}(Q'(D_\Omega))$ **do**
    $V \leftarrow V \cup \{v_t\}$; $\ell \leftarrow \ell \cup \{(v_t, +)\}$ ▷ Allocate a fresh node v_t
    $\phi(t) \leftarrow v_t$
  end for
  for $t \in Q'(D_\Omega)$ **do**
    $E \leftarrow E \cup \{(\phi'(t), \phi(\pi_{\vec{A}} t))\}$
  end for
  Correct nodes with in-degrees > 2 by appending an equivalent fan-in two tree instead
else if Q is $Q_1 \cup Q_2$ **then** ▷ **Case 4**: Q is a Bag Union
  $\langle V, E, \phi_1, \ell \rangle \leftarrow LC(Q_1, D_\Omega, V, E, \ell)$
  $\langle V, E, \phi_2, \ell \rangle \leftarrow LC(Q_2, D_\Omega, V, E, \ell)$
  $\phi \leftarrow \phi_1 \cup \phi_2$
  for $t \in \textsc{Dom}(\phi_1) \cap \textsc{Dom}(\phi_2)$ **do**
    $V \leftarrow V \cup \{v_t\}$; $\ell \leftarrow \ell \cup \{(v_t, +)\}$ ▷ Allocate a fresh node v_t
    $\phi(t) \leftarrow v_t$
    $E \leftarrow E \cup \{(\phi_1(t), v_t), (\phi_2(t), v_t)\}$
  end for
else if Q is $Q_1 \bowtie \ldots \bowtie Q_m$ **then** ▷ **Case 5**: Q is a m-ary Join
  for $i \in [m]$ **do**
    $\langle V, E, \phi_i, \ell \rangle \leftarrow LC(Q_i, D_\Omega, V, E, \ell)$
  end for
  for $t \in \textsc{Dom}(\phi_1) \bowtie \ldots \bowtie \textsc{Dom}(\phi_m)$ **do**
    $V \leftarrow V \cup \{v_t\}$; $\ell \leftarrow \ell \cup \{(v_t, \times)\}$ ▷ Allocate a fresh node v_t
    $\phi(t) \leftarrow v_t$
    $E \leftarrow E \cup \left\{ (\phi_i(\pi_{sch(Q_i(D_\Omega))}(t)), v_t) \mid i \in [n] \right\}$
  end for
  Correct nodes with in-degrees > 2 by appending an equivalent fan-in two tree instead
end if

Proof. We show that the bound of Proposition E.1 holds for the circuit constructed by Algorithm 4. First, observe that Algorithm 4 is (recursively) invoked exactly once for every relational operator or base relation in Q; It thus suffices to show that a call to Algorithm 4 adds at most $O_k(\log(n))$ to the depth of a circuit produced by any recursive invocation. Second, observe that modulo the logarithmic fan-in of the projection and join cases, the depth of the output is at most one greater than the depth of any input (or at most 1 in the base case of relation atoms). For the join case, the number of in-edges can be no greater than the join width, which itself is bounded by k. The depth thus increases by at most a constant factor of $\lceil \log(k) \rceil = O_k(1)$. For the projection case, observe that the fan-in is bounded by $|Q'(D_\Omega)|$, which is in turn bounded by n^k. The depth increase for any projection node is thus at most $\lceil \log(n^k) \rceil = O(k \log(n))$, as desired. ◀

E.2.2 Circuit size vs. runtime

▶ **Lemma E.2.** *Given a $\mathbb{N}[\mathbf{X}]$-encoded PDB $\mathcal{D}_{\mathbb{N}[\mathbf{X}]}$ with deterministic bounding database D_Ω, and an $\mathcal{RA}^+$ query Q, the runtime of Q over D_Ω has the same or greater complexity as the size of the lineage of $Q(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})$. That is, we have $\left|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}\right| \leq kT^*_{det}(Q, D_\Omega) + 1$, where $k \geq 1$ is the maximal degree of any polynomial in $Q(\mathcal{D}_{\mathbb{N}[\mathbf{X}]})$.*

Proof. We prove by induction that $\left|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}} \setminus \{v_0\}\right| \leq kT^*_{det}(Q, D_\Omega)$. For clarity, we implicitly exclude v_0 in the proof below.

The base case is a base relation: $Q = R$ and is trivially true since $|V_{R,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| = |D_\Omega.R| = T^*_{det}(R, D_\Omega)$ (note that here the degree $k = 1$). For the inductive step, we assume that we have circuits for subqueries $Q_1, \ldots, Q_m$ such that $|V_{Q_i,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| \leq k_i T^*_{det}(Q_i, D_\Omega)$ where k_i is the degree of Q_i.

Selection. Assume that $Q = \sigma_\theta(Q_1)$. In the circuit for Q, $|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| = |V_{Q_1,D_\Omega}|$ vertices, so from the inductive assumption and $T^*_{det}(Q, D_\Omega) = T^*_{det}(Q_1, D_\Omega)$ by definition, we have $|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| \leq kT^*_{det}(Q, D_\Omega)$.

Projection. Assume that $Q = \pi_{\mathbf{A}}(Q_1)$. The circuit for Q has at most $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1|$ vertices.

$$|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| \leq |V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1|$$

(From the inductive assumption)

$$\leq kT^*_{det}(Q_1, D_\Omega) + |Q_1|$$

(By definition of $T^*_{det}(Q, D_\Omega)$)

$$\leq kT^*_{det}(Q, D_\Omega).$$

Union. Assume that $Q = Q_1 \cup Q_2$. The circuit for Q has $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |V_{Q_2,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1 \cap Q_2|$ vertices.

$$|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| \leq |V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |V_{Q_2,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + |Q_1| + |Q_2|$$

(From the inductive assumption)

$$\leq k(T^*_{det}(Q_1, D_\Omega) + T^*_{det}(Q_2, D_\Omega)) + (|Q_1| + |Q_2|)$$

(By definition of $T^*_{det}(Q, D_\Omega)$)

$$\leq k(T^*_{det}(Q, D_\Omega)).$$

m-ary Join. Assume that $Q = Q_1 \bowtie \ldots \bowtie Q_m$. Note that $k = \sum_{i=1}^{m} k_i \geq m$. The circuit for Q has $|V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + \ldots + |V_{Q_k,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + (m-1)|Q_1 \bowtie \ldots \bowtie Q_k|$ vertices.

$$|V_{Q,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| = |V_{Q_1,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + \ldots + |V_{Q_k,\mathcal{D}_{\mathbb{N}[\mathbf{X}]}}| + (m-1)|Q_1 \bowtie \ldots \bowtie Q_k|$$

From the inductive assumption and noting $\forall i : k_i \leq k$ and $m \leq k$

$$\begin{aligned}
&\leq kT^*_{det}(Q_1, D_\Omega) + \ldots + kT^*_{det}(Q_k, D_\Omega) + \\
&\quad (m-1)|Q_1 \bowtie \ldots \bowtie Q_m| \\
&\leq k(T^*_{det}(Q_1, D_\Omega) + \ldots + T^*_{det}(Q_m, D_\Omega) + \\
&\quad |Q_1 \bowtie \ldots \bowtie Q_m|)
\end{aligned}$$

(By definition of $T^*_{det}(Q, D_\Omega)$ and assumption on $T_{join}(\cdot)$)

$$\leq kT^*_{det}(Q, D_\Omega).$$

The property holds for all recursive queries, and the proof holds. ◀

E.2.3 Runtime of LC

We next need to show that we can construct the circuit in time linear in the deterministic runtime.

▶ **Lemma E.3.** *Given a query Q over a deterministic bounding database D_Ω and the $\mathtt{C}^*$ output by Algorithm 4, the runtime $T_{LC}(Q, D_\Omega, \mathtt{C}^*) \leq O(T^*_{det}(Q, D_\Omega))$.*

Proof. By analysis of Algorithm 4, invoked as $\mathtt{C}^* \leftarrow LC(Q, D_\Omega, \{v_0\}, \emptyset, \{(v_0, 0)\})$.

We assume that the vertex list V, edge list E, and vertex label list ℓ are mutable accumulators with $O(1)$ ammortized append. We assume that the tuple to sink mapping ϕ is a linked hashmap, with $O(1)$ insertions and retrievals, and $O(n)$ iteration over the domain of keys. We assume that the n-ary join $\textsc{Dom}(\phi_1) \bowtie \ldots \bowtie \textsc{Dom}(\phi_n)$ can be computed in time $T_{join}(\textsc{Dom}(\phi_1), \ldots, \textsc{Dom}(\phi_n))$ (Definition 2.10) and that an intersection $\textsc{Dom}(\phi_1) \cap \textsc{Dom}(\phi_2)$ can be computed in time $O(|\textsc{Dom}(\phi_1)| + |\textsc{Dom}(\phi_2)|)$ (e.g., with a hash table).

Before proving our runtime bound, we first observe that $T^*_{det}(Q, D) \geq \Omega(|Q(D)|)$. This is true by construction for the relation, projection, and union cases, by Definition 2.10 for joins, and by the observation that $|\sigma(R)| \leq |R|$.

We showthat $T^*_{det}(Q, D_\Omega)$ is an upper-bound for the runtime of Algorithm 4 by recursion. The base case of a relation atom requires only an $O(|D_\Omega.R|)$ iteration over the source tuples. For the remaining cases, we make the recursive assumption that for every subquery Q', it holds that $O(T^*_{det}(Q', D_\Omega))$ bounds the runtime of Algorithm 4.

Selection. Selection requires a recursive call to Algorithm 4, which by the recursive assumption is bounded by $O(T^*_{det}(Q', D_\Omega))$. Algorithm 4 requires a loop over every element of $Q'(D_\Omega)$. By the observation above that $T^*_{det}(Q, D) \geq \Omega(|Q(D)|)$, this iteration is also bounded by $O(T^*_{det}(Q', D_\Omega))$.

Projection. Projection requires a recursive call to Algorithm 4, which by the recursive assumption is bounded by $O(T^*_{det}(Q', D_\Omega))$, which in turn is a term in $T^*_{det}(\pi_{\vec{A}} Q', D_\Omega)$. What remains is an iteration over $\pi_{\vec{A}}(Q(D_\Omega))$ (lines 13–16), an iteration over $Q'(D_\Omega)$ (lines 17–19), and the construction of a fan-in tree (line 20). The first iteration is $O(|Q(D_\Omega)|) \leq$

$O(T^*_{det}(Q, D_\Omega))$. The second iteration and the construction of the bounded fan-in tree are both $O(|Q'(D_\Omega)|) \leq O(T^*_{det}(Q', D_\Omega)) \leq O(T^*_{det}(Q, D_\Omega))$, by the the observation above that $T^*_{det}(Q, D) \geq \Omega(|Q(D)|)$.

Bag Union. As above, the recursive calls explicitly correspond to terms in the expansion of $T^*_{det}(Q_1 \cup Q_2, D_\Omega)$. Initializing ϕ (line 24) can be accomplished in $O(\textsc{Dom}(\phi_1) + \textsc{Dom}(\phi_2)) = O(|Q_1(D_\Omega)| + |Q_2(D_\Omega)|) \leq O(T^*_{det}(Q_1, D_\Omega) + T^*_{det}(Q_2, D_\Omega))$. The remainder requires computing $Q_1 \cup Q_2$ (line 25) and iterating over it (lines 25–29), which is $O(|Q_1| + |Q_2|)$ as noted above — this directly corresponds to terms in $T^*_{det}(Q_1 \cup Q_2, D_\Omega)$.

m-ary Join. As in the prior cases, recursive calls explicitly correspond to terms in our target runtime. The remaining logic involves (i) computing $\textsc{Dom}(\phi_1) \bowtie \ldots \bowtie \textsc{Dom}(\phi_m)$, (ii) iterating over the results, and (iii) creating a fan-in tree. Respectively, these are:
(i) $T_{join}(\textsc{Dom}(\phi_1), \ldots, \textsc{Dom}(\phi_m))$
(ii) $O(|Q_1(D_\Omega) \bowtie \ldots \bowtie Q_m(D_\Omega)|) \leq O(T_{join}(\textsc{Dom}(\phi_1), \ldots, \textsc{Dom}(\phi_m)))$ (Definition 2.10)
(iii) $O(m|Q_1(D_\Omega) \bowtie \ldots \bowtie Q_m(D_\Omega)|)$ (as (ii), noting that $m \leq k = O(1)$) ◀

F Higher Moments

We make a simple observation to conclude the presentation of our results. So far we have only focused on the expectation of Φ. In addition, we could e.g. prove bounds of the probability of a tuple's multiplicity being at least 1. Progress can be made on this as follows: For any positive integer m we can compute the m-th moment of the multiplicities, allowing us to e.g. use the Chebyschev inequality or other high moment based probability bounds on the events we might be interested in. We leave further investigations for future work.

G The Karp-Luby Estimator

Computing the marginal probability of a tuple in the output of a set-probabilistic database query has been studied extensively. To the best of our knowledge, the current state of the art approximation algorithm for this problem is the Karp-Luby estimator [30], which first appeared in MayBMS/Sprout [38], and more recently as part of an online "anytime" approximation algorithm [20, 15].

The estimator works by observing that for any ℓ random binary (but not necessarily independent) events $\mathbf{W}_1, \ldots, \mathbf{W}_\ell$, the probability of at least one event occurring (i.e., $Pr(\mathbf{W}_1 \vee \ldots \vee \mathbf{W}_\ell)$) is bounded from above by the sum of the independent event probabilities (i.e., $Pr(\mathbf{W}_1 \vee \ldots \vee \mathbf{W}_\ell) \leq Pr(\mathbf{W}_1) + \ldots + Pr(\mathbf{W}_\ell)$). Starting from this ('easily' computable and large) value, the estimator proceeds to correct the estimate by estimating how much of an over-estimate it is. Specifically, if $\mathcal{P}$ is the joint distribution over $\mathbf{W}$, the estimator computes an approximation of:

$$\mathcal{O} = \mathop{\mathbb{E}}_{\mathbf{W} \sim \mathcal{P}} \Big[|\{ i \mid \mathbf{W}_i = 1, i \in [\ell] \}| \Big].$$

The accuracy of this estimate is improved by conditioning $\mathcal{P}$ on a W_i chosen uniformly at random (which ensures that the sampled count will be at least 1) and correcting the resulting estimate by $Pr(W_i)$. With an estimate of $\mathcal{O}$, it can easily be verified that the probability of the disjunction can be computed as:

$$Pr(\mathbf{W}_1 \vee \ldots \vee \mathbf{W}_\ell) = Pr(\mathbf{W}_1) + \ldots + Pr(\mathbf{W}_\ell) - \mathcal{O}$$

The Karp-Luby estimator is employed on the SMB representation[14] of $\mathtt{C}$ (to solve the set-PDB version of Problem 1.6), where each W_i represents the event that one monomial is true. By simple inspection, if there are ℓ monomials, this estimator has runtime $\Omega(\ell)$. Further, a minimum of $\left\lceil \frac{3 \cdot \ell \cdot \log(\frac{2}{\delta})}{\epsilon^2} \right\rceil$ invocations of the estimator are required to achieve $1 \pm \epsilon$ approximation with probability at least $1 - \delta$ [38], entailing a runtime at least quadratic in ℓ. As an arbitrary lineage circuit $\mathtt{C}$ may encode $\Omega\left(|\mathtt{C}|^k\right)$ monomials, the worst case runtime is at least $\Omega\left(|\mathtt{C}|^{2k}\right)$ (where k is the 'degree' of lineage polynomial encoded by $\mathtt{C}$). By contrast note that by the discussion after Lemma 4.8 we can solve Problem 1.6 in time $O\left(|\mathtt{C}|^2\right)$ for all BIDB circuits *independent* of the degree k.

H Parameterized Complexity

In Sec. 3, we utilized common conjectures from fine-grained complexity theory. The notion of $\#\mathsf{W}[1] - hard$ is a standard notion in *parameterized complexity*, which by now is a standard complexity tool in providing data complexity bounds on query processing results [22]. E.g. the fact that k-matching is $\#\mathsf{W}[1] - hard$ implies that we cannot have an $n^{\Omega(1)}$ runtime. However, these results do not carefully track the exponent in the hardness result. E.g. $\#\mathsf{W}[1] - hard$ for the general k-matching problem does not imply anything specific for the 3-matching problem. Similar questions have led to intense research into the new sub-field of *fine-grained complexity* (see [48]), where we care about the exponent in our hardness assumptions as well– e.g. Conjecture 3.3 is based on the popular *Triangle detection hypothesis* in this area (cf. [34]).

I Response to first cycle reviewer comments

This paper is a resubmission of our submission to the ICDT first cycle. We thank the reviewers for their insightful comments, which we believe has helped improve the presentation of the paper tremendously. We use this section to document the changes that have been made since our prior submission, and in particular, how we have addressed reviewer comments (reviewer comments are shaded followed by our responses).

I.1 Meta Review

> Problem definition not stated rigorously nor motivated. Discussion needed on the standard PDB approach vs your approach.

We rewrote Sec. 1 to specifically address this concern. The opening paragraph precisely and formally states the query evaluation problem in bag-PDBs. We use a series of problem statements to clearly define the problem we are addressing as it relates to the query evaluation problem. We made the concrete problem statements more precise by more clearly formalizing $T^*_{det}(Q, D_\Omega)$ and stating our runtime objectives relative to it (**??**, 1.5, 1.6).

We have included a discussion of the standard approach, e.g. see the paragraph **Relationship to Set-Probabilistic Query Evaluation** on page 4.

> Definition 2.6 on reduced BIDB polynomials seem not the right tool for the studied problem.

[14] Note that since we are in the set semantics, in the lineage polynomial/formula, addition is logical OR and multiplication is logical AND.

We have chosen to stick with a less formal, ad-hoc definition (please see Definition 1.3 and **??**) of the general problem as suggested by both Reviewer 1 and Reviewer 2. Our earlier proof of the current **??** (in the appendix) had a small bug, which also has been fixed.

> The paper is very difficult to read. Improvements are needed in particular for the presentation of the approximation results and their proofs. Also for the notation. Missing definitions for used notions need to be added. Ideally use one instead of three query languages (UCQ, RA+, SPJU).

We have chosen one specific query language throughout the paper ($\mathcal{RA}^+$) and made a concerted effort to use clean, defined, non-ambiguous notation. We have also simplified the notation by limiting the paper's use of provenance semirings (which are needed solely for proofs) to the appendix. To the best of our examination, all notation conflicts have been addressed and definitions for used notions are added (see e.g. Definition C.4 appears before Lemma C.6 and Lemma C.8).

After the rewrite of Sec. 1, we had even less space for Sec. 4. However, we have modified Sec. 4 so that it flows better. In particular, we start off with the algorithm idea first (paragraph **Overview of our Techniques** in Sec. 1 also has more details on the intuition behind the approximation algorithm) and then state the results (with more details on how we argue the claimed runtime). Finally, we clearly state Corollary 4.9 for which queries our linear-time approximation result holds.

I.2 Reviewer 1

> l.24 "is #W[1]-hard": parameterized by what?

> l.103 and l.105: again, what is the parameter exactly?

While the above references do not exist in the revised Sec. 1 anymore, all theorem statements and claims on #W[1] runtime have been stated in a way so as to avoid ambiguity in the parameter. Please see e.g. Theorem 3.1 and Theorem 3.6.

> You might want to explain your title somewhere (probably in the introduction): in the end, what exactly should be considered harmful and why?

We have modified the title to be more descriptive.

> l.45 when discussing Dalvi and Suciu's dichotomy, you might want to mention that they consider *data complexity*. Currently the second sentence of your introduction ("take a query Q and a pdb D") suggests that you are considering combined complexity.

We have made an explicit mention of data complexity when alluding to Dalvi and Suciu's dichotomy. We have further rewritten Sec. 1 in such a way as to explicitly note the type(s) of complexity we are considering (mostly it's parameterized complexity).

> l.51 "Consider ... and tuples are independent random event": so this is actually a set PDB... You might want to use an example where the input PDB is actually a bag PDB. The last sentence before the example makes the reader *expect* that the example will be of a bag PDB that is not a set PDB

Our revision has removed the example referred to above. While the paper considers inputs to queries that are equivalent to set-PDB, this is not limiting. Please see **??** on **??**. Furthermore, we have added a discussion to the appendix that expands on why our results do extend beyond set inputs (Appendix A).

> - In the case of set semantics, the lineage of a tuple can be defined for *any* query: it is the unique Boolean function that satisfies the if and only if property that you mention on line 70. For bag semantics however, to the best of my knowledge there is no general definition of what is a lineage for an arbitrary query. On line 73, it is not clear at all how the polynomial should be defined, since this will depend on the type of query that you consider

Note that lineage for a set semantics query is as a positive Boolean formula is defined for positive relational algebra. For instance, for aggregate queries a more powerful model ([4]) is needed. The definition of the lineage polynomial (bag PDB) semantics over an arbitrary $\mathcal{RA}^+$ query Q is now given in Fig. 1. We also note that these semantics are not novel (e.g., similar semantics appear for both provenance [25] and probabilistic database [32, 19] contexts). However, as we were unable to find a formal proof of the equivalence between the expectation of the query multiplicity and of the lineage polynomial in related work, we have included a proof of Proposition 2.5.

> l.75 "evaluating the lineage of t over an assignment corresponding to a possible world": here, does the assignment assigns each tuple to true or false? In other words, do the variables X still represent individual tuples? From what I see later in the article it seems that no, so this is confusing if we compare to what is explained in the previous paragraph about set TIDB

The discussion after Problem 1.2 (in particular, the paragraph TIDB**s**) specifically address these questions. While values for possible worlds assigned are from $\{0, 1\}$, which is analog to Boolean, this is not limiting. Please see **??** (**??**) and the new appendix section Appendix A.

> - l.135 "polynomial Q(X)": Q should be reserved for queries... You could use φ or ϕ or... anything else but Q really

We now use $\Phi(\mathbf{X})$ for (lineage) polynomials.

> - If we consider data complexity (as did Dalvi and Suciu) and fix an UCQ Q, given as input a bag TIDB PDB we can always compute the lineage in $O(|D|^{|Q|})$ in SOP form and from there compute the expected multiplicity with the same complexity, so in polynomial time. How does this relate to your hardness result? Is it that you are only interested in combined complexity? Why one shouldn't be happy with this running time? Usually queries are much smaller than databases and this motivates studying data complexity.

We have rewritten Sec. 1 in a way to stress that we are are primarily interested in data complexity, but we cannot stop there. As the reviewer has noted, the problem we explore requires further analysis, where we require parameterized and fine grained complexity analysis to provide a theoretical foundation for the question we ask in **??**. We have discussed this in the prose following Problem 1.2.

> A discussion is missing about the difference between the approach usually taken in PDB literature and your approach. In which case would one be more interested in the expected multiplicity or in the marginal probability of a tuple? This should be discussed clearly in the introduction, as currently there is no clear "motivation" to what you do. There is a section about related work at the end but it is mostly a set of facts and there is no insightful comparison to what you do.

We provide more motivating examples in the first paragraph, and include a more detailed discussion of the relationship to sets in paragraph **Relationship to Set-Probabilistic Query Evaluation** after **??**. For example, expected multiplicities can model expectation

of a `COUNT(*)` query, while in many contexts computing the probability that this count is non-zero is not that useful.

As we now explain in the introduction, another motivation for generalizing marginal probability to expected multiplicity is that it is a natural generalization. The marginal probability of a tuple t is the expectation of a Boolean random variable that is assigned 1 in every world where tuple t exists and 0 otherwise. For bag-PDBs the multiplicity of a query result tuple can be modeled as a natural-number random variable that for a world D is assigned the multiplicity of the tuple in D. Thus, a natural generalization of the marginal probability (expectation of a Boolean random variable) to bags is the expectation of this variable: the tuple's expected multiplicity.

> l.176 "N[X] relations are closed under RA+": is this a *definition* of what it means to take an RA+ query and evaluate it over an N[X] database, or does this sentence say something more? Also, I think it would be clearer to use UCQs in the whole paper instead of constantly changing between UCQs, RA+ and SPJU formalisms

To make the paper more accessible and general, we found it better to not use $\mathbb{N}[\mathbf{X}]$-DBs. While we wanted to use UCQ, we found the choice of $\mathcal{RA}^+$ to be more amenable to the presentation of the paper, and have, as suggested stuck with one query formalism.

> There are too many things undefined in from l.182 to the end of page. l.182 and in Proposition 2.1 N-PDBs are not defined, the function mod is undefined, etc. The article should be self-contained: important definitions should be in the article and the appendix should only be used to hide proof details. I think it would not take a lot of space to properly define the main concepts that you are using, without hiding things in the appendix

All material in Sec. 2 that is proof-related is in the appendix, while Sec. 2 (modulo the proofs) is itself now self-contained.

> l.622 and l.632-634: so a N-PDB is a PDB where each possible world is an N-database, but an N[X]-PDB is not a PDB where each possible world is an N[X]-database... Confusing notation

The text now refers to latter as an $\mathbb{N}[\mathbf{X}]$-encoded PDBs.

> If you want to be in the setting of bag PDBs, why not consider that the value of the variables are integers rather that Boolean? I.e., consider valuations $\nu : X \to \mathrm{N}$ (or even to R, why not?) instead of $X \to \{0, 1\}$; this would seem more natural to me than having this ad-hoc "mix" of Boolean and non-Boolean setting. If you consider this then your "reduced polynomial" trick does not seem to work anymore.

Our objective is to establish the feasibility of bag-probabilistic databases as compared to existing deterministic query processing systems. Accordingly, we take our input model from production database systems like Postgresql, Oracle, DB2, SQLServer, etc. (e.g., see **??** on **??**), where duplicate tuples are represented as independent entities. As a convenient benefit, this leads to a direct translation of TIDBs (which are defined over $\{0, 1\}$ inputs). Finally, as we mention earlier, an easy generalization exists to encode a bag-PDB in a set-PDB (which then allows for bag inputs). Appendix A.

> - l.656 "Thus, from now on we will solely use such vectors...": this seems to be false. Moreover you keep switching notation which makes it very hard to read... Sometimes it is φ, sometimes it is small w, sometimes it is big W (l.174 or l.722), sometimes the database is $\varphi(D)$, sometimes it is $\varphi_w(D)$, other times it is $D_{[w]}$ (l.671), and so on.

We have made effort to be consistent with the use of notation, following standard usage whenever possible.

> l.658 "we use $\varphi(D)$ to denote the semiring homomorphism $\mathbb{N}[\mathbf{X}] \to \mathbb{N}$ that...": I don't understand why you need a database to extend an assignment to its semiring homomorphism from $\mathbb{N}[\mathbf{X}] \to \mathbb{N}$

φ [25] lifts the valuation function (with kind $\mathbb{N}[\mathbf{X}] \to \mathbb{N}$) to databases (i.e., a mapping from an $\mathbb{N}[\mathbf{X}]$-DB to a deterministic $\mathbb{N}$-DB). We note that the main body of the paper no longer references $\mathbb{N}[\mathbf{X}]$-DBs, and thus φ is discussed exclusively in the appendix.

> Figure 2, K is undefined

We have updated Fig. 1 (originally figure 2) to not need K.

> l.178 "Q_t", l.189 "Q will denote a polynomial": this is a very poor choice of notation

> l.242 "and query Q": is Q a query or a lineage?

We have reserved Q to mean an $\mathcal{RA}^+$ query and nothing else.

> Section 2.1.1: here you are considering set semantics no? Otherwise, one would think that for bag semantics the annotation of a tuple could be 0 or something of the form c × X, where X is a variable and c is a natural number

Please see Appendix A for a discussion on going beyond set inputs.

> Proof of Proposition A.3. I seems the proof should end after l.687, since you already proved everything from the statement of the proposition. I don't understand what it is that you do after this line.

This text is an informal proof of Proposition 2.5 originally intended to motivate Proposition B.3. We agree that this should not be part of the proof of the later, and have removed the text.

> l.686 "The closure of ... over K-relations": you should give more details on this part. It is not obvious to me that the relations from l.646 hold.

The core of this (otherwise trivial) argument, that semiring homomorphisms commute through queries, was already proven in [25]. We now make this reference explicit.

We apologize for not explaining this in more detail. In universal algebra [24], it has been proven (the HSP theorem) that for any variety, the set of all structures (called objects) with a certain signature that obey a set of equational laws, there exists a "most general" object called the *free object*. The elements of the free objects are equivalence classes (with respect to the laws of the variety) of symbolic expressions over a set of variables $\mathbf{X}$ that consist of the operations of the structure. The operations of the free object are combining symbolic expression using the operation. It has been shown that for any other object K of a variety, any assignment $\phi : \mathbf{X} \to K$ uniquely extends to a homomorphism from the free object to K by substituting variables for based on ϕ in symbolic expression and then evaluating the resulting expression in K.

Commutative semirings form a variety where $\mathbb{N}[\mathbf{X}]$ is the free object. Thus, for any polynomial (element of $\mathbb{N}[\mathbf{X}]$), for any assignment $\phi : \mathbf{X} \to \mathbb{N}$ (also a semiring) there exists a unique semiring homomorphism $\text{Eval}_\phi : \mathbb{N}[\mathbf{X}] \to \mathbb{N}$. Homomorphisms by definition commute with the operations of a semiring. Green et al. [**?**] did prove that semiring homomorphisms extend to homomorphisms over K-relations (by applying the homomorphism to each tuple's annotation) and these homomorphisms over K-relations commute with queries.

> l.711 "As already noted...": ah? I don't see where you define which subclass of N[X]-PDBs define bag version of TIDBs. If this is supposed to be in Section 2.1.1 this is not clear, since the world "bag" does not even appear there (and as already mentioned everything seems to be set semantics in this section). I fact, nowhere in the article can I see a definition of what are bag TIDBs/BIDBs

The new text precisely defines TIDBs (Sec. 1), and the BIDB generalization (Sec. 2.1.1). The specific text referenced in this comment has now been moved to the appendix and restructured to reference Definition B.2 (which defines an $\mathbb{N}[\mathbf{X}]$-encoded PDB defined over variables $\mathbf{X}$) and relate it to the formal structure of BIDBs in Sec. 2.1.1.

> - l.707 "the sum of the probabilities of all the tuples in the same block b is 1": no, traditionally it can be less than 1, which means that there could be no tuple in the block.

The reviewer is correct and we have updated our appendix text accordingly.

> it is not clear to me how you can go from l.733 to l.736, which is sad because this is actually the whole point of this proof. If I understand correctly, in l.733, Q(D)(t) is the polynomial annotation of t when you use the semantics of Figure 2 with the semiring K being N[X], so I don't see how you go from this to l.736

This result follows from the inner sum looping only over $\mathbf{w}$ s.t. $\psi_{\mathbf{w}}(\mathcal{D}_{\mathbb{N}[\mathbf{X}]}) = D$. As a consequence of this constraint, we have $Q(D_{\mathbb{N}[\mathbf{X}]})(t)(\mathbf{w}) = Q(D)(t)$. The latter term is independent of the summation, and so can be pulled out by distributivity of addition over multiplication.

We agree with the reviewer that this could be presented more clearly, and have now split the distributivity argument into a separate step.

> l.209-227: so you define what is a polynomial and what is the degree of a polynomial (things that everyone knows), but you don't bother explaining what "taking the mod of Q(X) over all polynomials in S" means? This is a bit weird.

Based on this and other reviewer comments, we removed the earlier definition of $\widetilde{\Phi}(\mathbf{X})$ and have defined it in a more ad-hoc manner, as suggested by the reviewers, including the comment immediately following.

> Definition 2.6: to me, using polynomial long division to define $\tilde{Q}$(X) seems like a pedantic way of reformulating something similar to Definition 1.3, which was perfectly fine and understandable already! You could just define $\tilde{Q}$(X) to set all exponents in the SOP that are >1 to 1 and to remove all monomials with variables from the same block, or using Lemma A.4 as a definition?

As alluded to above, we have incorporated the reviewer's suggestion, c.f. Definition 1.3 and **??**.

> Definition 2.14. It is not clear what is the input exactly. Are the query Q and database D fixed? Moreover, I have the impression that your hardness results have nothing to do with lineages and that you don't need them to express your results. I think the problem you should consider is simply the following: Expected Multiplicity Problem: Input: query Q, N[X]-database D, tuple t. Output: expected multiplicity of t in Q(D). Your main hardness result would then look like this: the Expected Multiplicity problem restricted to conjunctive queries is #W[1]-hard, parameterized by query size. Indeed if I look at the proof, all you need is the queries Q^k_G. The problem is #W[1]-hard and it should not matter how one tries to solve it: using an approach with lineages or using anything else. Currently it is confusing because you make it look like the problem is hard only when you consider general arithmetic circuits, but your hardness proof has nothing to do with circuits. Moreover, it is not surprising that computing the expected output of an arithmetic circuit is hard: it is trivial, given a CNF ϕ, to build an arithmetic circuit C such that for any valuation ν of the variables the formula ϕ evaluates to True under ν if C evaluates to 1 and the formula ϕ evaluates to False under ν if C evaluates to 0, so this problem is #P-hard anyways.

The reviewer is correct. Our hardness results are now stated independently of circuits. We note that the hardness result alluded to at the end of the comment above is not applicable in our case since for fixed queries Q, **??** and Problem 1.2 can be solved in polynomial time.

Further, as we point out in Sec. 1 what is new in our hardness results is that we show a query Q^k such that $T^*_{det}(Q^k, D_\Omega)$ is small (linear in $|D_\Omega|$ but solving **??** and Problem 1.2 is hard. We note that it is well-known that one can reduce the problem of counting k-cliques or k-matchings to a query Q for which computing $Q(D_\Omega)$ is #W[1]-hard. So our contribution to come up with a different reduction from counting k-matchings so that the hardness manifests itself in the probabilistic computing part of our problem.

> Section 3.3. It seems to me the important part of this section is not so much the fact that we have fixed values of p but that the query is now fixed and that you are looking at the fine-grained complexity. If what you really cared about was having fixed value of p, then the result of this section should be exactly like the one in Theorem 3.4, but starting with "fix p". So something like "Fix p. Computing $\tilde{Q}^k_G$ for arbitrary G is #W1-hard".

We agree with the reviewer that the result on fixed value of p is mostly of (narrow) theoretical interest. We have added a discussion summarizing the reviewer's point above below Theorem 3.7.

> General remark: The story of the paper I think should be this: we can always compute the expected multiplicity for a UCQ Q and N[X]-database D and tuple t by first computing the lineage in SOP form and then using linearity of expectation, which gives an upper bound of (roughly) $O(|D|^{|Q|})$. We show that this exponential dependence in |Q| is unavoidable by proving that this problem is #W1 hard parameterized by |Q| (which implies that we cannot solve it in $f(|Q|)|D|^c$). Furthermore we obtain fine-grained superlinear lower bounds for a fix conjunctive query Q. (Observe how up to here, there is no need to talk about lineages at all). We then obtain an approximation algorithm for this problem for [this class of queries] and [that class of bag PDBs] with [that running time (Q,D)]. The method is to first compute the lineage as an arithmetic circuit C in [this running time (Q,D)], and then from the arithmetic circuit C compute in [running time(C)] an approximation of its expected output. Currently I don't understand to which queries your approximation algorithm can be applied (see later comments).

We have restructured Sec. 1 to more or less follow the reviewer's outline above. The only deviation is that we still introduce lineage polynomials. We do this because the polynomial

view is very helpful in the proofs of our hardness result (in addition to the obvious relevance for the approximation algorithm). We have also clarified that our approximation result applied to all $\mathcal{RA}^+$ queries (see Corollary 4.9).

> l.381: Here again, I think it would be simpler to consider that the input of the problem is the query, the database and a tuple and claim that you can compute an approximation of the expected multiplicity in linear time. The algo is to first compute the lineage as an arithmetic circuit, and then to use what you currently use (which could be put in a lemma or in a proposition).

We have implemented the above overview in Sec. 1 when we move from **??** to Problem 1.6. For the approximation algorithm we focus on Problem 1.6, which still takes a circuit as an input.

> Definition 4.2: would you mind giving an intuition of what this is? It is not OK to define something and just tell the reader to refer the appendix to understand what this is and why this is needed; the article should be understandable without having to look at the appendix. It is simply something that gives the coefficient of each monomial in the reduced polynomial?

We have provided an example in directly after Definition 4.1 as well as a sentence pointing out why this definitions is useful.

> - l.409: how does it matter that the circuit C is the lineage of a UCQ? Doesn't this work for any arithmetic circuit?

The reviewer is correct that the earlier Theorem 4.9 works for any circuit (this result is now in the appendix).

> l.411: what are $|C|^2(1,...,1)$ and $|C|(1,...,1)$?

We clarify this overloaded notation immediately after Definition 4.2.

> Sometimes you consider UCQs, sometimes RA+ queries. I think it would be simpler if you stick to one formalism (probably UCQs is cleaner?)

As alluded to previously, we have followed the reviewer's suggestion and have found $\mathcal{RA}^+$ queries to be most amenable for this work.

> l.432 what is an FAQ query?

We actually no longer need that result since Lemma 4.8 now has a bound on $|\mathtt{C}|(1,\ldots,1)$ in terms of $\textsc{depth}(\mathtt{C})$ and the latter is used in Corollary 4.9 for all $\mathcal{RA}^+$ queries. Please see Lemma 4.8 and the followup discussion for more on this.

> Generally speaking, I think I don't understand much about Section 4, and the convolutedness of the appendix does not help to understand. I don't even see in which result you get a linear runtime and to which queries the linear runtime applies. Somewhere there should be a corollary that clearly states a linear time approximation algorithm for some queries.

We have re-organized sec:algo to address the above comments as follows:

- We now start off Sec. 4.2 with the algorithm idea.
- We give a quick overview of how the claimed runtime follows from the algorithm idea mentioned above.
- Added Corollary 4.9 that clearly states that we get an $O(T^*_{det}(Q, D_\Omega))$ for *all* $\mathcal{RA}^+$ queries Q.

> In section 5, it seems you are arguing that we can compute lineages as arithmetic circuits at the same time as we would be running an ordinary query evaluation plan. How is that different from using the relations in Figure 2 for computing the lineage?

There is not a major difference between the two. This observation has persuaded us to eliminate $\mathbb{N}[\mathbf{X}]$-DB query evaluation and have only an algorithm for lineage.

We have also re-organized the earlier Section 5 and moved the definition of $T^*_{det}(\cdot)$ (earlier denoted as $\mathbf{cost}(\cdot)$) to Sec. 2.3 and moved the rest of the material to the appendix.

> l.679 where do you use $max(D_i)$ later in the proof?

Thank you. This reference was unnecessary and has been removed.

> l.688 That sentence is hard to parse, consider reformulating it

As the reviewer notes above, this paragraph is unnecessary and we have removed it.

> it seems you are defining N[X]-PDB at two places in the appendix: once near l.632, and another time near l.652

Thank you. The latter definition has been removed.

I.3 Reviewer 2

> First, the paper should state rigorously the problem definition. There are three well-known definitions in database theory: data complexity, combined complexity, and parameterized complexity. If I understand correctly, Theorem 3.4 refers to the parameterized complexity, Theorem 3.6 refers to the data complexity (of a fixed query), while the positive results in Sec. 4 (e.g. Th. 4.8) introduce yet another notion of complexity, which requires discussion.

We have addressed the concerns in rewriting the entirety of Sec. 1, explicitly mentioning complexity metrics considered, while forming a series of problem statements that describe the exact problem we are considering, and the complexity metrics considered. We have also adjusted the phrasing of the said theorems and definitions to eliminate the ambiguity described.

> The problem definition is supposed to be in Definition 2.14, but this definition is sloppy. It states that the input to the problem is a circuit C: but then, what is the role of the PDB and the query Q? Currently Definition 2.14 reads as follows: "Given a circuit C defining some polynomial Q(X), compute E[Q(W)]", and, thus, the PDB and the query play no role at all. All results in Section 4 seem to assume this simplified version of Definition 2.14. On the other hand, if one interprets the definition in the traditional framework of data complexity (Q is fixed, the inputs are D and C) then the problem is solvable in PTIME (and there is no need for C), since E[Q(W)] is the sum of expectations of the monomials in Q (this is mentioned in Example 1.2).

We have rephrased Definition 2.9 to qualify data complexity. The paper (especially in Sec. 1) builds up the fact that we aren't stopping at polynomial time, but exploring parameterized complexity and fine grained analysis (as the reviewer aptly noted in the first comment).

> Second, Definition 2.6 of Reduced BIDB polynomials is simply wrong. It uses "mod" of two multivariate polynomials, but "mod" doesn't exists for multivariate polynomials...Either state Definition 2.6 directly, in an ad-hoc manner (which seems doable), or do a more thorough job grounding it in the ring of multivariate polynomials and its ideals.

The reviewer is correct in their comment on the "mod" part– we apologize for the error. We have implemented the reviewer's ad-hoc suggestion in light of Reviewer 1's similar suggestions.

> the paper uses three notations (UCQ, RA+, SPJU) for the same thing, and never defines formally any of them.

We have chosen $\mathcal{RA}^+$ for consistent use throughout the paper. We have included **??** on **??** for an explicit definition of $\mathcal{RA}^+$ queries.

> G^ℓ is used in Lemma 3.8 but defined only in the Appendix (Def. B.2), without even a forward pointer. This is a major omission: Lemma 3.8 is a key step for a key result, but it is impossible to read.

We have fixed this mistake. Unfortunately, because of the changes in the paper (especially expanding on Sec. 1), the earlier Lemma 3.8 had to be moved to the appendix.

> Definition 2.7. "valid worlds η". This is confusing. A "possible world" is an element of Ω: this is not stated explicitly in the paper, but it is implicit on line 163, so I assumed that possible worlds refer to elements of Ω. If I assumed correctly, then calling η a "world" in Def. 2.7 is misleading, because η is not an element of Ω. More, it is unclear to me why this definition is needed: it is used right below, in Lemma 2.8, but that lemma seems to continue to hold even if w is not restricted.

We agree with the reviewer that this notation is confusing; η is meant to cope with the fact that tuples from the same group in a BIDB can not co-exist, even though our $\{0,1\}$-input vectors can encode such worlds. We now address this constraint by embedding it directly into the reduced polynomial with **??**.

> line 305: please define what is an "occurrence of H in G". It could mean: a homomorphic image, a subgraph of G isomorphic to H, an induced subgraph of G isomorphic to H, or maybe something else.

We agree with the reviewer's suggestion and have rephrased the wording to be clear. Please see the beginning of Sec. 3.1.

> If the proofs are given in the appendix, please say so. Lemmas 3.5 and 3.8 are stated without any mention, and one has to guess whether they are obvious, or proven somewhere else. On this note: I found Lemma 3.5 quite easy, since the number of k-matching is the coefficient of the leading monomial (of degree 2k) in $Q^k(p, p, ..., p)$, while Lemma 3.8 appears much harder. It would help to briefly mention this in the main body of the paper.

We have implemented the reviewer's suggestion. Please see the last sentence of Sec. 1 (as well as the expanded discussion on the hardness result in the **Overview of our Techniques** part of Sec. 1).

> line 177: what is $\Omega_{\mathbb{N}[\mathbf{X}]}$?

We have eliminated the use of $\mathbb{N}[\mathbf{X}]$-DBs in the paper proper, using them only when necessary in the proofs of the appendix.

> line 217. The polynomial $X^2 + 2XY + Y^2$ is a poor choice to illustrate the degree. There are two standard definitions of the degree of a multivariate polynomial, and one has to always clarify which one is meant. One definition is the total degree (which is Def. 2.3 in the paper), the other is the maximum degree of any single variable. It is nice that you are trying to clarify for the reader which definition you are using, but the polynomial $X^2 + 2XY + Y^2$ is worst choice, since here the two coincide.

We have adjusted the example to account for the reviewer's correct observation.

> line 220. "we consider only finite degree polynomials". This is a surprise. Polynomials, by definition, are of finite degree; there are extensions (I'm aware of powerseries, maybe you have other extensions in mind), but they are usually not called polynomials, plus, nothing in the paper so far suggests that it might refer to those extensions.

We have removed the redundant terminology the reviewer has pointed out, and refined the discussion surrounding (and including) Eq. (1) to be explicit to the novice reader that polynomials are by definition of finite degree.

> "Note that our hardness results even hold for the expression trees". At this point we haven't seen the hardness results, nor their proofs, and we don't know what expression trees are. It's unclear what we can note.

Our hardness results are now stated independently of circuits so the above statement no longer appears in the paper.

> paragraph at the top of pp.10 is confusing. My guess is that it is trying to this say: "there exists a query Q, such that, for each graph G, there exists a database D s.t. the lineage of Q on D is the polynomial Q_G."

Our revision has eliminated this statement.

I.4 Reviewer 3

> The overall study is then extended to a multiplicative approximation algorithm for the expectation of polynomial circuits in linear time in the size of the polynomial. It was much harder to read this part, and I found the examples and flow in the appendix quite helpful. I suggest to include these examples into the body of the paper.

In our revision we expanded on Sec. 1 to give a better overview of the problems we are considering in this paper. This meant we had to cut out material in later sections, which unfortunately meant we did not have space in Sec. 4 to include any examples that the reviewer suggested above. However, we have tried to make Sec. 4 more readable as a whole.

> While ApproximateQ is linear in the size of the circuit, it is quadratic in epsilon and so we need quadratically many samples for the desired accuracy – overall runtime is not linear therefore and it may be better to elaborate this. It may also be helpful to comment on how this relates to Karp, Luby, Madras algorithm [1] for #DNF which is also quadratic in epsilon.

In Problem 1.5 we note explicitly that we care about linear dependence on $T^*_{det}((Q, D_\Omega))$ and do not care about the exact dependence on ϵ. While it would be nice to design an approximation algorithm that is linear in $1/\epsilon$ as well, we believe it is out of scope of this initial work.

> The coverage of related work is adequate. Fink et. al seems as the closest related work to me and I would appreciate a more elaborate comparison with this paper. My understanding is that Fink et. al considers exact evaluation only and focuses on knowledge compilation techniques based on decompositions. They also note that "Expected values can lead to unintuitive query answers, for instance when data values and their probabilities follow skewed and non-aligned distributions" attributed to [2]. Does this apply to the current work? Can you please comment on this?

The work is indeed quite close to our own. It targets a broader class of queries (aggregates include COUNT/SUM/MIN/MAX, rather than our more narrow focus on COUNT), but has significantly less general theoretical results. Most notably, their proof of linear runtime in the

size of the input polynomial is based on a tree-based encoding of the polynomial. Tree-based representation representation (and hence the Fink et. al. algorithm's runtime) is, as we note several times, superlinear in $T^*_{det}(Q, D_\Omega)$. This result is also limited to a specific class of (hierarchical) queries, devolving to exponential time (as in [20]) in general. By contrast, our results apply to all of $\mathcal{RA}^+$. Our revised related work section now addresses both points.

> I assume the authors focus on parameterized complexity throughout the paper, and even this is not stated unambiguously. The authors should make an extra effort to make the paper more accessible by using the explanations and examples from the appendix in the body of the paper. It is also important to highlight the differences with the complexity of standard query evaluation over PDBs.

Our revision has focused on explicitly mentioning the complexity metrics we are interested in. This can be seen in e.g. Sec. 1 and formal statement (theorems, lemmas, etc.), which have been rewritten to eliminate ambiguities. We have also taken pains to be promote accessibility, keeping the paper self-contained, and using examples for difficult or potentially unclear concepts. This can be seen in e.g. eliminating unnecessary machinery (e.g. $\mathbb{N}[\mathbf{X}]$-DB machinery from the paper proper), providing/modifying examples (c.f. Definition 4.1, Definition 4.4), and ensuring consistency in notational use, e.g. using one query evaluation formalism ($\mathcal{RA}^+$).

We decided to focus on beefing up Sec. 1 and cleaning up definitions of problems, which unfortunately meant we ran out of space to bring back examples from the appendix (especially into Sec. 4).

I.5 Reviewer 4

> I wonder whether the writing could be revisited to give the reader a better overview of the technical challenges, motivation, and the high level ideas of the algorithm and hardness results. The current exposition seems slightly too tailored for the expert in probabilistic databases rather than the average ICDT attendee. Also the current exposition is structured such that the reader needs to get through quite a few definitions and technical lemmas until they get to the new ideas in the paper.

We have (as noted throughout this section) revised the writing to provide precision and clarity to the problem we explore as well as the results we obtain. Part of this revision was a complete rewriting of Sec. 1 where we sought to be extremely precise in language and through a series of problem statements to help the reader navigate and understand the problem we explore as well as how we have gone about exploring that problem coupled with the validity of the exploration strategy. We have simultaneously sought to make the paper more accessible by assuming the average ICDT attendee and defining or explaining concepts that might not be known to them. Finally, we have expanded on the **Overview of our Techniques** part of Sec. 1 to provide more intuition on how we prove our lower and upper bounds.