Proceedings of Machine Learning Research vol 275:874–893, 2025 4th Conference on Causal Learning and Reasoning

The Causal-Effect Score in Data Management

Felipe Azúa FAZUA@ALUMNOS.UAI.CL
Universidad Adolfo Ibáñez, Santiago, Chile.

Leopoldo Bertossi* BERTOSSI@SCS.CARLETON.CA
Universidad San Sebastián, Santiago, Chile.

Editors: Biwei Huang and Mathias Drton

Abstract

The *Causal Effect* (CE) is a numerical measure of causal influence of variables on observed results. Despite being widely used in many areas, only preliminary attempts have been made to use CE as an attribution score in data management, to measure the causal strength of tuples for query answering in databases. In this work, we introduce, generalize and investigate the so-called *Causal-Effect Score* in the context of classical and probabilistic databases.

Keywords: Causality, Probabilistic Databases, Attribution Scores

1. Introduction

Users of a database management system could expect *explanations*, for example, for the answers to a query, for a violation of an integrity constraint, etc. So as in AI and machine-learning, explanations may come in different forms, in particular, as an *attribution score*, that is, as a quantitative degree of relevance of a piece of data in a database (DB) in relation to the observed output (or for the feature values in input entities in the case of AI and ML). This work concentrates on *local attribution scores* as explanations in DBs, that is, for a particular query answer, and at the DB tuple level.

Several scores have been proposed and investigated in DBs. *Responsibility* (Chockler and Halpern, 2004) is based on *actual causality* and *counterfactual interventions* (Halpern and Pearl, 2005; Halpern, 2015, 2016). The latter can be seen as hypothetical changes performed on a (causal) model, to identify other changes. By doing so, one can detect cause-effect relationships. In the context of actual causality, *Responsibility* (RESP) has been applied in DBs to quantify the relevance of individual tuples, and attribute values in them, for a query result (Meliou et al., 2010a,b; Bertossi and Salimi, 2017a,b). The *Shapley Value* of *Coalition Game Theory* (Shapley, 1953; Roth, 1988), as a measure of contribution of individual players to a shared wealth or game function, has been applied as an explanation score in DBs, with tuples acting as players, and the query (Boolean or aggregate) as the game function (Livshits et al., 2021a,b; Deutch et al., 2022; Bertossi et al., 2023b). The Shapley Value is as the only measure that satisfies a given set of desired properties (or *axioms*) (Roth, 1988). A related game-theoretic measure, the *Banzhaf Power Index* (BPI) (Banzhaf, 1964), has been applied in DBs as the *Banzhaf Score* (Livshits et al., 2021a; Abramovich et al. , 2024).

The *Causal-Effect Score* (CES) can be traced back to *causality in observational studies* (Rubin, 1974; Holland, 1986), where one usually cannot predefine and build control groups, but they have to be recovered from the available data (Gelman and Hill, 2007; Pearl, 2009; Roy and Salimi, 2023). The CES is also based on interventions. In (Salimi et. al., 2016), an appropriate form of the CES was first used in data management, when RESP did not provide intuitive results. Furthermore, CES

* Professor Emeritus, Carleton University, Ottawa, Canada

© 2025 F. Azúa & L. Bertossi.

is more appropriate than RESP for explaining aggregate queries. RESP would only consider, say under a counterfactual tuple-deletion, if there is a change in the numerical value, no matter how small. CES would take the amount of change into consideration.

In order to apply the CES with a query posed to a DB D (Salimi et. al., 2016), D is first transformed into a *tuple-independent probabilistic DB* (TID) D^p (Suciu et al., 2011) that *independently* and *uniformly* assigns a probability of $\frac{1}{2}$ to all tuples. The "counterfactual game" is played on D^p. The CES can accommodate both *endogenous* and *exogenous* tuples, which have also been considered with RESP and the Shapley Value in DBs (Meliou et al., 2010a; Bertossi and Salimi, 2017a; Livshits et al., 2021a). Sometimes we will denote this case of the CES with CES(I,U($\frac{1}{2}$)).

In this work, we propose and investigate a *Generalized Causal-Effect Score* (GCES) in DBs. Generalizing CES(I,U($\frac{1}{2}$)), it can be applied to arbitrary probabilistic DBs. We introduce the GCES on the basis of *counterfactually intervened distributions*. The GCES becomes interesting in several situations, among them, when: (a) One cannot assume independence or a uniform distribution on tuples, in particular with *block-independent* PDBs (Suciu et al., 2011). (b) Additional domain semantics or domain knowledge can be taken into account, e.g. integrity constraints on the DB (this was investigated for RESP in (Bertossi and Salimi, 2017b)). ICs can be compiled into an updated joint distribution (Bertossi, 2023a). (c) There are explicit correlations among database tuples (Sen and Deshpande, 2007; Kanagal and Deshpande, 2010). (d) There are quantitative or qualitative stochastic (in)dependencies among attributes provided by a Bayesian or causal network (Sen et al., 2009). (e) With crowd-sourced data that come with different trustworthiness scores or; and, more generally, when data from sources of varying reliability are combined (Ratner et al., 2020).

The GCES can be defined for any query that returns a numerical value (0 or 1 for Boolean query, and a number for a scalar aggregate query). However, we restrict its analysis to monotone queries, which include those most common queries. This is a common practice in database research, as a first step, to gain insight before moving to more general queries. We investigate the data complexity of computing causal-effect scores as explanations for query results, establishing a dichotomy for Boolean Conjunctive Queries (BCQs). Furthermore, taking inspiration from the categorical axiomatization of the Shapley value (Roth, 1988), we uncover an axiomatic characterization of the GCES as applied to DBs and monotone queries. We compare these properties with those of the Shapley Value and the BPI. Notice that, for query explanations in DBs, the latter coincides with CES(I,U($\frac{1}{2}$)) (Livshits et al., 2021a).

This paper is structured as follows. Section 2 provides background on databases, probabilistic databases, and the Shapley Value and BPI. It also describes related work. Section 3 introduces the Generalized Causal Effect Score. Section 4 investigates the complexity of computing the GCES. Section 5 analyzes the general properties of the GCES, and its axiomatization. Section 6 points to ongoing and future research. Proofs can be found in the Appendix.

2. Background

Relational Databases. For causality purposes, some of the tuples in a DB D are considered to be *endogenous*. They can be subject to causal (more precisely, counterfactual) *interventions*, in this case, deletions, insertions, or value updates. The other tuples are *exogenous*, and are taken as given. They may participate in query answering, but they are not subject to interventions. D^{en}, D^{ex} denote the subinstances containing the endogenous, resp. exogenous tuples. With $Adom(D)$ we denote the *active domain* of D, formed by all the constants appearing in D.

In this work, we will consider *Boolean conjunctive queries* (BCQs), unions thereof (UBCQs), which all take values 0 or 1 on DBs; and also, numerical aggregations on CQs. $D \models \mathcal{Q}$ denotes that instance D makes $\mathcal{Q}$ true. $\mathcal{Q}[D]$ denotes the answer to query $\mathcal{Q}$ on instance D. A query $\mathcal{Q}$ is *monotone* when, for every $D_1 \subseteq D_2$, it holds $\mathcal{Q}[D_1] \leq \mathcal{Q}[D_2]$. BCQs and UBCQs are monotone. All the computational complexity results refer to *data complexity*, i.e. in the size of the database.

Probabilistic Databases. We recall here what we need about probabilistic databases (PDBs). See (Suciu et al., 2011) for a deeper treatment. For an initial intuition, we can conceive a PDB as a regular relational DB whose relations have an extra attribute accommodating probability values associated to the corresponding tuples. Table 1 shows a PDB of this kind. For example, the first tuple in the relation R indicates that tuple τ_1 belongs to the relation with probability p_1.

R	A	B	C	P
τ_1	a_1	b_1	c_1	p_1
τ_2	a_2	b_2	c_2	p_2
τ_3	a_3	b_3	c_3	p_3
τ_4	a_4	b_4	c_4	p_4
τ_5	a_5	b_5	c_5	p_5

Table 1: A TID

The semantics of a PDB, D^p, is a *possible world semantics*, in this case a collection $\mathcal{W}$ of subinstances, W, of D^p whose relations, R_W, do not have a probabilistic attribute, but a global probability, $p(R_W)$. Different semantics differ on how the instances D are built and how probabilities are assigned to their relations.

The most common case is that of a *tuple-independent* PDBs (TIDs). In Table 1, each tuple is (or is not) in the DB independently from the other tuples in the relation and in other relations. Each tuple has a probability assigned. In a possible world $W \in \mathcal{W}$, the corresponding (non-probabilistic) relation R_W will contain only some of the tuples in R, and the probability associated to R_W is defined by: $p(R_W) := \Pi_{\tau_i \in R_W} p_i \ \times \ \Pi_{\tau_j \in (R \smallsetminus R_W)}(1-p_j)$. For example, if in a possible world W relation R_W contains only tuples τ_1 and τ_3, it will have the probability $p_1 \times (1-p_2) \times p_3 \times (1-p_4) \times (1-p_5)$. Tuple independence beyond the single-relation level leads to the overall probability assigned to a possible world: $p(W) := \Pi_{R_W} p(R_W)$.

More generally, for the purpose of this work, and not necessarily in the TID case, a PDB D^p can be identified with a discrete *probability space* $\langle \mathcal{W}, p\rangle$, where $\mathcal{W}$ is the collection of *possible worlds* W that are regular relational subinstances of D^p, p is defined on $\mathcal{W}$, and $\sum_{W \in \mathcal{W}} p(W) = 1$. Since, D^p is finite, $\mathcal{W}$ and every world W in it are also finite.

Given a PDB D^p, and a tuple τ, the probability of τ being in D^p, is defined by:

$$P(\tau) \ := \sum_{W \in \mathcal{W}\colon\, \tau \in W} p(W). \tag{1}$$

Remark 1 If there is a partition (D^{en}, D^{ex}) of D^p, we assume that, for every $W \in \mathcal{W}$ with $D^{ex} \not\subseteq W$, it holds $p(W) = 0$. From this we obtain that, for every $\tau \in D^{ex}$, $P(\tau) = 1$. ■

A BCQ on a PDB becomes a Bernoulli random variable, taking values 0 or 1 on the outcomes $W \in \mathcal{W}$. There are several query-answering semantics that have been considered in the literature. We briefly mention the one that becomes relevant in our work.

Let $\mathcal{Q}$ be a query for schema $\mathcal{S}$. If $\mathcal{Q}$ is a BCQ, the probability of $\mathcal{Q}$ (being true) is $P(\mathcal{Q}, D^p) := \sum_{W\in\mathcal{W}:\, W\models\mathcal{Q}} p(W)$. If $\mathcal{Q}(\bar{x})$ is an open query, and $\bar{a}$, a sequence of constants, $P(\bar{a}) := P(\mathcal{Q}[\bar{a}])$. Under this semantics, each answer comes with a probability (of being an answer). Notice that for a Boolean query, $P(\mathcal{Q}, D^p) = \mathbb{E}(\mathcal{Q})$, which invites us to define, for an *aggregate query* $\mathcal{Q}$ a probabilistic answer: $\mathcal{Q}[D^p] := \mathbb{E}(\mathcal{Q}) = \sum_{W\in\mathcal{W}} p(W) \times \mathcal{Q}[W]$. We will often omit the database when it is clear from the context, simply writing $P(\mathcal{Q})$.

Shapley Value and Banzhaf Power-Index in Databases. The Shapley Value (Shapley, 1953) was used in (Livshits et al., 2021a) to quantify the contribution of tuples to a query answer (see also (Livshits et al., 2021b; Bertossi et al., 2023b)), as follows:

$$Shapley(D, \mathcal{Q}, \tau) = \sum_{S\subseteq(D^{en}\smallsetminus\{\tau\})} \frac{|S|!\cdot(|D^{en}|-|S|-1)!}{|D^{en}|!} \cdot \Delta(\mathcal{Q}, S, \tau), \tag{2}$$

with $\Delta(\mathcal{Q}, S, \tau) := \mathcal{Q}[S \cup D^{ex} \cup \{\tau\}] - \mathcal{Q}[S \cup D^{ex}]$. When $\mathcal{Q}$ is Boolean, $\mathcal{Q}[S]$ is 0 or 1. When it is a numerical aggregation, $\mathcal{Q}[S]$ is the resulting value. The *Bahnzhaf Power-Index* (BPI) (Banzhaf, 1964) is defined, for queries in DBs, by:

$$BPI(D, \mathcal{Q}, \tau) := \sum_{S\subseteq D^{en}\smallsetminus\{\tau\}} \frac{1}{2^{|D^{en}|-1}} \cdot \Delta(\mathcal{Q}, S, \tau). \tag{3}$$

The Shapley *is the only measure* of contribution that satisfies certain desirable properties (Shapley, 1953; Roth, 1988). The more relaxed definition of the BPI (permutations are not considered) makes it miss some of the properties of the Shapley value (see Section 5).

Related Work. In (Livshits et al., 2021a), the Shapley value and the BPI were applied and investigated in data management, to quantify the contribution of individual tuples to a query answer (see also (Livshits et al., 2021b; Bertossi et al., 2023b). In (Bienvenu et al., 2024), the complexity of the Shapley value in DBs is further explored by establishing a connection with the Generalized Model Counting Problem (the number of subinstances of a given size that satisfy the query). In this line of research, (Deutch et al., 2022) concentrates on computational aspects of the Shapley value applied to query answering.

In (Arad et al., 2022), the *rankings* induced by the Shapley values are investigated, rather than on the values themselves. In (Davidson et al., 2022) the interest is in the combination of *data provenance* and the Shapley values, for the computation of the latter. In (Abramovich et al. , 2024), research delves more deeply into experimental results around the BPI for query answering. In (Kara et al., 2024), the problem of computing the Shapley value of variables in Boolean circuits is connected with query evaluation in PDBs, obtaining dichotomy results (similar to those in (Dalvi and Suciu, 2012)) for the complexity of Shapley value computation for query answering in DBs.

Applications of the responsibility score in data management (DM) have been less explored than those of the Shapley value. References for the use of responsibility in DM are (Meliou et al., 2010b; Bertossi and Salimi, 2017a,b), among others. In all those papers, semantic and computational problems were addressed. Close to responsibility as used in DM, we find the notion of *resilience*, whose computational aspects have been investigated (Makhija and Gatterbauer, 2023).

The use of the *causal effect* in DM has been explored, and only superficially and in its simplest form, i.e. the CES(I,U($\frac{1}{2}$)), in (Salimi et. al., 2016). The connection with the BPI was first established in (Livshits et al., 2021a). The complexity of CES(I,U($\frac{1}{2}$)) has been partially and indirectly studied in (Livshits et al., 2021a), through its connection to the BPI.

Closer in spirit to our work, in that a probabilistic setting is assumed at the start, (Karmakar et al., 2024) investigate the expected value of the Shapley value on tuple-independent PDBs.

3. The Causal-Effect Score in Databases

In data management, we usually want to explain *why* a given query becomes true or returns a particular answer. In this work, we want to provide *explanations at the tuple-level*. In (Salimi et. al., 2016), it was shown that the *Causal-Effect Score* (in its CES(I,U($\frac{1}{2}$)) form) can be sensibly applied as an explanation score that reflects the causal strength of a tuple for the query answer.

3.1. Interventional Distributions on PDBs

The causal effect relies on *interventions*, which in DBs become hypothetical insertions or deletions of tuples. If we start with a PDB D^p, with a given distribution p, an intervention induces a new, *interventional distribution* (see Definition 2). Interventions will be denoted with $do(\tau\ \mathsf{in})$ and $do(\tau\ \mathsf{out})$, with the intuitive meaning that tuple τ is *made true*, i.e. it is inserted into D^p (if it is not already in it). Similarly, $do(\tau\ \mathsf{out})$ means that τ is *made false*, i.e. removed from D^p. Interventions are applied only with endogenous tuples, which have an initial probability of being (true) in D^p. Interventions are applied to detect if making a tuple true or false affects a query answer. They can also be applied with sets of tuples $\mathcal{T}$, e.g. $do(\mathcal{T}\ \mathsf{in})$.[1]

We will use expressions of the form $P(\mathcal{Q} = 1 \mid do(\tau\ \mathsf{in}))$, etc., where $\mathcal{Q}$ is a Boolean query, and P is the intervened distribution on the query range associated to a PDB D^p. Intuitively, it means "the probability of the query being true given that tuple τ is made true". This common notation may be misleading, suggesting a conditional probability, which strictly speaking is not.

More precisely, in Definition 2, we start with a distribution p on the class of possible worlds $\mathcal{W}$ associated to D^p, and we create the intervened distributions $p^{+\tau}(W) := p(W \mid do(\tau\ \mathsf{in}))$ and $p^{-\tau}(W) := p(W \mid do(\tau\ \mathsf{out}))$ on a collection of possible worlds. They can be seen as modifications of the original distribution p.[2]

Definition 2 Given a PDB $D^p = \langle \mathcal{W}, p\rangle$, and a tuple τ: (a) The positive intervention with τ on D^p is the PDB: $D^p(do(\tau\ \mathsf{in})) := \langle \mathcal{W}^{+\tau}, p^{+\tau}\rangle$, with $\mathcal{W}^{+\tau} := \{W \cup \{\tau\} \mid W \in \mathcal{W}\}$; and, for each $W' \in \mathcal{W}^{\tau +}$, $p^{+\tau}(W') := \sum_{W\cup\{\tau\}=W'} p(W)$.
(b) Similarly, for a negative intervention: $D^p(do(\tau\ \mathsf{out})) := \langle \mathcal{W}^{-\tau}, p^{-\tau}\rangle$, with $\mathcal{W}^{-\tau} := \{W \smallsetminus \{\tau\} \mid W \in \mathcal{W}\}$; and, for each $W' \in \mathcal{W}^{\tau -}$, $p^{-\tau}(W') := \sum_{W\smallsetminus\{\tau\}=W'} p(W)$.
(c) For a Boolean query $\mathcal{Q}$, its intervened probabilities are: $P(\mathcal{Q} = 1 \mid do(\tau\ \ \mathsf{in})) := p^{+\tau}(\{W' \in \mathcal{W}^{+\tau} \mid W' \models \mathcal{Q}\})$, $P(\mathcal{Q} = 0 \mid do(\tau\ \ \mathsf{in})) := p^{+\tau}(\{W' \in \mathcal{W}^{+\tau} \mid W' \not\models \mathcal{Q}\})$, $P(\mathcal{Q} = 1 \mid do(\tau\ \mathsf{out})) := p^{-\tau}(\{W' \in \mathcal{W}^{-\tau} \mid W' \models \mathcal{Q}\})$, and $P(\mathcal{Q} = 0 \mid do(\tau\ \mathsf{out})) := p^{-\tau}(\{W' \in \mathcal{W}^{-\tau} \mid W' \not\models \mathcal{Q}\})$. ∎

1. It is common to apply interventions on variables of a causal model (Pearl, 2009). In the case of databases, to the *lineage of the query*, that can act as the model, with random propositional variables X_τ, which are made *true* or *false* via $do(X_\tau = 1)$ or $do(X_\tau = 0)$ (Salimi et. al., 2016). It should be noted that the notion of causality considered here does not attempt to model some kind of causal structure in the "real world", but is based on the causal structure between the input and output of a database query inside the computer.
2. The starting set $\mathcal{W}$ of possible world contains subinstances of instance D^p whose probabilities add up to 1. However, it does not have to contain all subinstances (but commonly it does). Similarly, for the collection of possible worlds after an intervention.

It holds: (similarly for other cases)

$$p^{+\tau}(\{W' \in \mathcal{W}^{+\tau} \mid W' \models \mathcal{Q}\}) \;=\; \sum_{W \in \mathcal{W}:\; W \cup \{\tau\} \models \mathcal{Q}} p(W).$$

In case (c), P is the probability induced by, e.g., $p^{+\tau}$ on the range $\{0, 1\}$ of the query.

Remark 3 (a) It follows from (1) that, for an endogenous tuple τ: $P(\tau \mid do(\tau\ \mathsf{in})) = 1$; and $P(\tau \mid do(\tau\ \mathsf{out})) = 0$, which captures the original intuition. (b) For a TID D^p and two tuples $\tau, \tau' \in D^{en}$, it holds: $P(\tau' \mid do(\tau\ \mathsf{in})) = P(\tau' \mid do(\tau\ \mathsf{out})) = p(\{\tau'\})$, that is, an intervention $do(\tau\ \mathsf{in})$ ($do(\tau\ \mathsf{out})$, resp.) on a TID translates in changing the probability of τ to 1 (0, resp.), leaving all other probabilities unchanged. Furthermore, an intervened TID becomes a TID. ∎

The same definitions can be applied to a scalar aggregate query $\mathcal{Q}$, which also becomes a random variable over a PDB. Similarly, and "componentwise", for aggregate queries with group-by.

3.2. The Generalized Causal Effect Score

In (Salimi et. al., 2016), in order to apply the CES(I,U($\frac{1}{2}$)), the original DB was converted into a uniform TID. The CES(I,U($\frac{1}{2}$)) can be generalized by considering an arbitrary PDB D^p.

Definition 4 Let $D^p = \langle \mathcal{W}, p \rangle$ be a PDB , and $\mathcal{Q}$ a Boolean or scalar aggregate query. The *generalized causal effect score* (GCES) of $\mathcal{T} \subseteq D^{en}$ on $\mathcal{Q}$ is:

$$CE(D^p, \mathcal{Q}, \mathcal{T}) \;:=\; \mathbb{E}(\mathcal{Q} \mid do(\mathcal{T}\ \mathsf{in})) - \mathbb{E}(\mathcal{Q} \mid do(\mathcal{T}\ \mathsf{out})).$$ ∎

This definition relies on the probability distribution of D^p. When D^p is a TID associated to a regular relational DB D, we write $CE^I(D, \mathcal{Q}, \mathcal{T})$. CES(I,U($\frac{1}{2}$)) is a particular case where endogenous tuples have a uniform distribution $U(\frac{1}{2})$ (see Remark 1). In this case we use the notation $CE^{UI}(D, \mathcal{Q}, \mathcal{T})$ (omitting the $\frac{1}{2}$).

The role of exogenous tuples is twofold. First, they are excluded from counterfactual interventions; and, secondly, they influence the probabilities used in the expected values by contributing with probability 1. There are no technical difficulties in defining causal-effects for exogenous tuples. However, we leave them outside the scope of explanations for query answers.

Remark 5 **Special cases and notation.** (a) For a Boolean query $\mathcal{Q}$: $CE(D^p, \mathcal{Q}, \mathcal{T}) = P(\mathcal{Q} = 1 \mid do(\mathcal{T}\ \mathsf{in})) - P(\mathcal{Q} = 1 \mid do(\mathcal{T}\ \mathsf{out}))$. (b) $CE(D^p, \mathcal{Q}, \tau)$ denotes the GCES for a single endogenous tuple τ. ∎

In (Salimi et. al., 2016) it was shown that, for a MBQ $\mathcal{Q}$, and $\tau \in D^{en}$, τ is an *actual cause* (Halpern and Pearl, 2005) for $\mathcal{Q}$ in D iff $CE^{UI}(D, \mathcal{Q}, \tau) > 0$. In (Livshits et al., 2021a), it was shown that the CES coincides with the BPI: $CE^{UI}(D, \mathcal{Q}, \tau) = BPI(D, \mathcal{Q}, \tau)$.

Example 1 Let D be a database instance with the relations E and S here below (used in (Salimi et. al., 2016)), with all their tuples endogenous. Let us build a *uniform* TID by defining: $p(\tau) := \frac{1}{2}$ for every tuple τ. Consider the Boolean query $\mathcal{Q}_1$ asking if there exists a path from a to b according to relation E. It can be expressed in Datalog (or a UCQs for a fixed instance), which makes it monotone.

E	A	B
τ_1	a	b
τ_2	a	c
τ_3	c	b
τ_4	a	d
τ_5	d	e
τ_6	e	b

S	A	C
τ_7	a	1
τ_8	a	2
τ_9	b	0
τ_{10}	a	3
τ_{11}	b	1
τ_{12}	b	10

It holds: $CE^{UI}(D, \mathcal{Q}_1, \tau_1) = 0.65625$, $CE^{UI}(D, \mathcal{Q}_1, \tau_2) = CE^{UI}(D, \mathcal{Q}_1, \tau_3) = 0.21875$, and $CE^{UI}(D, \mathcal{Q}_1, \tau_4) = CE^{UI}(D, \mathcal{Q}_1, \tau_5) = CE^{UI}(D, \mathcal{Q}_1, \tau_6) = 0.09375$. As noticed in (Salimi et. al., 2016), these scores significantly differ from the responsibility scores, which are all $\frac{1}{3}$, despite the fact that they make the query true through paths of different lengths.

Consider now the scalar aggregate query $\mathcal{Q}$ defined by: $Ans^{\mathcal{Q}}(sum(y)) \leftarrow S(x, y)$. Its answer is 17. When $Dom(B) \subseteq \mathbb{R}^+$, this is a monotone query. CES(I,U($\frac{1}{2}$)) for a tuple $\tau \in S$ is computed using Definition 4. We need to compute the expected value of the query when intervening the tuple τ. Denote with $\tau[C]$ the restriction of a tuple τ to attribute C; in this case, the numerical value. Consider that, for any $\tau' \in D$ with $\tau \neq \tau'$, the probability of τ' stays the same when intervening τ, i.e. $P(\tau' \mid do(\tau \text{ in})) = P(\tau' \mid do(\tau \text{ out})) = p(\{\tau'\})$, and for τ, it holds $P(\tau \mid do(\tau \text{ in})) = 1$ and $P(\tau \mid do(\tau \text{ out})) = 0$. Now, and since the aggregate query $\mathcal{Q}$ is just adding the value of the attribute C from each tuple, the CES becomes: $CE^{UI}(D, \mathcal{Q}, \tau_7) = \mathbb{E}(\mathcal{Q} \mid do(\tau_7 \text{ in})) - \mathbb{E}(\mathcal{Q} \mid do(\tau_7 \text{ out})) = \tau_7[C] = 1$, a result in line with the intuition: the average expected contribution of tuple τ_7 for a query that is adding up all tuples in the relation would be the attribute value of the tuple itself, $\tau_7[C]$.

Let us now define a PDB D^p (restricted to relation E) by: For $W_1 = \{\tau_1, \tau_3, \tau_4, \tau_6\}$, $p(W_1) := 0.20$; for $W_2 = \{\tau_1, \tau_2, \tau_3\}$, $p(W_2) := 0.25$; for $W_3 = \{\tau_2, \tau_3, \tau_6\}$, $p(W_3) := 0.15$; for $W_4 = \{\tau_2, \tau_6\}$, $p(W_4) := 0.40$; and for any other $W \subseteq E$, $p(W) := 0$. Here, we are starting with a probability distribution over the possible worlds.

According to (1): $P(\tau_1) = 0.2 + 0.25 = 0.45$, $P(\tau_2) = 0.8$, $P(\tau_3) = 0.8$, $P(\tau_4) = 0.2$, $P(\tau_5) = 0$, $P(\tau_6) = 0.75$. If we compute the probabilities of possible worlds using these tuple probabilities assuming independence, we obtain, e.g. for W_1: $P'(W_1) := 0.2 \times (1 - 0.8) \times 0.8 \times 0.2 \times (1 - 0) \times 0.75 = 0.00288$, showing that D^p we started with is not a TID.

We can compute the causal effect of τ_3 without explicitly appealing to the intervened worlds and distributions. It differs from $CE^{UI}(D, \mathcal{Q}_1, \tau_3)$: (see Remark 5(a))

$$CE(D^p, \mathcal{Q}_1, \tau_3) = P(\mathcal{Q}_1 = 1 \mid do(\tau_3 \text{ in})) - P(\mathcal{Q}_1 = 1 \mid do(\tau_3 \text{ out})) = \sum_{W \in \mathcal{W},\, W \cup \{\tau_3\} \models \mathcal{Q}} p(W) - \sum_{W \in \mathcal{W},\, W \smallsetminus \{\tau_3\} \models \mathcal{Q}} p(W) = (p(W_1) + p(W_2) + p(W_3) + p(W_4)) - (p(W_1) + p(W_2)) = p(W_3) + p(W_4) = 0.55.$$

For illustration, let us consider the intervened distributions $P(\cdot \mid do(\tau_3 \text{ in}))$ and $P(\cdot \mid do(\tau_3 \text{ out}))$ for the computation of the same causal effect: $\mathcal{W}^{+\tau_3} = \{W_1', W_2', W_3'\}$, with $W_1' = W_1$, $W_2' = W_2$ and $W_3' = \{\tau_2, \tau_3, \tau_6\}$; and $\mathcal{W}^{-\tau_3} = \{W_1^\star, W_2^\star, W_3^\star\}$ with $W_1^\star = W_1$, $W_2^\star = W_2$ and $W_3^\star = \{\tau_2, \tau_6\}$. Notice that $p(W_1) = p^{+\tau_3}(W_1') = p^{-\tau_3}(W_1^\star)$ and $p(W_2) = p^{+\tau_3}(W_2') = p^{-\tau_3}(W_2^\star)$. Also, $W_3 \cup \{\tau_3\} = W_4 \cup \{\tau_3\} = W_3'$ and $W_3 \smallsetminus \{\tau_3\} = W_4 \smallsetminus \{\tau_3\} = W_3^\star$. Then, $p^{+\tau_3}(W_3') = p^{-\tau_3}(W_3^\star) = p(W_3) + p(W_4) = 0.55$.

$$CE(D^p, \mathcal{Q}_1, \tau_3) = \mathbb{E}(\mathcal{Q}_1 \mid do(\tau_3 \text{ in})) - \mathbb{E}(\mathcal{Q}_1 \mid do(\tau_3 \text{ out})) = \sum_{W' \in \mathcal{W}^{+\tau_3},\, W' \models \mathcal{Q}_1} p^{+\tau_3}(W') - \sum_{W^\star \in \mathcal{W}^{-\tau_3},\, W^\star \models \mathcal{Q}_1} p^{-\tau_3}(W^\star) = (p^{+\tau_3}(W_1') + p^{+\tau_3}(W_2') + p^{+\tau_3}(W_3')) - (p^{-\tau_3}(W_1^\star) + p^{-\tau_3}(W_2^\star)) = p^{+\tau_3}(W_3') = p(W_3) + p(W_4) = 0.55. \quad \blacksquare$$

4. Computational Complexity of CES

In this section we investigate the complexity of computing the CES for an arbitrary TID, i.e. $CE^I(D, \mathcal{Q}, \tau)$, and BCQs (see Remark 5). We obtain a dichotomy result similar to that for query answering on TIDs (Suciu et al., 2011). For its formulation, we need some preliminary notions.

For a BCQ $\mathcal{Q}$, $Atoms(\mathcal{Q})$ denotes the set of all atoms in $\mathcal{Q}$, i.e. its conjuncts. For a variable v in $\mathcal{Q}$, $Atoms(v)$ denotes the set of atoms of $\mathcal{Q}$ where v appears. A BCQ $\mathcal{Q}$ is *hierarchical* if, for any two variables x, y in $\mathcal{Q}$, one of the following holds: (a) $Atoms(x) \subseteq Atoms(y)$, (b) $Atoms(y) \subseteq Atoms(x)$ or (c) $Atoms(x) \cap Atoms(y) = \emptyset$. Otherwise, the query is called *non-hierarchical* (Suciu et al., 2011). The evaluation of a BCQ without self-joins on a TID D: (a) Can be done in polynomial-time when $\mathcal{Q}$ is hierarchical, and (b) Is $\#P$-hard when $\mathcal{Q}$ is non-hierarchical. All these results are in data complexity (Suciu et al., 2011). Deciding if a BCQ is hierarchical or not can be done with an efficient syntactic test.

Definition 6 (a) Let $\mathcal{Q}$ be a BCQ; $\mathcal{G}$ the undirected graph with nodes in $Atoms(\mathcal{Q})$ and edges of the form $\{A_i, A_j\}$, where A_1 and A_2 share at least one variable; and $\mathcal{C}(\mathcal{Q})$ the sets of connected components in $\mathcal{G}$. Accordingly, two different sets of atoms in $\mathcal{C}(\mathcal{Q})$ do not share variables. (b) For an instance D compatible with the schema of $\mathcal{Q}$, the *fresh expansion of D via $\mathcal{Q}$* is the instance $D \cup \mathcal{T}$, where $\mathcal{T}$ contains exactly one new tuple τ^U for each $U \in Atoms(\mathcal{Q})$, with the same predicate of U and with the variables inherited from $\mathcal{Q}$ all replaced simultaneously by new constants not appearing in $Adom(D)$. Now, a *selection function* σ assigns, to each $C \in \mathcal{C}(\mathcal{Q})$, exactly one $\tau^U \in \mathcal{T}$, with $U \in C$. ■

Remark 7 For a TID D^p, and a BCQ $\mathcal{Q}$, it holds: $P(\mathcal{Q}, D^p) = \prod_{C_i \in \mathcal{C}(\mathcal{Q})} P(\bar{\exists}\mathcal{Q}_i, D)$, where $\bar{\exists}\mathcal{Q}_i$ is the existential closure of the conjunction of the atoms in component C_i. For example, for the query $\mathcal{Q}: \exists x \exists y(R_1(x) \wedge R_2(x) \wedge R_3(y))$, $C_1 = \{R_1(x), R_2(x)\}$, $C_2 = \{R_3(y)\}$; $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are $(R_1(x) \wedge R_2(x))$ and $R_3(y)$, resp.; and $P(\mathcal{Q}, D^p) = P(\exists x(R_1(x) \wedge R_2(x)), D^p) \times P(\exists y R_3(y), D^p)$. ■

Proposition 8 Let $\mathcal{Q}$ be a BCQ, $D^p = \langle \mathcal{W}, p \rangle$ a TID associated to relational instance D, and $D' := D \cup \mathcal{T}$ a fresh expansion of D via $\mathcal{Q}$. A TID $(D')^p = \langle \mathcal{W}', p' \rangle$, with $\mathcal{W}' = \mathcal{P}(D')$, the power set of D', and $\mathcal{T} \subseteq (D')^{en}$, can be constructed in constant time, in such a way that, for every selection function σ:

$$P(\mathcal{Q}, D^p) = \prod_{C \in \mathcal{C}(\mathcal{Q})} (1 - CE^I(D, \mathcal{Q}, \sigma(C))). \quad (4)$$

■

Remark 9 (a) Notice that $P(\mathcal{Q}, D^p)$ can be obtained using any selection function. That is, there are different but coincident representations of this probability. (b) The number of factors in the product is $|\mathcal{C}(\mathcal{Q})|$, and depends only on the query. ■

The following example illustrates Definition 6 and Proposition 8.

Example 2 Consider the TID $D^p = \langle \mathcal{W}, p \rangle$ here below, and the BCQ $\mathcal{Q}: \exists x \exists y(R_1(x, y) \wedge R_2(y) \wedge R_3(z))$.

R_1	A	B	P
τ_1	a	a	0.9
τ_2	b	b	0.3
τ_3	c	b	0.8

R_2	A	P
τ_4	b	0.5

R_3	A	P
τ_5	d	0.9
τ_6	e	0.2

TID $(D')^p := \langle \mathcal{W}', p' \rangle$ is built as follows: Introduce fresh constants c_1, c_2, c_3 for variables x, y, z in $\mathcal{Q}$, and create new endogenous tuples $\tau_1^U: R_1(c_1, c_2)$, $\tau_2^U: R_2(c_2)$, and $\tau_3^U: R_3(c_3)$; all with probability of 1. The *fresh expansion* of D via $\mathcal{Q}$ is $D' = D \cup \{\tau_1^U, \tau_2^U, \tau_3^U\}$, with the distribution on D extended to D' to make the latter a TID. Query $\mathcal{Q}$ has two components: $C_1 = \{R_1(x,y), R_2(y)\}$ and $C_2 = \{R_3(z)\}$. With the selection function $\sigma(C_1) := \tau_2^U$ and $\sigma(C_2) := \tau_3^U$, it holds: $P(\mathcal{Q}, D^p) = \prod_{C \in \mathcal{C}(\mathcal{Q})}(1 - CE(D', \mathcal{Q}, \sigma(C))) = (1 - CE(D', \mathcal{Q}, \tau_2^U)) \times (1 - CE(D', \mathcal{Q}, \tau_3^U)) = (1 - 0.57) \times (1 - 0.02) = 0.3965$. With the selection function $\sigma'(C_1) := \tau_1^U$ and $\sigma'(C_2) := \tau_3^U$, the result is the same since $CE^I(D', \mathcal{Q}, \tau_1^U) = CE^I(D', \mathcal{Q}, \tau_2^U) = 0.57$. ■

Notice that the extra tuples in Proposition 8, although endogenous, behave as exogenous tuples in that they have probability 1 (see Remark 1).

Theorem 10 Let $\mathcal{Q}$ be a BCQ without self-joins. (a) If $\mathcal{Q}$ is hierarchical, then, for every TID $D^p = \langle \mathcal{W}, p \rangle$ and endogenous tuple τ, computing $CE^I(D, \mathcal{Q}, \tau)$ is in $PTIME$. (b) If $\mathcal{Q}$ is non-hierarchical, then computing $CE^I(D, \mathcal{Q}, \tau)$ on TIDs is $\#P$-hard.[3] ■

Remark 11 (a) Theorem 10 applies in particular to the computation of the CES on a TID with an arbitrary uniform distribution on endogenous tuples (and probability 1 for exogenous tuples), and also to CES(I,U($\frac{1}{2}$)), i.e. when endogenous tuples have probability $\frac{1}{2}$.[4] (b) Since CES(I,U($\frac{1}{2}$)) coincides with the BPI on DBs, we reobtain here, in a different way, the dichotomy result for BPI and BCQs (Livshits et al., 2021a). (c) With some caveats, the results in this Section can be extended to UBCQs, but considering the notions of *safe* and *unsafe* queries (Dalvi and Suciu, 2012) (see also (Suciu, 2020; Deutch et al., 2022; Bertossi et al., 2023b)). (c) For the hardness part of Theorem 10 we used the existence of pseudo-exogenous tuples. A natural question is whether it is possible to obtain the same result without appealing to them, but only to "properly" endogenous tuples. ■

As a consequence of the analysis in this section, it is clear that computing the Generalized Causal-Effect Score is also intractable.

5. Characterizing Properties of the GCES

In this section we provide an axiomatic characterization of the Generalized Causal-Effect Score (GCES), that is, we show that this score is the only function satisfying a given set of properties. We consider the most general version of GCES, i.e. for an arbitrary probability distribution over all possible worlds of an instance. However, we restrict ourselves to *monotone Boolean queries* (MBQs), which includes BCQ and UBCQs. The axioms depend on dealing with MBQs.

In (Dubey and Shapley, 1979), an axiomatic characterization was given for the Banzhaf Power Index (BPI). It is a general axiomatization in the context of game-theory. As already mentioned, in (Livshits et al., 2021a) it was shown that, in the context of query answering in DBs, CES(I,U($\frac{1}{2}$)), i.e. the CES with independent tuples and uniform distribution with parameter $\frac{1}{2}$ coincides with BPI; the latter applied with the endogenous tuples as players in a coalition game. From this connection, it follows that there is an axiomatic characterization for this particular case of the CES.

In the following, we will show that only some of the characterizing properties, or axioms, for the BPI still hold for the GCES. We will also propose extra properties for the GCES, and will establish,

3. Meaning that computing $CE^I(D, \mathcal{Q}, \tau)$ over inputs of the form $\langle D^p, \tau \rangle$, with D^p a TID for D and $\tau \in D^{en}$, is a $\#P$-hard problem (Suciu, 2020).
4. For discussions and results on the complexity of probabilistic query answering for these different cases, see (Amarilli and Kimelfeld, 2022; Suciu, 2020; Kening and Suciu, 2021)).

in Theorem 13, that the resulting set of axioms turns out to be *categorical*, i.e. fully characterizing for the GCES.[5] We will start introducing some useful notions and notation that will allow us to formulate the characterizing properties of the GCES as appropriate for query answering in PDBs.

Definition 12 Let $\mathcal{Q}$ be a monotone Boolean query, $D^p = \langle \mathcal{W}, p\rangle$ a PDB, and $W \in \mathcal{W}$.

(a) For $W \subsetneqq D^{en}$ its *power* is:

$$\begin{aligned} Power(D^p, \mathcal{Q}, W) &:= \sum_{\tau \in D^{en}} \Delta(\mathcal{Q}, W, \tau), \text{ with} \\ \Delta(\mathcal{Q}, W, \tau) &:= \mathcal{Q}[W \cup D^{ex} \cup \{\tau\}] - \mathcal{Q}[W \cup D^{ex}]. \end{aligned}$$

(b) For a tuple $\tau \in D^{en}$, its *power* is:

$$Power(D^p, \mathcal{Q}, \tau) := \sum_{W \,\subseteq\, (D^{en} \smallsetminus \{\tau\})} \Delta(\mathcal{Q}, W, \tau),$$

with $\Delta(\mathcal{Q}, W, \tau)$ as in (a).

(c) The *weighted power* of $\tau \in D^{en}$ is:

$$Power^w(D^p, \mathcal{Q}, \tau) := \sum_{W \subseteq (D^{en} \smallsetminus \{\tau\})} \Delta(\mathcal{Q}, W, \tau) \times p(W \cup D^{ex}).$$

(d) The *total power* of the pair $(D^p, \mathcal{Q})$, is:

$$Power(D, \mathcal{Q}) := \sum_{W \,\subsetneqq\, D^{en}} Power(D^p, \mathcal{Q}, W). \quad \blacksquare$$

The *power* of a tuple τ represents the number of subsets W for which, adding τ to them, produces a change in the value of $\mathcal{Q}$. Similarly, the power of a subset W represents the number of tuples that, when added to W, change the query answer. Notice that, equivalently, $Power(D^p, \mathcal{Q}) = \sum_{\tau \in D^{en}} Power(D^p, \mathcal{Q}, \tau)$.

Example 3 Consider the instance (with tuple-identifiers) $D^p = \{\tau_1 : R(a,b),\ \tau_2 : R(b,c),\ \tau_3 : S(a,a),\ \tau_4 : S(a,d)\}$ with exogenous tuples in $D^{ex} = \{\tau_1\}$. For the BCQ $\mathcal{Q} : \exists x \exists y \exists z (R(x,y) \wedge S(x,z))$ and $W_1 = \{\tau_2\}$: $Power(D^p, \mathcal{Q}, W_1) = \sum_{\tau \in \{\tau_2,\tau_3,\tau_4\}} \Delta(\mathcal{Q}, W_1, \tau) = 0 + 1 + 1 = 2$.

The *power* of τ_2 (as a tuple, not as the set $\{\tau_2\}$) is: $Power(D^p, \mathcal{Q}, \tau_2) = \sum_{W \subseteq (D^{en} \smallsetminus \{\tau_2\})} \Delta(\mathcal{Q}, W, \tau_2) = \Delta(\mathcal{Q}, \emptyset, \tau_1) + \Delta(\mathcal{Q}, \{\tau_3\}, \tau_2) + \Delta(\mathcal{Q}, \{\tau_4\}, \tau_2) + \Delta(\mathcal{Q}, \{\tau_3, \tau_4\}, \tau_2) = 0 + 0 + 0 + 0 = 0$.

With $W_2 = \{\tau_3\}, W_3 = \{\tau_4\}, W_4 = \{\tau_2, \tau_3\}, W_5 = \{\tau_2, \tau_4\}, W_6 = \{\tau_3, \tau_4\}, W_7 = \{\tau_2, \tau_3, \tau_4\}$, $W_8 = \emptyset$, the *power* of W_i is $Power(D^p, \mathcal{Q}, W_i) = 2$ if $i = 1, 8$, but 0, otherwise. Then, the *total power* of $(D^p, \mathcal{Q})$ is: $Power(D^p, \mathcal{Q}) = \sum_{W \subseteq D^{en}} Power(D^p, \mathcal{Q}, W) = Power(D^p, \mathcal{Q}, W_1) + \cdots + Power(D^p, \mathcal{Q}, W_8) = 2 + 0 + \cdots + 0 + 2 = 4$.

As mentioned above, the *total power* can also be obtained through the *power* of the tuples. The *power* of τ_3 and τ_4 are $Power(D^p, \mathcal{Q}, \tau_3) = Power(D^p, \mathcal{Q}, \tau_4) = 2$. Then, $Power(D^p, \mathcal{Q}) = \sum_{\tau \in \{\tau_2,\tau_3,\tau_4\}} Power(D^p, \mathcal{Q}, \tau) = 0 + 2 + 2 = 4$. $\blacksquare$

5. In Mathematical Logic, a *categorical theory* has a single model (modulo isomorphism).

In the following, we will assume that we have a fixed PDB $D^p = D^{en} \cup D^{ex}$, with $D^{en} = \{\tau_1, \ldots, \tau_N\}$, and a probability distribution p on $\mathcal{W}$, its set of possible worlds. We restrict ourselves to the class MBQ *of monotone Boolean queries for the schema of D^p.*

Now, depending on a query $\mathcal{Q}$, each tuple in D^{en} will take a real number as a score that represent its relevance for the answer to $\mathcal{Q}$ from D^p. This is represented as a "vectorial score-function", ψ, from MBQ to $\mathbb{R}^N$: $\psi(\mathcal{Q}) = \langle \psi_{\tau_1}(\mathcal{Q}), \ldots, \psi_{\tau_N}(\mathcal{Q})\rangle$.

Based on (Dubey and Shapley, 1979), we first analyze the following potential properties of ψ:

- DUM: (for "dummy") If $\tau \in D^{en}$ is a *dummy tuple*, meaning that $Power(D^p, \mathcal{Q}, \tau) = 0$, then $\psi_\tau(\mathcal{Q}) = 0$.[6]
- EFF: (for "efficiency") $\sum_{\tau \in D^{en}} \psi_\tau(\mathcal{Q}) = Power(D^p, \mathcal{Q})/(2^{N-1})$.
- SYM: (for "symmetry") If $\mathcal{Q}[W \cup D^{ex} \cup \{\tau\}] = \mathcal{Q}[W \cup D^{ex} \cup \{\tau'\}]$ for every $W \subseteq D^{en} \smallsetminus \{\tau, \tau'\}$, then $\psi_\tau(\mathcal{Q}) = \psi_{\tau'}(\mathcal{Q})$.
- LIN: (for "linearity") For MBQs $\mathcal{Q}$ and $\mathcal{Q}'$, $\psi(\mathcal{Q} \vee \mathcal{Q}') + \psi(\mathcal{Q} \wedge \mathcal{Q}') = \psi(\mathcal{Q}) + \psi(\mathcal{Q}')$.

According to (Dubey and Shapley, 1979), as applied to our setting, there is a unique function $\psi\colon \mathsf{MBQ} \to \mathbb{R}^N$ that satisfies the properties DUM, EFF, SYM and LIN, and it corresponds to the BPI as defined in (3). For the formulation of the properties above we do not using the possibly arbitrary distribution of the PDB D^p. However, the BPI on DBs coincides, in its turn, with the CES(I,U($\frac{1}{2}$)) case of the Causal-Effect Score in DBs (Livshits et al., 2021a). This leaves open the question about properties that characterize GCES, the general case of the score, that we address in the rest of this section. We would expect the distribution p of D^p to play a role in them.

It is worth mentioning that the Shapley value also satisfies DUM, SYM and LIN, but not EFF. However, it does satisfy slightly modified version of EFF, where the sum of the value for all tuples is equal to $\mathcal{Q}[D]$ (Shapley, 1953; Aumann and Shapley, 2015).

Example 4 below shows that the GCES does not (always) satisfy SYM and EFF. It also illustrates that it satisfies DUM and LIN. Actually, Theorem 13 will show that the GCES always satisfies DUM, LIN, and modified versions of SYM and EFF.

Example 4 (ex. 3 cont.) Consider the PDBs $D^p = \langle \mathcal{W}, p\rangle$ and $D^{p'} = \langle \mathcal{W}, p'\rangle$ associated to the same instance of Example 3. Here, $N = 3$, the number of endogenous tuples.

The probability distributions are defined as follows: $p(W_i \cup \{\tau_1\}) = 1/8$ for $i = 1, \ldots, 8$; and $p'(W_j \cup \{\tau_1\}) = 1/12$ for $j = 1, 2, 3, 4$, and $p'(W_k \cup \{\tau_1\}) = 1/6$ for $k = 5, 6, 7, 8$; and $p(W_i) = p'(W_i) = 0$, for $i = 1, \ldots, 8$, because those W_i do not contain D^{ex}.

Notice that p creates a TID with distribution $U(\frac{1}{2})$, because: (a) for each $\tau \in D^{en}$, $P(\tau) = \sum_{W, \tau \in W} p(W) = \frac{1}{2}$; and (b) for $W \subseteq D^{en}$, $p(W \cup D^{ex}) = \prod_{\tau \in W} P(\tau) \times \prod_{\tau \in (D^{en} \smallsetminus W)} (1 - P(\tau))$ (see Section 2). However, p' is neither independent nor uniform.

First, we check DUM. Notice that τ_2 is the only *dummy tuple*, i.e. $Power(D^p, \mathcal{Q}, \tau_2) = 0$, as in Example 3. It holds $CE(D^p, \mathcal{Q}, \tau_2) = CE(D^{p'}, \mathcal{Q}, \tau_2) = 0$, and therefore, both distributions satisfy this property.

Next, we check EFF by computing the GCES of τ_3 and τ_4, for both PDBs, obtaining that p satisfies EFF, but p' does not:

6. For monotone queries, this condition is equivalent to $\mathcal{Q}[W \cup D^{ex}] = \mathcal{Q}[W \cup D^{ex} \cup \{\tau\}]$ for all $W \subseteq D^{en} \smallsetminus \{\tau\}$, a more common formulation.

$$\sum_{\tau \in \{\tau_2,\tau_3,\tau_4\}} CE(D^p, \mathcal{Q}, \tau) \ = \ 0 \ + \ 1/2 \ + \ 1/2 \ = \ \frac{Power(D^p, \mathcal{Q})}{2^{N-1}},$$

$$\sum_{\tau \in \{\tau_2,\tau_3,\tau_4\}} CE(D^{p'}, \mathcal{Q}, \tau) \ = \ 0 \ + \ 5/12 \ + \ 1/2 \ \neq \ \frac{Power(D^{p'}, \mathcal{Q})}{2^{N-1}}.$$

Let us check SYM. Tuples τ_3 and τ_4 should have the same GCES, because, for every $W \subseteq \{\tau_2\}$, $\mathcal{Q}[W \cup D^{ex} \cup \{\tau_3\}] = \mathcal{Q}[W \cup D^{ex} \cup \{\tau_4\}]$. This holds for p, but not for p'.

Lastly, we check LIN. Consider the query $\mathcal{Q}' : \exists w S(w,w)$. We only show that GCES satisfies LIN for τ_3 (it is similar for the other tuples). Computing its GCES with queries Q', $Q \wedge Q'$ and $Q \vee Q'$, we obtain: $CE(D^p, \mathcal{Q}', \tau_3) = CE(D^p, Q \wedge Q', \tau_3) = 1$, $CE(D^p, Q \vee Q', \tau_3) = 1/2$; and $CE(D^{p'}, \mathcal{Q}', \tau_3) = CE(D^{p'}, \mathcal{Q}', \tau_3) = 1$, $CE(D^{p'}, \mathcal{Q}', \tau_3) = 5/12$. It is easy to verify that LIN is satisfied with both distributions. ■

This example motivates modifying EFF and SYM, proposing their generalized versions:

G-EFF: (for "generalized-efficiency")

$$\sum_{\tau \in D^{en}} \psi_\tau(\mathcal{Q}) = \sum_{W \subsetneq D^{en}} \sum_{\tau \in (D^{en} \smallsetminus W)} \Delta(\mathcal{Q}, W, \tau) \times (p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})) .$$ [7]

G-SYM: (for "generalized-symmetry") If $\Delta(\mathcal{Q}, W, \tau) = \Delta(\mathcal{Q}, W, \tau') = 0$ for every $W \subseteq D^{en} \smallsetminus \{\tau, \tau'\}$, then $\psi_\tau(\mathcal{Q}) - Power^w(D^p, \mathcal{Q}, \tau) = \psi_{\tau'}(\mathcal{Q}) - Power^w(D^p, \mathcal{Q}, \tau')$.[8]

G-EFF is similar to EFF, but involves an arbitrary distribution p. If the latter is independent and uniform $U(\frac{1}{2})$, then $(p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\}))$ becomes $1/2^{N-1}$ ($N = |D^{en}|$), for every $\tau \in D^{en}$ and $W \subsetneq D^{en}$.

The equalities in the consequents of G-SYM and SYM are similar, but, for the former, we remove the tuples' powers. With an independent and $U(\frac{1}{2})$ distribution, if two tuples satisfy the hypothesis, the term $Power(D^p, \mathcal{Q}, \tau)$ is the same for both, reobtaining SYM.

Theorem 13 Let $D^p = \langle \mathcal{W}, p \rangle$ be a PDB. There is a unique score function ψ from MBQ, the class of monotone Boolean queries, to $\mathbb{R}^N$, with $N = |D^{en}|$, that satisfies the properties DUM, G-EFF, G-SYM and LIN. Moreover, this function coincides with the GCES. ■

This theorem does not make any assumption on the distribution p. In particular, neither tuple-independence nor uniformity are required. It is worth noticing that the property of monotonicity of the queries is used in the proof of the Theorem, particularly for uniqueness, and for GCES' satisfaction of G-SYM.

7. The sums on the RHS can be given in terms of $W \subseteq D^{en}$ and $\tau \in D^{en}$, because $\Delta(Q, W, \tau) = 0$ for every $\tau \in W$.
8. One could argue that the property should also hold under the "symmetric" condition "$\Delta(\mathcal{Q}, W, \tau) = \Delta(\mathcal{Q}, W, \tau') = 1$ for all possible $W \subseteq D^{en} \smallsetminus \{\tau, \tau'\}$". However, if that were the case, we would have $Q[W] = 0$ and $Q[W \cup \{\tau\}] = Q[W \cup \{\tau'\}] = 1$. Since $W \cup \{\tau\}$ and $W \cup \{\tau'\}$ are distinct sets, possibly with different probabilities, it is not sensible to impose this kind of symmetry.

6. Discussion and Conclusions

In between, and in parallel work, we have managed to investigate in detail the conditions under which CES(I,U($\frac{1}{2}$)) (i.e. GCES with independence and $U(\frac{1}{2})$ distribution), Shapley and RESP are aligned, i.e. they produce the same rankings (a property also called "ordinal equivalence" (Freixas, 2010)). For some of our results see (Azua and Bertossi, 2024). Experimentation with cases of queries whose scores are not aligned is left for ongoing and future work, so as other open research directions we mention next.

Beyond Monotone Queries: We have restricted ourselves to monotone queries, and this assumption play an important role in the characterization of the GCES. It is left open to find a characterization for interesting classes of non-monotone queries, e.g. conjunctive with negative literals.

Beyond Boolean Queries: For the axiomatic characterization of the GCES, we have restricted ourselves to Boolean queries. It would be interesting to have such an axiomatization for more general monotone numerical queries, e.g. some aggregate queries.

GES Monotonicity: It would be interesting to investigate more deeply the monotonicity properties (or the lack thereof) of the GCES. For example, under what changes in the database, or in the tuple probabilities, the GCES of tuples change accordingly (or the other way around).

Score Computation: Due to the intrinsic high complexity of computing the GCES, it is worth exploring efficient approximate algorithms, possibly for some interesting cases of queries.

Aggregate Queries: In this work, we considered mostly conjunctive queries. A natural extension is a deeper investigation of queries with aggregations on CQs. We made early on the case for the convenience of the GCES for this kind of queries.

Semantic Constraints: RESP has been formalized and investigated in the presence of integrity constraints, and its behavior changes (Bertossi and Salimi, 2017b). This is something to investigate for the GCES, in its basic and generalized versions. Particularly interesting becomes dealing with constraints on probabilistic DBs (Suciu, 2020).

GES Robustness: It would be interesting to analyze the *robustness* of the Generalized CES, under small variations of parameters, such as the distribution of the probabilistic DB.

Attribute-Level GCES: We defined the GCES at the tuple level. It would be interesting to extend its definition and investigation in order to quantify the causal effect of an attribute value in a tuple. This case is challenging in that it is not only about making an attribute value true or false anymore. Making the latter false may lead to consider multiple alternative values, as has been done for RESP (Bertossi et al., 2020; Bertossi, 2023a).

Acknowledgments

Felipe Azua has been supported by the Millennium Institute for Foundational Research on Data (IMFD, Chile). Part of this work was done while Felipe Azua was visiting LIMOS (U. Clermont-Ferrand, France). Leopoldo Bertossi has been supported by NSERC- DG 2023-04650, the IMFD, and the "National Center for AI Research" (CENIA, Chile, Basal ANID FB210017). Excellent and useful comments by one of the anonymous reviewers are extremely appreciated.

References

Abramovich, O., Deutch, D., Frost, N., Kara, A. and Olteanu, D. Banzhaf Values for Facts in Query Answering. In *Proc. SIGMOD*, 2024.

Amarilli, A. and Kimelfeld, B. Uniform Reliability of Self-Join-Free Conjunctive Queries. *Log. Methods Comput. Sci.*, 18(4), 2022.

Arad, D., Deutch, D. and Frost, N. LearnShapley: Learning to Predict Rankings of Facts Contribution Based on Query Logs. In *Proc. CIKM*, pages 4788-4792, 2022.

Aumann, R. and Shapley, L. Values of Non-Atomic Games. In *Appendix A*. Princeton University Press, 2015.

Azua, F. and Bertossi, L. The Generalized Causal-Effect Score in Data Management (short paper). In. *Proc. Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI (Guide-AI'24)*, collocated with SIGMOD'24, pages 32-35, 2024.

Banzhaf, J. Weighted Voting Doesn't Work: A Mathematical Analysis. *Rutgers L. Rev.*, 19(31), 1964.

Bertossi, L. Attribution-Scores in Data Management and Explainable Machine Learning. In *Proc. ADBIS*, pages 16–33. Springer LNCS 13985, 2023.

Bertossi, L and Salimi, B. From Causes for Database Queries to Repairs and Model-Based Diagnosis and Back. *Theory Comput. Syst.*, 61(1):191–232, 2017a.

Bertossi, L. and Salimi, B. Causes for Query Answers from Databases: Datalog Abduction, View-Updates, and Integrity Constraints. *Int. J. Approx. Reasoning*, 90:226–252, 2017b.

Bertossi, L., Li, J., Schleich, M., Suciu, D. and Vagena, Z. Causality-Based Explanation of Classification Outcomes. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, pages 1–10, 2020.

Bertossi, L., Livshits, E., Kimelfeld, B. and Monet, M. The Shapley Value in Database Management. *ACM Sigmod Record*, 52(3):6–17, 2023.

Bienvenu, M., Figueira, D. and Lafourcade, P. When is Shapley Value Computation a Matter of Counting?. Proc. ACM Manag. Data, 2024, 2(2):105.

Chockler, H. and Halpern, J. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Intell. Res.*, 22:93–115, 2004.

Dalvi, N. and Suciu, D. The Dichotomy of Probabilistic Inference for Unions of Conjunctive Queries. *J. ACM*, 59(6):30:1–30:87, 2012.

Davidson, S., Deutch, D., Frost, N., Kimelfeld, B., Koren, O. and Monet, M. ShapGraph: An Holistic View of Explanations through Provenance Graphs and Shapley Values. In *Proc. SIGMOD*, 2022.

Deutch, D., Frost, N., Kimelfeld, B. and Monet, M. Computing the Shapley Value of Facts in Query Answering. In *Proc. SIGMOD*, pages 1570–1583, 2022.

Dubey, P. and Shapley, L. Mathematical Properties of the Banzhaf Power Index. *Math. Oper. Res.*, 4(2):99–131, 1979.

Freixas, J. On Ordinal Equivalence of the Shapley and Banzhaf Values for Cooperative Games. *Int. J. Game Theory*, 2010, 39:513–527.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, 2007.

Halpern, J. A Modification of the Halpern-Pearl Definition of Causality. In *Proc. IJCAI*, pages 3022–3033, 2015.

Halpern, J. *Actual Causality*. MIT Press, 2016.

Halpern, J. Y. and Pearl, J. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986.

Kanagal, B. and Deshpande, A. Lineage Processing over Correlated Probabilistic Databases. In *Proc. SIGMOD*, pages 675–686, 2010.

Karmakar, P., Monet, M., Senellart, P. and Bressan, S. Expected Shapley-Like Scores of Boolean functions: Complexity and Applications to Probabilistic Databases. Proc. ACM Manag. Data, 2024, 2(2):92.

Kara, A., Olteanu, D. and Suciu, D. From Shapley Value to Model Counting and Back. Proc. ACM Manag. Data, 2024, 2(2):79.

Kenig, B. and Suciu, D. A Dichotomy for the Generalized Model Counting Problem for Unions of Conjunctive Queries. In *Proc. PODS*, pages 312-324, 2021.

Livshits, E., Bertossi, L., Kimelfeld, B. and Sebag, M. The Shapley Value of Tuples in Query Answering. *Log. Methods Comput. Sci.*, 17(3), 2021a.

Livshits, E., Bertossi, L., Kimelfeld, B. and Sebag, M. Query Games in Databases. *ACM Sigmod Record*, 50(1):78–85, 2021b.

Makhija, N. and Gatterbauer, W. A Unified Approach for Resilience and Causal Responsibility with Integer Linear Programming and LP Relaxations. In *Proc. SIGMOD*, 2023.

Meliou, A., Gatterbauer, W., Moore, K. F. and Suciu, D. The Complexity of Causality and Responsibility for Query Answers and Non-Answers. *Proc. VLDB Endow.*, 4(1):34–45, 2010a.

Meliou, A., Gatterbauer, W., Halpern, J., Koch, Ch., Moore, K. E. and Suciu, D. Causality in Databases. *IEEE Data Eng. Bull.*, 33(3):59–67, 2010b.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

Ratner, A., Bach, S. T., Ehrenberg, H., Fries, J., Wu, S. and Ré, Ch. Snorkel: Rapid Training Data Creation with Weak Supervision. *The VLDB Journal*, 2020, 29:709–730.

Roth, A. E, editor. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

Roy, S. and Salimi, B. Causal Inference in Data Analysis with Applications to Fairness and Explanations. In *Reasoning Web. Causality, Explanations and Declarative Knowledge*, pages 105–131. Springer LNCS 13759, 2023.

Rubin, D. B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Salimi, B., Bertossi, L., Suciu, D and Van den Broeck, G. Quantifying Causal Effects on Query Answering in Databases. In *Proc. TaPP*. USENIX Association, 2016.

Sen, P. and Deshpande, A. Representing and Querying Correlated Tuples in Probabilistic Databases. In *Proc. ICDE*, pages 596–605, 2007.

Sen, P., Deshpande, A. and Getoor, L. PRDB: Managing and Exploiting Rich Correlations in Probabilistic Databases. *VLDB Journal*, 18:1065–1090, 2009.

Shapley, L. A Value for an n-Person Game. In *Contributions to the Theory of Games II*, page 307–331. Princeton Univ. Press, 1953.

Suciu, D. Probabilistic Databases for All. In *Proc. PODS*, pages 19–31, 2020.

Suciu, D., Olteanu, D., Ré, Ch. and Koch, Ch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

Appendix: Proofs

Appendix A. Proofs of Section 4

Proof of Proposition 8: Let $D^p = \langle \mathcal{W}, p\rangle$ be a TID associated to relational instance with D. Now, consider a *fresh expansion* of D over $\mathcal{Q}$, $D' = D \cup \mathcal{T}$. Each new tuple $\tau^U \in \mathcal{T}$ will have a probability of 1 and will be independent from every other tuple. The tuples in $D \cap D'$ retain there original (independent) probabilities. Now, consider a new tuple $\tau^U \in \mathcal{T}$ created from an atom of the component $C_k \in \mathcal{C}(\mathcal{Q})$, we can compute its CES using the new probability distribution p': $CE^I(D', \mathcal{Q}, \tau^U) = \mathbb{E}(\mathcal{Q} \mid do(\tau^U \ \mathsf{in})) - \mathbb{E}(\mathcal{Q} \mid do(\tau^U \ \mathsf{out})) = P(\mathcal{Q} \mid do(\tau^U \ \mathsf{in})) - P(\mathcal{Q} \mid do(\tau^U \ \mathsf{out})) = 1 - P(\mathcal{Q} \mid do(\tau^U \ \mathsf{out}))$, the last probability corresponds to interventions on p'.

It holds: (a) $P(\mathcal{Q}, D') = 1$ since $(D' \smallsetminus D) \models \mathcal{Q}$, that is, the set of new tuples τ^U is sufficient to make the query true, and each new tuple has a probability of 1. This probability will not change by the intervention $do(\tau^U \ \mathsf{in})$. (b) By Remark 7, we can compute $P(\mathcal{Q} \mid do(\tau^U \ \mathsf{out}))$ by multiplying the probability of the sub-query of each component of $\mathcal{Q}$, and since each component, will have a probability of 1 (because each tuple in $\mathcal{T}$ has a probability of 1), $P(\mathcal{Q} \mid do(\tau^U \ \mathsf{out})) = P(\mathcal{Q}_k, D)$, with $\mathcal{Q}_k$ the sub-query of component C_k. Note that, if $\tau^U, \tau^{U'} \in \mathcal{T}$ were created from atoms of the same component, say C_k, then $CE^I(D', \mathcal{Q}, \tau^U) = CE^I(D', \mathcal{Q}, \tau^{U'})$. Consider a selection function σ as in Definition 6. Then, by Remark 7, we can rewrite the probability of the sub-query of each component as $P(\mathcal{Q}_i, D) = 1 - CE^I(D, \mathcal{Q}, \sigma(C_i))$. It follows that $P(\mathcal{Q}, D^p) = \prod_{C \in \mathcal{C}(\mathcal{Q})}(1 - CE^I(D, \mathcal{Q}, \sigma(C)))$. Since two tuples from atoms of the same component will have the CES, then any selection function σ would produce the same result. ■

Proof of Theorem 10: Assume $\mathcal{Q}$ is hierarchical. Its causal effect is given in Remark 5(a). Since D^p is a TID, the two probabilities there are also for TIDs (see Remark 3(b)). As a consequence, they can be computed in polynomial-time.

Assume $\mathcal{Q}$ is non-hierarchical. We appeal to Proposition 8. Let D^p be a TID that makes computing $P(\mathcal{Q}, D^p)$ a $\#P$-hard case. Now, for D^p, consider the TID $(D')^p$ as in Proposition 8, for which there are endogenous tuples satisfying (4), which shows that there are tuples for which computing CE^I is as hard as computing $P(\mathcal{Q}, D^p)$. ■

Appendix B. Proof of Theorem 13

In order to prove Theorem 13, we need some notions and a technical result. First, we need to introduce a particular query and its notation. For a fixed instance D, and $S \subseteq D$, $\mathcal{Q}_S$ denotes the following monotone Boolean query:[9] For $W \in \mathcal{W}$,

$$\mathcal{Q}_S[W] := \begin{cases} 1 & \text{, if } S \subseteq W \\ 0 & \text{, otherwise.} \end{cases} \tag{5}$$

Example 5 (ex. 3 cont.) Query $\mathcal{Q}_{W_3}$ is defined by $\mathcal{Q}_{W_3}[W] := 1$ if $W_3 \subseteq W$; and 0 otherwise. It can be expressed as a the conjunction of the tuples (atoms) in W_3: $\mathcal{Q}_{W_3}$: $(R(a,b) \wedge S(b))$. ■

9. Notice that, since S is fixed and finite, it can be expressed in the FO language of the schema. It can be written as the conjunction of the tuples, as ground atoms, of the set S.

Definition 14 Consider a MBQ $\mathcal{Q}$ and an instance D. A $W \subseteq D$ is a *minimal satisfiable set* if $\mathcal{Q}[W] = 1$, and $\mathcal{Q}[W \smallsetminus \{\tau\}] = 0$, for every $\tau \in W$. $\mathit{MSS}(D, \mathcal{Q})$ denotes the set of all minimal satisfiability sets for D and $\mathcal{Q}$. ∎

Lemma 15 Let D be an instance, $\mathcal{Q}$ be a MBQ, and $\mathit{MSS}(\mathcal{Q}) = \{S_1, \dots, S_m\}$. $\mathcal{Q}$ can be expressed as follows: For $W \in \mathcal{W}(D)$, $\mathcal{Q}[W] = (\mathcal{Q}_{S_1} \vee \cdots \vee \mathcal{Q}_{S_m})[W]$, with the queries $\mathcal{Q}_{S_i}$ defined as in (5).

Proof of Lemma 15: Consider a MBQ $\mathcal{Q}$ and an instance D. The statement of the lemma is equivalent to: $\mathcal{Q}[W] = 1$ iff there exists $S_i \in \mathit{MSS}(D, \mathcal{Q})$, such that $S_i \subseteq W$. Now we prove this claim (both directions) by contradiction.

($\Rightarrow$) Let $W^\star$ be a set such $\mathcal{Q}[W^\star] = 1$, but, for every $S_i \in \mathit{MSS}(\mathcal{Q})$, $S_i \not\subseteq W^\star$. Two cases arise: (a) If $\mathcal{Q}[W^\star \smallsetminus \{\tau\}] = 0$ for every $\tau \in W^\star$, then $W^\star \in \mathit{MSS}(D, \mathcal{Q})$; and (b) There is a tuple τ, such that $\mathcal{Q}[W^\star \smallsetminus \{\tau\}] = 1$. Case (a) is clearly a contradiction. In case (b), we can remove tuples from $W^\star$ until case (a) occurs, eventually leading to a contradiction. Therefore, such a $W^\star$ cannot exist.

($\Leftarrow$) Assume that there exists $W \subseteq D$ such $\mathcal{Q}[W] = 0$, and $S \subseteq W$ with $S \in \mathit{MSS}(D, \mathcal{Q})$. Since $S \in \mathit{MSS}(D, \mathcal{Q})$, $\mathcal{Q}[S] = 1$, which cannot happen, because $\mathcal{Q}$ is monotone. So, such a W cannot exist. ∎

Proof of Theorem 13:[10] First, we prove the uniqueness of the function ψ, and then, that the GCES satisfies all properties. Consider a monotone Boolean query $\mathcal{Q}$, and its minimal satisfiable sets, say $\mathit{MSS}(D, \mathcal{Q}) = \{S_1, \dots, S_m\}$. By Lemma 15, $\mathcal{Q}$ can be decomposed in queries of the form (5). That is, for every $W \subseteq D$: $\mathcal{Q}[W] = (\mathcal{Q}_{S_1} \vee \cdots \vee \mathcal{Q}_{S_m})[W]$.

Now, consider any of the individual queries in the disjunction, say $\mathcal{Q}_{S_i}$. By property G-EFF, the following holds:

$$\sum_{\tau \in D^{en}} \psi_\tau(\mathcal{Q}) = \sum_{W \subsetneqq D^{en}} \sum_{\tau \in (D^{en} \smallsetminus W)} \Delta(\mathcal{Q}, W, \tau) \times \left(p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})\right).$$

Three observations: (a) We can express $\sum_{\tau \in D^{en}} \psi_\tau(\mathcal{Q}_{S_i})$ as:

$$\sum_{\tau \in D^{en}} \psi_\tau(\mathcal{Q}_{S_i}) = \sum_{\tau \in D^{en}} \sum_{W \subseteq (D^{en} \smallsetminus \{\tau\})} \Delta(\mathcal{Q}_{S_i}, W, \tau) \times p(W \cup D^{ex}) \; + \\ \sum_{\tau \in D^{en}} \sum_{W \subseteq (D^{en} \smallsetminus \{\tau\})} \Delta(\mathcal{Q}_{S_i}, W, \tau) \times p(W \cup D^{ex} \cup \{\tau\}).$$

(b)

$$\sum_{\tau \in D^{en}} \mathit{Power}^w(D^p, \mathcal{Q}_{S_i}, \tau) = \sum_{\tau \in D^{en}} \sum_{W \subseteq (D^{en} \smallsetminus \{\tau\})} \Delta(\mathcal{Q}_{S_i}, W, \tau) \times p(W \cup D^{ex}).$$

(c) For any two tuples $\tau, \tau' \in (S_i \cap D^{en})$, $\Delta(\mathcal{Q}_{S_i}, W, \tau) = \Delta(\mathcal{Q}_{S_i}, W, \tau') = 0$, whenever $W \subseteq (D^{en} \smallsetminus \{\tau, \tau'\})$. Therefore, by G-SYM, it holds:

$$\psi_\tau(\mathcal{Q}_{S_i}) - \mathit{Power}^w(D^p, \mathcal{Q}_{S_i}, \tau) \;=\; \psi_{\tau'}(\mathcal{Q}_{S_i}) - \mathit{Power}^w(D^p, \mathcal{Q}_{S_i}, \tau').$$

10. This proof is inspired by a proof in (Dubey and Shapley, 1979) for the BPI.

From (a) and (b), we obtain: $\sum_{\tau \in D^{en}}(\psi_\tau(\mathcal{Q}_{S_i}) - Power^w(D^p, \mathcal{Q}_{S_i}, \tau)) = k$, where k is

$$k = \sum_{\tau \in D^{en}} \sum_{W \subseteq (D^{en} \setminus \{\tau\})} \Delta(\mathcal{Q}_{S_i}, W, \tau) \times p(W \cup D^{ex} \cup \{\tau\})$$

.

Furthermore, each endogenous tuple $\tau \notin S_i$ is *dummy*, and, therefore, by property DUM, $\psi_\tau(\mathcal{Q}_{S_i}) = 0$. Then, by (c): $\psi_\tau(\mathcal{Q}_{S_i}) - Power^w(D^p, \mathcal{Q}_{S_i}, \tau) = \frac{k}{|D^{en} \cap S_i|}$. This expression uniquely defines the value of function ψ_τ for query $\mathcal{Q}_{S_i}$:

$$\psi_\tau(\mathcal{Q}_{S_i}) = \begin{cases} \frac{k}{|D^{en} \cap S_i|} + Power^w(D^p, \mathcal{Q}_{S_i}, \tau) & \text{, if } \tau \in S \cap D^{en} \\ 0 & \text{, otherwise.} \end{cases}$$

Now, by property LIN, we can recursively obtain for $\psi(\mathcal{Q})$:

$$\psi(\mathcal{Q}) = \psi(\mathcal{Q}_l \vee \mathcal{Q}_r) = \psi(\mathcal{Q}_l) + \psi(\mathcal{Q}_r) - \psi(\mathcal{Q}_l \wedge \mathcal{Q}_r),$$

where $\mathcal{Q}_l[W] := (\mathcal{Q}_{S_1} \vee \cdots \vee \mathcal{Q}_{S_k})[W]$ and $\mathcal{Q}_r[W] := (\mathcal{Q}_{S_{k+1}} \vee \cdots \vee \mathcal{Q}_{S_m})[W]$, with $1 \leq k < m$ and $W \subseteq D^p$.[11] Then, as each $\psi_\tau(\mathcal{Q}_{S_i})$ is uniquely defined for any $S_i \in MSS(D^p, \mathcal{Q})$ and $\tau \in D^{en}$, we conclude that $\psi_\tau(\mathcal{Q})$ is uniquely defined for each $\tau \in D^{en}$.

Having established uniqueness, we now turn to proving that the GCES satisfies all the given properties for a monotone query $\mathcal{Q}$ and a PDB D^p with distribution p. First, notice that, for a given MBQ $\mathcal{Q}$, and an endogenous tuple $\tau \in D^p$, the GCES for can be written as:

$$CE(D^p, \mathcal{Q}, \tau) = \sum_{W \subseteq (D^{en} \setminus \{\tau\})} \Delta(\mathcal{Q}, W, \tau) \times (p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})). \quad (6)$$

The property DUM is trivial, since if a tuple τ does not contribute to any tuple, then $\mathcal{Q}[W] - \mathcal{Q}[W \setminus \{\tau\}] = 0$ for every world W.

For property G-EFF, we can group the sum of all the scores by the subsets of D^p, that is:

$$\begin{aligned} \sum_{\tau \in D^{en}} CE(D^p, \mathcal{Q}, \tau) &= \sum_{\tau \in D^{en}} \sum_{W \subseteq (D^{en} \setminus \{\tau\})} \Delta(\mathcal{Q}, W, \tau) \times (p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})) \\ &= \sum_{W \subsetneq D^{en}} \sum_{\tau \in D^{en}} \Delta(\mathcal{Q}, W, \tau) \times (p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})). \end{aligned}$$

Since $\Delta(\mathcal{Q}, W, \tau) = 0$ for every $\tau \in W$, we obtain the desired expression for G-EFF.

For property G-SYM, consider that:

$$CE(D^p, \mathcal{Q}, \tau) - Power^w(D^p, \mathcal{Q}_{S_i}, \tau) = \sum_{W \subseteq (D^{en} \setminus \{\tau\})} \Delta(\mathcal{Q}, W, \tau) \times p(W \cup D^{ex} \cup \{\tau\})$$

Then, to verify G-SYM, we need to show that, for any two tuples $\tau, \tau' \in D^{en}$ satisfying $\Delta(\mathcal{Q}, W, \tau) = \Delta(\mathcal{Q}, W, \tau') = 0$ for every $W \subseteq D^{en} \setminus \{\tau, \tau'\}$, the following holds:

$$\sum_{W \subseteq D^{en} \setminus \{\tau\}} \Delta(\mathcal{Q}, W, \tau) \times p(W \cup D^{ex} \cup \{\tau\}) = \sum_{W \subseteq (D^{en} \setminus \{\tau'\})} \Delta(\mathcal{Q}, W, \tau') \times p(W \cup D^{ex} \cup \{\tau'\}). \quad (7)$$

11. Any query of the form $\mathcal{Q}_{S_i} \wedge \mathcal{Q}_{S_j}$ can be expressed as $\mathcal{Q}_{S_i \cap S_j}$.

Notice that, for τ, $\Delta(\mathcal{Q}, W, \tau) = \Delta(\mathcal{Q}, W, \tau') = 0$ for all $W \subseteq (D^{en} \setminus \{\tau, \tau'\})$. By this, we can express the sums as:

$$\sum_{W \subseteq (D^{en} \setminus \{\tau,\tau'\})} \Delta(\mathcal{Q}, W \cup \{\tau'\}, \tau) \times p(W \cup \{\tau'\} \cup D^{ex} \cup \{\tau\}) \; = \sum_{W \subseteq (D^{en} \setminus \{\tau,\tau'\})} \Delta(\mathcal{Q}, W \cup \{\tau\}, \tau') \times p(W \cup \{\tau\} \cup D^{ex} \cup \{\tau'\}).$$

From the initial condition $\Delta(\mathcal{Q}, W, \tau) = \Delta(\mathcal{Q}, W, \tau')$ for every $W \subseteq D^{en} \setminus \{\tau, \tau'\}$, and the fact that $\mathcal{Q}$ is monotone, we obtain $\mathcal{Q}[W \cup D^{ex} \cup \{\tau\}] = \mathcal{Q}[W \cup D^{ex} \cup \{\tau'\}]$. From this, it follows $\Delta(\mathcal{Q}, W \cup \{\tau'\}, \tau) = \Delta(\mathcal{Q}, W \cup \{\tau\}, \tau')$ for every $W \subseteq D^{en} \setminus \{\tau, \tau'\}$. We obtain that the equality (7) holds, and therefore, GCES satisfies G-SYM.

For LIN, let $\mathcal{Q}$ and $\mathcal{Q}'$ be two monotone Boolean queries. Notice that the expression $p(W \cup D^{ex}) + p(W \cup D^{ex} \cup \{\tau\})$ does not depend on the query, only on the possible world W. Then, it is immediate that, for a given PDB instance D^p and any given endogenous tuple $\tau \in D^{en}$, the following holds:

$$CE(D^p, \mathcal{Q} \wedge \mathcal{Q}', \tau) + CE(D^p, \mathcal{Q} \vee \mathcal{Q}', \tau) = CE(D^p, \mathcal{Q}, \tau) + CE(D^p, \mathcal{Q}', \tau). \qquad \blacksquare$$