

Credit Distribution through Data Provenance in Relational Scientific Databases

Dennis Dosso^a, Susan B. Davidson^b, Gianmaria Silvello^a

^a*Department of Information Engineering, University of Padua, Italy*

^b*Department of Computer and Information Science, University of Pennsylvania, USA*

Abstract

Digital data is an important form of research product for which citation, and the generation of credit or recognition for authors, is still not well understood. The notion of *data credit* has therefore recently emerged as a new metric, defined and based on data citation theory.

Data credit is a real value that represents the importance of data cited by a paper or by another research entity. Credit can be used to annotate data contained in a curated scientific database, and then used as a measure for the importance and impact of that data in the research world. As such, it is a new method that, together with traditional citations, helps recognize the value of data and its creators.

In this paper we explore the problem of Data Credit Distribution, the process by which credit is distributed to the database parts responsible for the production of data being cited by a research entity.

We adopt as use case the IUPHAR/BPS Guide to Pharmacology (GtoPdb), a widely-used curated scientific relational database. We define four new distribution strategies, the first two based on two forms of data provenance, why-provenance and how-provenance, the third based on the concept of responsibility, the fourth on the Shapley value.

Using these distribution strategies we show how credit can highlight frequently used database areas and how it can be used as a new bibliometric measure for data and their corresponding curators. In particular, credit rewards data and authors based on their research impact, not merely on the number of citations. We also show how these distribution strategies vary in their sensitivity to the role of an input tuple in the generation of the output data, and reward input tuples differently.

Keywords: Data Citation, Data Credit

1. Introduction

Citations are an essential component of scientific research, enabling research products to be found as well as the relationships between them to be created and understood. They form a basis on which to give credit to authors, papers, and venues [20, 21, 62]. Citations are used, among other things, to decide on tenure, promotion, hiring, and funding of grants for researchers [22, 35, 41, 45].

Science and research are increasingly digital, and there are numerous curated databases that are at the core of scientific research efforts [12]. It is therefore generally accepted that data must be cited and citable [15, 42], and that data citations should contribute to the scientific reputation of researchers, scientists, data curators, and creators [4, 57]. It is also accepted that data citations should be counted alongside of traditional citations, and contribute to bibliometrics indicators [7, 49].

A central problem in the data citation process is how to attribute credit to data creators and curators [11]. How to handle and count the credit generated by data citation, and how it contributes to traditional and new bibliometrics, are long-standing research issues [9, 30]. However, even when correctly applied, data citations and the bibliometrics computed using them do not always fully reward the creators of data used in a database. Data, in fact, is often cited at the “database level” or the “webpage level”. In the first case, the whole database is cited and therefore all credit goes to the key personnel of the database. In the second case, the database has a website with webpages that can be individually cited. The webpages use data extracted from the database, which is aggregated by topic and built to resemble a traditional research paper. Often the creators and curators of the webpage’s data are not credited or only marginally credited for their work [3].

Recently, the idea of *Data Credit Distribution* (DCD) [29, 40, 61] has emerged, built on top of methodologies for data citation. Data credit is a value that is computed based on the importance of the data being cited in a paper, and is a proxy for the impact of the data on the citing paper. The DCD problem consists of distributing this credit to elements in the databases in the citation graph that are responsible for the generation of the data being cited. The goal of DCD is to improve and expand the reach of data citation,

36 rather than being an alternative to it.

37 In this paper, we consider data credit as a measure of value for data in
38 a (curated) scientific database. Credit is a real value that can be assigned
39 to data of any kind and at any level of granularity. Therefore the concept
40 of “data” is left intentionally vague, although in this paper we focus on
41 relational databases. Credit acts as a proxy for the value of data based on
42 the measure of citations, accesses, clicks, downloads, or other surrogates for
43 data use.

44 We define DCD as *the process, method, or algorithm used to assign credit*
45 *to a given datum or dataset*. It differs from the traditional citation setting
46 since:

- 47 1. When a paper p_1 cites another paper p_2 , a +1 citation “credit” is given
48 to p_2 , and to all its authors. It does not matter why or how paper
49 p_1 cites paper p_2 ¹, the result is always +1 to the citation count of p_2
50 and of its authors. A different credit distribution strategy can assign a
51 quantity of credit to p_2 and its authors that is *proportional* to the role
52 played by p_2 in p_1 . Hence, we can weight the importance of the cited
53 entities and assign credit according to their role.
- 54 2. Traditional citations are *atomic*: a citation from p_1 to p_2 can never
55 be broken into pieces and assigned in part to p_2 and in part to other
56 papers or data that contributed to p_2 . In contrast, with data credit,
57 we use a *non-atomic* real value, which can be divided and distributed
58 to multiple components of a database.
- 59 3. Credit can be *transitive*, that is, it can be propagated through one
60 cited entity to other entities cited by it that contributed to its content.
61 Citations, traditionally, are not.

62 We study the DCD problem in the context of relational databases (RDBs)
63 since they are widely used² and are the main focus of current work in data
64 citation methods [12, 14, 50]. RDBs are also frequently a test-bed for new
65 methods that can be adapted to other databases, e.g., graphs or document
66 databases. The “portions” of data in an RDB that can be credited can be
67 defined at different levels of granularity, in particular: (i) the whole database,
68 (ii) tables, (iii) tuples, and (iv) attributes. The ability to specify different

¹Note that there is vast research on this topic and many alternative proposals, but none of them currently work at a large scale.

²The “relational database market alone has revenue upwards of \$50B” [1].



Figure 1: Overview of the credit distribution pipeline.

69 levels of granularity in a relational database allows us to define the DCD
70 problem at a particular level of granularity. In this paper, we focus on DCD
71 at the tuple level.

72 The DCD process that we use is summarized in Figure 1:

73 **Step 1** Scientists and experts contribute the curated information contained
74 in a scientific database. These are called the “Data Curators”.

75 **Step 2** Other researchers use the data in their research, and when possible,
76 cite them.

77 **Step 3** The citation to the data generates credit, that can be used as a
78 proxy for the impact of the data on the citing paper. This credit is
79 represented as a real value $k \in \mathbb{R}_{>0}$.

80 **Step 4** Given the database instance I and the query Q , the *data prove-*
81 *nance* of $Q(I)$ is computed. The data provenance of $Q(I)$ is a form of
82 metadata that captures how Q used I to generate the output [17].

83 **Step 5** Provenance is input to the *Credit Distribution Strategy* (CDS, also
84 referred only as *Distribution Strategy*, DS). CDS is a function f that

85 takes as input the credit value k , divides it and distributes it to the
 86 data in the input database I , and is defined on the basis of citation
 87 policies decided at the database administration level or at the domain
 88 community level.

89 **Step 6** Once the CDS is computed, it is used to distribute the given credit
 90 k to the parts of the database that are responsible for the generation
 91 of $Q(I)$. Transitively, this credit is also divided and given to the corre-
 92 sponding authors of those data.

93 This paper expands the work in [26] where we first defined the problem
 94 of DCD in relational databases, and proposed a viable Distribution Strategy
 95 (DS) based on *lineage* – the simplest form of *data provenance*. The lineage
 96 of a tuple t in the output $Q(I)$ is defined as the set of all and only the tuples
 97 in the database instance I that are “relevant” to the production of t . The
 98 corresponding strategy equally redistributes the credit k to the tuples in the
 99 lineage set, thus each tuple receives credit $k/|L_t|$, where L_t is the lineage set
 100 of t .

101 One may argue that this DS is too simplistic, since lineage does not convey
 102 any information about the role or importance of input tuples in the query.
 103 Therefore, one may desire to give more credit to the tuples that are more
 104 *important* to the production of the output, i.e. those tuples that, if removed,
 105 would prevent the output tuple from appearing in the final result, or those
 106 tuples used more than once by the query.

107 Therefore, in this paper, we expand the ideas in [26] by proposing new
 108 DSs based on two other forms of data provenance: why-provenance [13] and
 109 how-provenance [32]. Also, we propose other two DS based on the concepts
 110 of responsibility [46], and the Shapley value[25, 43]. We show how these DS
 111 differ from each other as well as the one based on lineage, and discuss why
 112 one may be preferred to another depending on the application and its goals.
 113 In particular, we show that the new proposed DSs are more sensitive than the
 114 one based on lineage to the *role* of a tuple in a query, i.e. how many times the
 115 tuple is used and how it is used. We also show that the DSs based on why-
 116 provenance and responsibility give more credit to tuples that are essential to
 117 the production of the result set, whereas the how-provenance-based DS takes
 118 into consideration the different ways in which a tuple is used. Finally, the DS
 119 based on the Shapley value sees the process of distribution as a competitive
 120 game where tuples that contribute more to the generation of the output are
 121 correspondingly rewarded more.

122 The evaluation is based on a well-known curated database, the IUPHAR/BPS³
123 Guide to Pharmacology [34], also known as GtoPdb⁴, which contains ex-
124 pertly curated information about diseases, drugs, cellular drug targets, and
125 their mechanisms of action. We chose GtoPdb for two main reasons: (i) it
126 is a widely-used and valuable curated relational database, (ii) many papers
127 in the literature use, and cite, its data (i.e., families, ligands, and receptors).
128 Real queries used in papers can therefore be seen as data citations which, in
129 turn, can be used to assign data credit.

130 We perform four sets of experiments. In the first, real queries are ex-
131 tracted from papers published in the British Journal of Pharmacology (BJP),
132 that represent data citations to GtoPdb, and are used to distribute credit in
133 the database using the three different provenance-based DSs. In the second
134 and third experiment we analyze the behavior of the different DS when com-
135 plex citation queries are employed. In the fourth set of experiments we use
136 both real and synthetic queries to assess the difference between traditional
137 citation and the notion of credit distribution in terms of rewarding those
138 responsible for the data, e.g. data curators.

139 **Contributions** of this work include:

- 140 • Four new Distribution Strategies based on why-provenance, how-provenance,
141 responsibility and the Shapley value.
- 142 • An in-depth analysis of the effects of credit distribution on real-world
143 curated data and of the differences between the five proposed Distri-
144 bution Strategies.
- 145 • A comparison between the behavior of traditional citations and data
146 credit in rewarding data curators.

147 **Outline.** The rest of the paper is organized as follows: Section 2 presents
148 background material and related work. Section 3 describes the GtoPdb use
149 case. Section 4 briefly presents the forms of provenance used in the paper.
150 Section 5 describes the credit distribution problem and the proposed dis-
151 tribution strategies. In Section 6 we present the experimental evaluation,
152 followed by a discussion of our design decisions in Section 7. Section 8 draws
153 some conclusions and outlines future work.

³International Union of Basic and Clinical Pharmacology/British Pharmacology Soci-
ety

⁴<https://www.guidetopharmacology.org/>

154 2. Background

155 *Data in Research.* The world of research is rapidly transitioning towards the
 156 *fourth paradigm of science* [36], that is, data-intensive scientific discovery,
 157 where data are important for scientific advances as well as for traditional
 158 publications [6].

159 The scientific community is promoting an *open research culture* [48],
 160 founded on methods and tools to share, discover, and access experimental
 161 data. The community has identified the FAIR principles (Findable, Acces-
 162 sible, Interoperable, and Reusable) [59], that should be enforced by every
 163 database. In particular, data should be accessible from the articles, journals,
 164 and papers that cite or use them [20]. Aspects such as the need for the *repro-*
 165 *ducibility* of experiments through the used data; the *availability* of scientific
 166 data; the *connections* between data and the scientific results are all needed
 167 aspects for the fourth paradigm, and are all relevant to the domain of *data*
 168 *citation* [37].

169 *Data Citation: Principles and Motivations.* Data Citation principles were
 170 proposed in [19], and later summarized and endorsed by the Joint Declaration
 171 of Data Citation Principles (JDDCP) [44]. The principles are divided into
 172 two groups [55]. The first group contains principles concerning the role of
 173 data citation in scholarly and research activities such as the (i) *importance*
 174 of data (why data citation is important and why data should be considered
 175 as first-class citizens); (ii) *credit* and *attribution* to the creators and curators
 176 of the data; (iii) *evidence*; (iv) *verifiability*; and *interoperability*, with these
 177 last three requiring data citation methods to be flexible enough to operate
 178 through different communities. The second group defines the main guidelines
 179 to establish a data citation systems, and contains principles such as the (i)
 180 *unique identification* of the data being cited; (ii) *(open) access* to data; (iii)
 181 guarantee of *persistence* and *availability* of citations even after the lifespan
 182 of the cited entity; the (iv) *specificity* of a citation, i.e. it must lead to the
 183 data set originally cited.

184 * SBD: Is the next paragraph necessary? Could we just say
 185 "The main motivations for data citation are outlined in [55]." *

186 It is possible to outline six main motivations for data citation [55]:

- 187 • *Data attribution:* identify the individuals that should be credited for
 188 data with variable granularity.

- 189 • *Data connection*: connect papers to the data being used.
- 190 • *Data Discovery*: citations helps to find data records and subsets that
191 would be otherwise not findable via search engines.
- 192 • *Data Sharing*: share data obtained by researchers within the whole
193 community.
- 194 • *Data Impact*: highlight the results obtained in writing papers using
195 specific data, the frequency and modality data were used.
- 196 • *Reproducibility*: data citation greatly impacts the reproducibility of
197 science [5]. Many authoritative journals ask to share data and provide
198 valid methodologies to reproduce experiments.

199 2.1. *Data Citation in Relational Databases*

200 Relational databases have been the target of data citation methods since
201 the surge of the data-centric research paradigm. The RDA “Working Group
202 on Data Citation: Making Dynamic Data Citable”⁵ [51] has developed guide-
203 lines for citing large, dynamic, and changing datasets which have now moved
204 on into adoption phase. The datasets considered by the Working Group are
205 often relational.

206 In one of its most recent sessions [52], the Working Group (WG) on
207 Data Citation reported that there are various implementations of its guide-
208 lines for Data Citation on MySQL/Postgres relational databases. Some of
209 these databases are: DEXHELPP⁶ (Social Security Records); NERC (ARGO
210 Global Array); EODC (Earth Observation Data Centre) [31]; LNEC (River
211 dam monitoring); MDS (Million Song Database) [8]; CBMI⁷ (Center for
212 Biomedical Informatics); VMC (Vermont Monitoring Cooperative); CCA⁸
213 (Climatic Change Center Austria); VAMDC (Virtual Atomic and Molecular
214 Data Center) [27, 63].

215 More examples of work on data citation in relational databases are [2,
216 12, 24, 60]. The website <https://fairsharing.org/> keeps an updated list

⁵<https://www.rd-alliance.org/groups/data-citation-wg.html>

⁶<http://www.dexhelpp.at/>

⁷<https://medicine.missouri.edu/centers-institutes-labs/center-for-biomedical-informatics>

⁸<https://ccca.ac.at/startseite>

217 of curated and scientific databases (many of which are relational or graph-
218 based) following FAIR guidelines. These databases are citable since they are
219 compliant with the most recent guidelines, and they are in the vast majority
220 of cases accessible via dynamically created Webpages. In all these databases
221 it is, therefore, possible to implement DCD on top of the existing infrastruc-
222 tures for citing data.

223 Data citation techniques are primarily applied to relational databases
224 because of their pervasiveness as well as the “identifiability” of the portions
225 of data that are to be cited: the whole database, a relation, a tuple, or
226 even an attribute. Many papers [2, 10, 12] consider more complex citable
227 units, recognizing that often the *views* of a database are the ones to be cited.
228 Generally, a *view* is a query on the database. To this end, [60] suggested
229 decomposing the database into a set of views, where each view is associated
230 with its citation.

231 At present, the most common practices to cite databases include:

- 232 1. A database cited as a whole, even though only parts of the databases
233 are used in the papers or datasets. Alternatively, the so-called “data pa-
234 pers” are cited, being traditional papers that describe a database [16].
235 In this case, all the credit from the citations goes to the database ad-
236 ministrators or to the authors of the data papers.
- 237 2. Subsets of data, obtained by issuing queries to a database, are individ-
238 ually cited. This is the solution adopted by the *Resource Data Alliance*
239 (RDA) working group on Data Citation [51]. In this case, the credit
240 generated from citations is distributed among the contributors of the
241 portions of data being cited, and/or to the database administrators.
- 242 3. The database is accessible via a series of Webpages that arrange the
243 content of the database by topic or theme. Examples in the life science
244 domain include the Reactome Pathway database [39], the GtoPdb [34],
245 and the VAMDC [63]. Every single Webpage is unequivocally identifi-
246 able and can be individually cited.

247 2.2. Data Credit

248 Data credit is related to data citation: they both aim to recognize the
249 work of data creators and curators. Data credit can therefore also be seen as
250 a by-product of data citation, since credit attribution is impossible without
251 the presence of data citations.

252 Katz [40] suggests the need for a *modified citation system* that includes
253 the idea of *transient* and *fractional credit*, to be used by developers of research

254 products as software and data. Two considerations are made: (i) research
 255 objects such as data and software are currently not formally rewarded or
 256 recognized by the community; (ii) even in traditional papers, the contribution
 257 of each author to the work is hard to understand, unless explicitly specified in
 258 the paper. This is even more true for data, where different groups of people
 259 work on the same database.

260 In [40] credit is defined as a “quantity” that describes the importance of a
 261 research entity, such as papers, software, or data, mentioned in a citation. It
 262 also proposed the idea of a *distribution* of credit from research entities, such
 263 as papers or data, to other research entities through citations. *Therefore,*
 264 *when discussing data credit, we need to consider credit computation – i.e.,*
 265 *the process to compute the quantity of credit generated by the citation – and*
 266 *credit distribution – i.e., the process to distribute credit and to assign it to*
 267 *the entities that contributed to the creation/curation of the cited data. In*
 268 *this paper we focus on the latter.*

269 *These two processes* are done by exploiting the structure of the *citation*
 270 *graph*, a directed graph whose nodes are publications and edges are citations.
 271 This graph is the model at the core of systems such as Google Scholar and
 272 the Web of Science. We add to this that the concept of credit can be built
 273 on top of the existing infrastructure handling traditional and data citations.

274 Katz [40] further explores the idea of a *distribution* of credit from research
 275 entities (i.e., papers and data) to other research entities through citations
 276 that connect them. Thanks to traditional citations and now also to data
 277 citations, this distribution is finally possible, at least between papers and
 278 data. Some problems related to traditional citations can thus be solved by
 279 citations:

- 280 1. Credit rewards research entities that to date are not (formally) recog-
 281 nized (a goal shared with data citation).
- 282 2. Credit can reward authors *proportionally* to their role in generating the
 283 entity. The more an author contributes to a paper, the more credit is
 284 given to him. Zou and Peterson [62] work on something similar with
 285 their zp-index, which includes in its formulation the position (and thus
 286 the role) of a publication author to represent its impact in the work
 287 itself.
- 288 3. Credit can be *transitively* channeled through a chain of papers citing
 289 each other, thus enabling the rewarding of older papers that are no
 290 more cited, since other papers summarize or report their content but

291 are nevertheless crucial in a research area for the influence of their
292 content.

293 Fang [29] presents a framework to distribute the credit generated by a
294 paper to its authors and to the papers in its reference list in a transitive way.
295 Let us consider the *citation graph* as the graph where the nodes are papers
296 and the links are the citations among them. In this graph, every paper is
297 a source of credit, which is then transferred to the neighboring nodes. The
298 quantity of credit received by each cited paper depends on its impact/role
299 in the citing paper. So far, this theoretical framework is limited to papers,
300 but it can be easily extended to a citation graph including both papers and
301 data.

302 Zeng et al. [61] proposes the first method to compute credit within a net-
303 work of papers citing data. Adopting a network flow algorithm, they simulate
304 a random walker to estimate a score for each dataset, leveraging real-world
305 usage data to compute the credit. This is the first step towards an automatic
306 credit computation procedure. This proposal is, however, limited to assign-
307 ing credit to whole datasets, and it does not deal with the granularity of data.
308 It does not work to assign credit to a single research entity within a dataset.
309 Differently from Zeng et al. [61], we do not treat the credit computation
310 process, but we focus on the distribution process.

311 2.3. Data Provenance

312 To distribute credit, we base our methods on *data provenance*. Data
313 provenance is information that describes the origin and the process of cre-
314 ation of data. It can also be seen as metadata pertaining to the derivation
315 history of the data. It is particularly useful to help users to understand
316 where data are coming from, and the process they went through. Data ci-
317 tation and data provenance are closely linked [3] since both are forms of
318 annotations on data retrieved through queries. Data provenance has been
319 widely studied in different areas of data management. In this paper, we fo-
320 cus on provenance for database management systems (DBMS). For further
321 details on data provenance, please refer to surveys like [17] and [56].

322 Cheney et al. [17] presents four main types of data citation for DBMS: *lin-*
323 *age* [23], *why-provenance* [13], *how-provenance* [32] and *where-provenance* [13].

324 Let us start with the first three provenances. Given a database instance
325 I , a query Q , and the result $Q(I)$, consider one tuple t of the output. Its
326 provenance is information about its generation through the tuples of the

input that are used by Q . Different types of provenance convey different levels of information. Since these three provenances are computed for each tuple of the output, they are also referred to as *tuple-based*.

Where-provenance, differently from the other three, is *attribute-based*, so we do not take it into account in this work since we consider the tuple as the finest citable unit.

2.4. Causality and Responsibility

We also consider the notions of causality and responsibility, as defined in [46]. Causality is an enrichment of lineage, and it is the attribution of a certain degree of importance to the tuples of the lineage based on their role in the generation of the output. Responsibility is a value given to the tuples of the lineage to rank them based on their degree of causality (the more important the role of a tuple in generating the output, the higher its responsibility).

While computing responsibility for general queries is hard [18], Meliou et al. [46] proved a dichotomy result for conjunctive queries: for each query without self-joins, either its responsibility can be computed in PTIME in the size of the database or checking if it has a responsibility below a given value is NP-hard.

2.5. Shapley value

The Shapley value is named after Lloyd Shapley, who introduced it for the first time in his 1952 work [54]. He considered a *cooperative game* played by a set A of players, defined by a *wealth function* v that assigns to each coalition set $B \subseteq A$ the wealth $v(B)$. The question behind the Shapley Value is how to quantify the contribution of each player to the overall wealth. Informally, the Shapley value is defined as follows [43]: assume that we select players randomly one by one and without replacement, starting with the empty set. Every time a player a is selected, its addition to the coalition B produces a change in the wealth of the coalition from $v(B)$ to $v(B \cup \{a\})$. The Shapley value of a is the expectation of change that a causes in this probabilistic process.

The Shapley value can be used in different research areas beyond cooperative games, such as economics, law, environmental science, and network analysis, and it has strong theoretical justifications. However, its use in databases as a metric for quantifying the influence of a tuple on the output of a query (thereby presenting an alternative to responsibility) has only

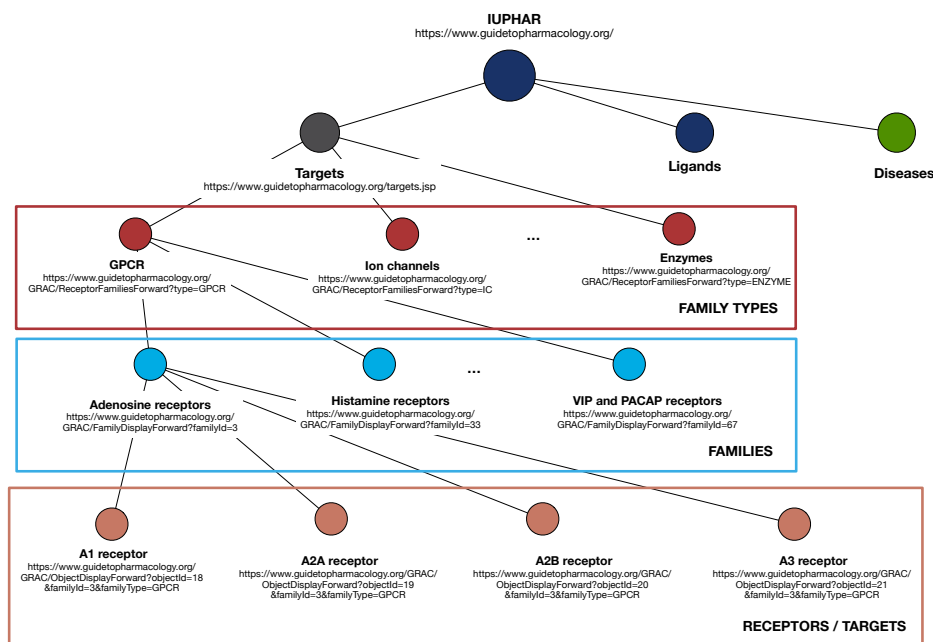


Figure 2: Partial map of the GtoPdb hierarchical structure grouping the targets into families and family types.

recently been proposed [43]. The initial theoretical analysis in [43] showed mainly lower bounds on the complexity of the problem, and did not suggest a feasible implementation. However, very recently, an efficient implementation for Boolean queries (queries that output true or false, or 1 or 0) has been provided [25], both in terms of an exact computations (which in practice works well for most queries) and in inexact one (which is extremely fast and provides the same ranking of tuples as the exact computation, but not necessarily the same values).

3. Use Case: GtoPdb

The IUPHAR/BPS Guide to Pharmacology [34] (GtoPdb⁹) is a well-known and well structured scientific relational database that contains expertly curated information about diseases, drugs in clinical use, their cellular targets, and the mechanisms of action on the human body. It is curated and

⁹<https://www.guidetopharmacology.org/>

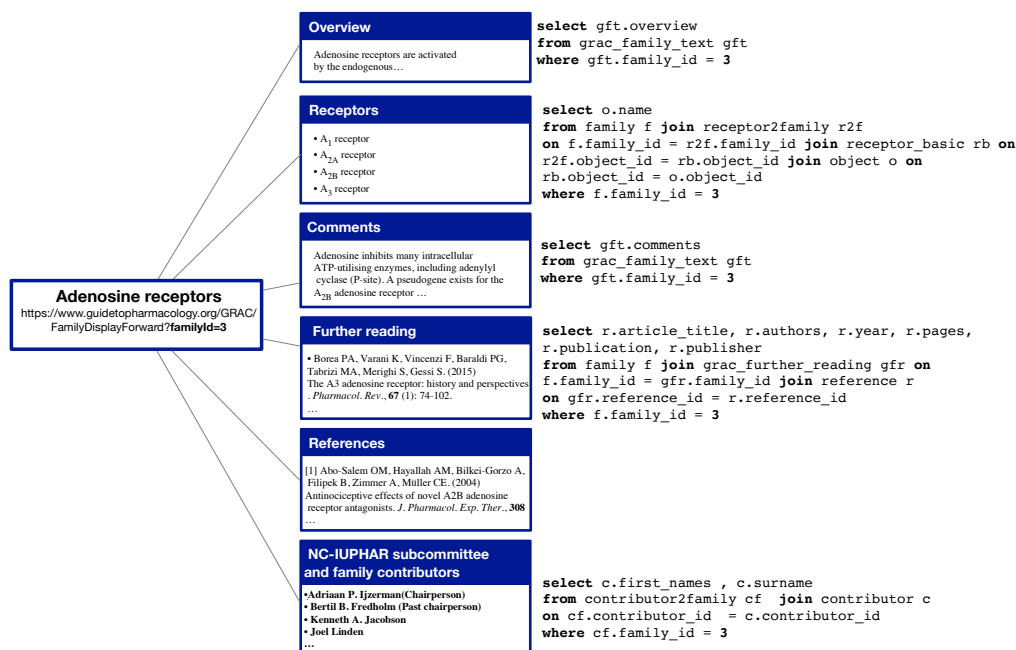


Figure 3: Basic web-page structure of “Adenosine receptors” family (ID 3), with queries used to retrieve the information contained in every section, except references.

376 maintained by the GtoPdb Committee and 96 subcommittees, comprising
 377 512 scientists collaborating with in-house curators who draw the information
 378 contained in the database from high-quality pharmacological and medicinal
 379 chemistry literature. Roughly 1000 researchers from all over the world have
 380 contributed to the database, and the curators wanted to give recognition to
 381 these contributors. This led to some early work on data citation [10].

382 GtoPdb is relational, but its logical structure is hierarchical as shown
 383 in Figure 2. The information contained in the database is also organized
 384 into webpages focused on specific diseases, targets or ligands, and families
 385 for easier access by users. As depicted in Figure 2, the database can be
 386 thought of as a tree where the root is the database; the first level consists
 387 of all targets, ligands, and diseases; and the lower levels consists of specific
 388 targets, ligands and diseases. In this paper, we focus on targets; thus the
 389 figure at the third level shows examples of family types, at the fourth level
 390 of specific families of targets (a finer level of granularity), and finally, at the
 391 last level, the single targets (also known as receptors).

392 GtoPdb provides access to the webpages corresponding to all these nodes

393 through URLs. The webpages corresponding to target families all present a
 394 similar structure, as shown in Figure 3 for the “Adenosine receptors” family.
 395 Each page has an *Overview*, a brief text describing the content of the page;
 396 a list of *Receptors* comprising the family; a section of *comments* about the
 397 family; the *References*, a list of the papers consulted by the curators of the
 398 page, similar to a reference list of a paper; the *further reading* list, reporting
 399 papers that an interested reader may want to consult to obtain more insight
 400 on the family; and a final section called *How to cite this family page*, con-
 401 taining text snippets useful to cite the specific page or the whole database.
 402 Figure 3 shows the SQL code that retrieves the information used to build the
 403 corresponding sections (apart from the References section). Therefore, each
 404 family page can be considered a full-fledged traditional publication, consist-
 405 ing of title, authors, abstract (the overview), content, and references.

406 In practice, many papers in the literature only reference GtoPdb (the
 407 root) without including a reference to the specific page being cited. That is,
 408 they only cite a paper describing GtoPdb as a whole (e.g., [34]) and refer
 409 to targets, ligands, diseases, etc. only by name. Thus, citations to specific
 410 families are *de-facto* “hidden” to citation systems such as Google Scholar,
 411 and useless for the computation of bibliometrics.

412 In certain “lucky” cases, as with papers available in PDF and published
 413 in the British Journal of Clinical Pharmacology ¹⁰ (BJCP), when a family,
 414 ligand, receptor name, etc. are used, they have a hyperlink pointing to the
 415 corresponding webpage in GtoPdb. Therefore, the citations to the families
 416 can be detected and counted using the URLs reported in the papers. How-
 417 ever, these citations to GtoPdb webpages are not counted as such by citation
 418 systems, so they are not converted into credit for curators and collaborators.

419 For our running example, consider Table 1. This simplified version of
 420 GtoPdb contains three tables: **family**, **contributor** and **contributor2family**.
 421 The first table, **family**, has tuples representing families with three attributes:
 422 the id of the family, its name, and type. Table **contributor** contains peo-
 423 ple who have helped generate the data in the database. The third table,
 424 **contributor2family**, serves as a link between the families and the people
 425 who contributed to them. For instance, “John Smith” (c_1) contributed to
 426 “Dopamine Receptors” (f_1) as well as to the “YANK Family” (f_4). Through-
 427 out the rest of the paper, we will use the **id** attribute of these tables as the

¹⁰<https://bpspubs.onlinelibrary.wiley.com/journal/13652125>

family			contributor2family		
id	name	type	id	family_id	contributor_id
f_1	Dopamine Receptors	gpcr	$c2f_1$	f_1	c_1
f_2	Bile Acid Receptor	gpcr	$c2f_2$	f_1	c_2
f_3	FAK Family	enzyme	$c2f_3$	f_2	c_3
f_4	YANK Family	enzyme	$c2f_4$	f_4	c_1

contributor		
id	Name	Country
c_1	John Smith	UK
c_2	Jim Doe	UK
c_3	Hans Zimmerman	Germany
c_4	Roberta Rossi	Italy

Table 1: Example of a database consisting of three tables. **family** contains receptor families; **contributor** contains the name and country of contributors; **contributor2family** connects contributors to the families they contributed to.

428 *provenance token* of its corresponding tuples, that is, as a symbol that serves
429 to identify a tuple when talking about provenance.

430 4. Data Provenances

431 We now describe the three types of provenance used in this paper – lin-
432 eage, why-provenance, and how-provenance – as well as the notion of Causal-
433 ity and Responsibility, and the Shapley value function.

434 4.1. Lineage

435 Lineage is the simplest form of provenance. It was first introduced by
436 Cui et al. [23], and can be thought of as the set of all tuples that are used
437 by the query to generate the output [17].

438 As an example, consider the following SQL query **Q1**, applied to the
439 database described in Table 1, asking for the names of families curated by
440 researchers based in the United Kingdom (UK):

```

441 Q1: SELECT DISTINCT f.name
442 FROM family AS f JOIN contributor2family AS c2f
443 ON f.id = c2f.family_id
444 JOIN contributor AS c ON c2f.contributor_id = c.id
445 WHERE c.country = 'UK'
```


I	database instance
L	lineage set of an output tuple
Γ	contingency set
ρ_t	responsibility of tuple t
Q	a query
I^n	set of endogenous tuples
I^x	set of exogenous tuples
\mathcal{W}	witness basis
W	a witness set
$\gamma(\mathcal{W}, t)$	set of witnesses in \mathcal{W} containing t
\mathcal{H}	provenance polynomial
M_i	a monomial in \mathcal{H}
t_j	a tuple in M_i
$c(\mathcal{H})$	sum of \mathcal{H} 's coefficients
$e(M_i)$	sum of M_i 's exponents
$mc(M_i)$	M_i 's coefficient
$te(t_j, M_i)$	exponent of t_j in M_i
$\gamma(t_j, \mathcal{H})$	set of monomials in \mathcal{H} containing t_j

Table 2: **Notations used in this paper.**

id	name	lineage
o_1	Dopamine Receptors	$\{f_1, c2f_1, c_1, c2f_2, c_2\}$
o_2	YANK Family	$\{f_4, c2f_4, c_1\}$

Table 3: Result of **Q1** over the database instance in Table 1 with the lineage of each output tuple. Attribute **id** is not part of the output, and was added to identify each tuple.

Table 3 shows the query output, which consists of two tuples. We add an extra attribute **id** so that we can easily refer to each result tuple. The lineage for tuple o_1 is the set $\{f_1, c2f_1, c_1, c2f_2, c_2\}$, since the tuple f_1 was joined with $c2f_1$ and then with c_1 , and was also joined with $c2f_2$ and c_2 . No other tuple is used in the database to produce o_1 . For tuple o_2 the lineage is $\{f_4, c2f_4, c_1\}$. Lineage is defined for each tuple of the output, and can differ between tuples.

4.2. Why-Provenance

Why-Provenance was first defined in terms of a deterministic semistructured data model and query language [13]. We use here its definition in terms of the relational model [17].

While lineage aims to find all and only the tuples in the input relevant to the production of an output tuple, why-provenance aims to find sub-instances of the input that “witness” a part of the output. Given a tuple t in the query’s output $Q(I)$, a *witness* is any sub-instance of the database that produces t , i.e., a set that guarantees the existence of t in $Q(I)$. In particular, the whole database and the lineage of t are both examples of witnesses of t . Since the definition of witness allows for the presence of “irrelevant” tuples, the set of all witnesses is finite (since the database instance I is finite), but it is potentially exponentially large [17].

Buneman et al. [13] defined the why-provenance of an output tuple t in the result $Q(I)$ as a special *subset* of the set of witnesses called the *witness basis*. The witnesses of the basis exclude tuples that are irrelevant to t being produced by Q , and thus the basis tends to be very small compared to the set of all possible witnesses [17].

In a sense, each witness in the witness basis captures one possible way in which a tuple in the output was generated by the query. To better understand this, consider the example in Table 4, where each tuple in the result of query **Q1** is annotated with its why-provenance.

id	name	why-provenance
o_1	Dopamine Receptors	$\{\{f_1, c2f_1, c_1\}, \{f_1, c2f_2, c_2\}\}$
o_2	YANK Family	$\{\{f_4, c2f_4, c_1\}\}$

Table 4: Result of Q1 over the database instance in Table 1 with the why-provenance of each output tuple.

475 The why-provenance of output tuple o_2 has only one witness, which co-
476 incides with its lineage. This happens because there is only one way this
477 output tuple can be produced, i.e., for tuple f_4 to be joined with $c2f_4$ and c_1 .
478 On the other hand, o_1 has a witness basis of two witnesses, since there are
479 two possible ways in which the query can generate o_1 . One possibility is that
480 f_1 is joined with $c2f_1$ and c_1 (the first witness), and the second possibility
481 is that f_1 is joined with $c2f_2$ and c_2 (the second witness). This means that
482 to generate o_1 , it is sufficient that only one of the two witnesses is present in
483 the input database.

484 4.3. How-Provenance

485 While why-provenance describes the source tuples that witness an output
486 tuple in the result of the query, it leaves out information about how the source
487 tuples are used. How-provenance was therefore defined in [32] to capture
488 this information using a *semiring* algebraic structure. It takes the form of
489 a polynomial, called *provenance polynomial*, where the variables are taken
490 from the set X of identifiers of the tuples (provided that each tuple in I has
491 an identifier) and the coefficients are drawn from the set of natural numbers
492 \mathbb{N} .¹¹

493 The key idea in Green et al. [32] is to use the two operators $+$ and \cdot to
494 represent two basic transformations that source tuples undergo as a result
495 of applying a relational query to a database [17]. Two tuples may either
496 be joined together (a join is represented with the \cdot operator) or merged via
497 union or projection (represented with the $+$ operator).

498 Table 5 shows the two output tuples of our running example annotated
499 with their respective how-provenances. Tuple o_2 was produced by a join of
500 the input tuples f_4 , $c2f_4$, and c_1 . The three provenance tokens are therefore
501 “multiplied” together. The case of o_1 is slightly more complex, as already
502 discussed. It can be obtained by the joins of two different sets of tuples,

¹¹This semiring is commonly referred as $\mathbb{N}[X]$ in the literature.

id	name	how-provenance
o_1	Dopamine Receptors	$f_1 \cdot c2f_1 \cdot c_1 + f_1 \cdot c2f_2 \cdot c_2$
o_2	YANK Family	$f_4 \cdot c2f_4 \cdot c_1$

Table 5: Result of Q1 over the database instance in Table 1 with the how-provenance polynomial of each output tuple.

so there are two monomials combined by + representing these alternative derivations. Each monomial corresponds, in a way, to the witnesses of the why-provenance of o_1 .

Provenance polynomials may also have monomials whose exponents and/or coefficients are greater than one, for example, $3f_1 \cdot c2f_1 \cdot c_1 + f_1 \cdot c2f_2^3 \cdot c_2^3$. This is a polynomial of a tuple produced by a query where the result of the join between the tuples f_1 , $c2f_1$, and c_1 is produced three times and then merged (e.g. as the result of a union), and the tuples $c2f_2$ and c_2 are used three times in the operation described by the second monomial (e.g., with nested queries).

513

514 4.4. Causality and Responsibility

A formal study of causality was introduced in [18, 33] and later expanded by Meliou et al. [46] to explain the causes of answers and non-answers to queries. In the following, we refer to the definition of causality and responsibility provided in [46]. In particular, we only focus on answers to a query since non-answers are not relevant in our context.

There are two types of “cause” tuples: counterfactual and actual. Let o be a tuple in the result of query q on the database instance I , and t a tuple in its lineage. We call t a *counterfactual cause* if, by removing t from I , o is also removed from the output (i.e., t is essential for the generation of t). We call t an *actual cause* if there is a set of tuples $\Gamma \subseteq I$ called a *contingency set*, such that t is a counterfactual cause in $I - \Gamma$. In other words, t is an actual cause if, even when removed from I , there is another set of tuples of the lineage that guarantees the presence of o .

Computing the causality of tuples is NP-complete for general queries [28], but for conjunctive queries can be computed in PTIME, as showed by Meliou et al. [46].

The notion of *responsibility* measures the degree of causality as a function of the size of the smallest contingency set [18]. This allows us to rank lineage

id	name	responsibility
o_1	Dopamine Receptors	$f_1 = 1, c_2f_1 = 0.5, c_2f_2 = 0.5, c_1 = 0.5, c_2 = 0.5$
o_2	YANK Family	$f_4 = 1, c_2f_4 = 1, c_1 = 1$

Table 6: Result of Q1 over the database instance in Table 1 with the responsibilities of lineage tuples.

533 tuples based on their degree of causality in generating the output.

Definition 4.1. *Responsibility [46]*

Let o be an output tuple in the result of query Q on I , and let t be a cause for o . The responsibility of t for the answer o is:

$$\rho_t = \frac{1}{1 + \min_{\Gamma} |\Gamma|}$$

534 where Γ ranges over all contingency sets for t .

535 Note that a counterfactual cause will have the maximum responsibility
536 of 1, and that the larger the minimum contingency of an actual cause is, the
537 smaller its responsibility will be since there are alternatives to guarantee the
538 presence of the answer o .

539 As an example, consider Table 5, where we reported the result set of Q1
540 and the tuples of the lineages with their responsibility values. Focusing on
541 o_1 : the lineage tuple f_1 is a counterfactual cause, since its contingency set is
542 empty (when removed from the database, o_1 disappears from the result set).
543 Consequently, its responsibility is 1. All the other tuples of the lineage are
544 actual causes. c_1 , for example, has as minimal contingency set $\{c_2f_2\}$, thus
545 its responsibility is 0.5. For the output tuple o_2 , all the tuples of the lineage
546 are counterfactual causes, thus their responsibility is 1.

547

548 4.5. Shapley value

549 To use the Shapley in the context of conjunctive queries for relational
550 databases, we use the definitions provided in [25]: given a query $q(\bar{x})$, a
551 database D , an input fact $f \in D$ (here seen as a player) and a tuple \bar{t} of same
552 arity as \bar{x} , the Shapley value of f in D intuitively represents the contribution
553 of f to the presence (or absence) of \bar{t} in the query result. Formally, the
554 Shapley value is defined as follows:

Definition 4.2. *Shapley value [?]*

Let the database instance I be partitioned into two sets of facts: a set I^x of

exogenous facts, and a set I^n of endogenous facts. Let Q be a Boolean query and $f \in I^n$ be an endogenous fact. The Shapley value of f in I for query Q is defined as:

$$\text{Shapley}(Q, I^n, I^x, f) = \sum_{B \subseteq I^n \setminus \{f\}} \frac{|B|! (|I^n| - |B| - 1)!}{|I^n|!} (Q(I^x \cup B \cup \{f\}) - Q(I^x \cup B))$$

555 The set I^n of endogenous facts can be thought as the set of tuples being
 556 taken into consideration, while I^x is the set of ignored tuples. The choice on
 557 I^n is usually application-dependent.

558 The sum in the definition of the Shapley value is performed on all pos-
 559 sible coalitions of fact B that do not contain the player f . Thus, the value
 560 $(Q(I^x \cup B \cup \{f\}) - Q(I^x \cup B))$ is the wealth brought by f when added to
 561 B . As we see, the Boolean query is used as wealth function v : its value is
 562 1 only when the set $I^x \cup B \cup \{f\}$ makes the query true, and the set $I^x \cup B$
 563 makes it false, i.e., when the addition of the fact f is determinant to make
 564 the Boolean query true. The value $|B|! (|I^n| - |B| - 1)!$ is the number of all
 565 the possible permutations over I^n where the facts in B come first, then f is
 566 added, and then all the remaining facts. Thus, the value $\frac{|B|! (|I^n| - |B| - 1)!}{|I^n|!}$ can
 567 be thought as a weight for the wealth brought by the addition of f to the
 568 coalition B .

569 To extend this definition to non-Boolean queries, we use the same straight-
 570 forward approach used in Deutch et al. [25]: the Shapley value of the fact f
 571 for the answer \bar{t} to $Q(\bar{x})$ is the value $\text{Shapley}(Q[\bar{x}/\bar{t}], I^n, I^x, f)$, where $Q[\bar{x}/\bar{t}]$
 572 is the Boolean query defined by $Q[\bar{x}/\bar{t}](I) = 1$ if and only if \bar{t} is in the output
 573 of $Q(\bar{x})$ on I , and 0 otherwise. *** DD: I added this paragraph down here**
 574 **hoping to make things clearer. If you think it fails to do so, feel**
 575 **free to delete it. *** In other words, the definition of $\text{Shapley}(Q, I^n, I^x, f)$ is
 576 extended to such queries $Q(\bar{x})$ with free variables by considering the Boolean
 577 query $Q[\bar{x}/\bar{t}]$ instead as value function. This query can be seen as a function
 578 that takes as input a set of facts and returns 1 if this set is a witness for \bar{t} ,
 579 and 0 otherwise.

580 As an example, consider table 7, that shows the Shapley values for the
 581 lineage's tuples of o_1 and o_2 , results of query Q1. Since the tuples of the
 582 lineage are the only one with a role in creating the output tuples, when
 583 computing the Shapley value we can use it as the set of endogenous facts.
 584 We note moreover that, to compute the Shapley value of an input tuple f in I

id	name	responsibility
o_1	Dopamine Receptors	$f_1 = \frac{7}{15}, c_2f_1 = \frac{2}{15}, c_2f_2 = \frac{2}{15}, c_1 = \frac{2}{15}, c_2 = \frac{2}{15}$
o_2	YANK Family	$f_4 = \frac{1}{3}, c_2f_4 = \frac{1}{3}, c_1 = \frac{1}{3}$

Table 7: Result of **Q1** over the database instance in Table 1 with the Shapley values of the tuples of the lineage. In this case D^n corresponds to the lineage.

is sufficient to compute and sum the values $\frac{|B|!(|I^n|-|B|-1)!}{|I^n|!}$ for all the possible sets B such that $B \cup \{f\}$ is a witness and B is not. Thus, suppose we want to compute the Shapley value of the tuple f_1 . Let us call \bar{Q}_{1,o_1} the Boolean query such that $\bar{Q}_{1,o_1}(I) = 1$ if and only if o_1 is in the output of **Q1**, and L_{o_1} the lineage of o_1 . Then the Shapley value of f_1 is given by:

$$\begin{aligned} \text{Shapley}(\bar{Q}_{1,o_1}, L, I \setminus L, f_1) &= \frac{2!2!}{5!} + \frac{2!2!}{5!} + \frac{3!}{5!} + \frac{3!}{5!} + \frac{3!}{5!} + \frac{3!}{5!} + \frac{4!}{5!} \\ &= \frac{7}{15} \end{aligned}$$

Where, for the first element of the sum the corresponding B is $\{c_2f_1, c_1\}$, for the second element it is $\{c_2f_2, c_2\}$, for the third it is $\{c_2f_1, c_2f_2, c_1\}$, for the fourth it is $\{c_2f_1, c_1, c_2\}$, for the fifth it is $\{c_2f_2, c_2, c_1\}$, for the sixth it is $\{c_2f_1, c_2f_2, c_2\}$, and for the seventh $\{c_2f_1, c_2f_2, c_1, c_2\}$. Every other possible coalition B would make the factor equal to 0.

Similarly, for tuple c_1 (and the other tuples of the lineage), the computation is:

$$\begin{aligned} \text{Shapley}(\bar{Q}_{1,o_1}, L, I \setminus L, c_1) &= \frac{2!2!}{5!} + \frac{3!}{5!} + \frac{3!}{5!} \\ &= \frac{2}{15} \end{aligned}$$

Similarly, it can be seen that for the tuples of o_2 's lineage the corresponding Shapley values are all equal to $1/3$, since they are all equally responsible for the generation of the output. As we can see, the sum of the Shapley values of all the tuples in an output tuple's lineage is always equal to 1 when using a Boolean query as wealth function.

5. Credit Distribution and Distribution Strategies

We now give formal definitions of data credit and Data Credit Distribution (DCD), and present three different Distribution Strategies (DSs) based on the forms of provenance discussed earlier: Lineage-based DS, Why-Provenance-based DS, How-Provenance-based DS, **responsibility-based DS**, and the **Shapley value-based DS**. We also show how these strategies distribute credit in the IUPHAR example discussed earlier.

607 5.1. Data Credit and Data Credit Distribution

608 Given a database instance I , a *recipient of credit* is a unit of information
 609 within I . In the case of relational databases, recipients may be (i) the whole
 610 database; (ii) a table; (iii) a tuple; or (iv) an attribute.

611 *Data credit* is a value $k \in \mathbb{R}_{>0}$. Every recipient in a database is annotated
 612 with a quantity of credit as a proxy for its importance. In this paper, we
 613 focus on *tuples* as recipients of credit.

614 Given a *distribution strategy* (DS), *Data Credit Distribution* (DCD) takes
 615 a database instance I , a quantity of credit k , and query Q over I , and it splits
 616 k among the recipients of credit in I .

617 In the following, we use the notation in Cheney et al. [17]: Given a
 618 database instance I , a *tuple location* (R, t) is a tuple t in relation R . With
 619 reference to the running example, $(\text{family}, \langle f_1, \text{Dopamine Receptors},$
 620 $\text{gpcr} \rangle)$ is the tuple location of the first tuple in the `family` relation. The set
 621 of all tuple locations in I is called *TupleLoc*. We use this to formally define
 622 DCD at the *tuple level*.

623 **Definition 5.1. Tuple Level Data Credit Distribution (DCD) [26]**

624 *Given a query Q over I and $k \in \mathbb{R}_{>0}$, DCD is defined by the function $f_{I,Q} :$
 625 $\text{TupleLoc} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ such that $f_{I,Q}(t, k) = h$ where $0 \leq h \leq k$ and
 626 $\sum_{t \in \text{TupleLoc}} f_{I,Q}(t, k) = k$. The function $f_{I,Q}$ is the distribution strategy (DS).*

627 As we can see, the DS is a function that annotates each tuple in the
 628 database with a real value, which is a fraction of the given quantity k . The
 629 only constraint is that the sum of the credit annotations on tuples must be
 630 k , i.e. that no credit is generated or destroyed during the distribution. Given
 631 I and Q , many different DSs may be defined as long as they sum up to k .

632 In what follows, we use information provided by data provenance to de-
 633 fine distribution functions. For simplicity, we assume that the credit k is
 634 distributed equally across the set of output tuples (i.e. the result of a query),
 635 and discuss how the credit of one output tuple o , k_o , is distributed across the
 636 instance I .

637 5.2. A Lineage-based Distribution Strategy

638 In the lineage-based distribution strategy, each tuple in the output of
 639 a query distributes credit equally to each input tuple that appears in its
 640 lineage. More formally:

Definition 5.2. *Lineage-based Distribution Strategy [26]*

Let I be a database instance, Q a query over I , $o \in Q(I)$ an output tuple and k_o the credit associated to o . Let L be the lineage of o and t be a tuple in I , then t receives credit equal to:

$$f_{I,Q}(t, k_o) = \begin{cases} 0 & \text{if } t \notin L \\ \frac{k_o}{|L|} & \text{if } t \in L \end{cases}$$

641 Note that lineage-based DS distributes credit only to input tuples that
642 have a role in creating o by the query Q , and that each receives an equal
643 share of credit. Thus, the more tuples in a lineage set, the less credit each
644 tuple receives.

645 As an example, consider the output tuples of Table 3, and assume that
646 each output tuple has credit $k_o = 1$. The lineage of the first tuple, o_1 , is
647 the set $\{f_1, c2f_1, c_1, c2f_2, c_2\}$. Therefore, each tuple in this set receives credit
648 $1/5$. The other tuples of the database receive zero credit. The lineage of the
649 second output tuple is $\{f_4, c2f_4, c_1\}$, therefore each of these tuples receives
650 credit $1/3$.

651 At the end of the process, tuples f_1 , $c2f_2$ and c_2 each receive credit $1/5$,
652 tuples f_4 and $c2f_4$ receive $1/3$, while tuple c_1 receives $8/15$. Note that if a
653 tuple appears in more than one lineage set, then it will accumulate credit
654 from the distribution associated with each one of these sets, implying that
655 it has a more significant role in the context Q , as is the case with c_1 in this
656 example.

657 Not all of the tuples in the lineage of an output tuple are necessary to be
658 present at the same time for the output tuple to appear in the query results.
659 For example, if the database only had the set of tuples $\{f_1, c2f_1, c_1\}$ or the set
660 $\{f_1, c2f_2, c_2\}$, the existence of o_1 would still be guaranteed. In other words,
661 while f_1 is always needed for o_1 to appear in the output, only one of the sets
662 of tuples $\{c2f_1, c_1\}$ and $\{c2f_2, c_2\}$ is required. One could therefore argue that
663 it would be more fair for f_1 to receive more credit than the other four tuples,
664 given its role in producing o_1 .

665 This highlights one limitation of the lineage-based DS: while able to find
666 all and only the relevant tuples of the output, it does not distinguish the
667 *importance* of tuples in the query computations. We therefore present four
668 other, more sophisticated, forms of distribution strategies based on why-
669 provenance, how-provenance, **responsibility**, and **Shapley value**.

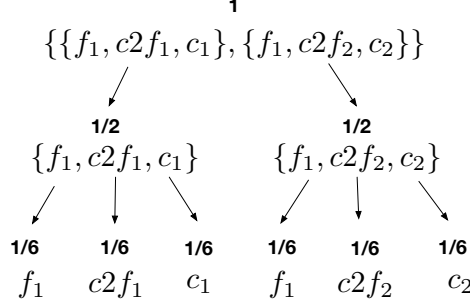


Figure 4: Distribution of credit using why-provenance-based DS for tuple o_1 .

5.3. A Why-Provenance-Based Distribution Strategy

The distribution strategy based on why-provenance first equally distributes the credit k_o among the witnesses of the witness basis for o , and then equally divides the credit of a witness among the tuples in the witness. Since a tuple may appear in more than one witness, it will receive more than one portion of credit from the same distribution. More formally:

Definition 5.3. Why-Provenance-based Distribution Strategy

Let I be a database instance, Q a query over I , $o \in Q(I)$ an output tuple and k_o the total credit associated to o . Let $\mathcal{W} = \text{Why}(Q, I, o)$ be the witness basis of o according to Q and I , and $W \in \mathcal{W}$ be a witness.

Then tuple t in I receives credit equal to:

$$f_{I,Q}(t, k_o) = \frac{k_o}{|\mathcal{W}|} \sum_{W \in \gamma(\mathcal{W}, t)} \frac{1}{|W|}$$

where γ is a function which returns all witnesses W in which t appears:

$$\gamma(\mathcal{W}, t) = \{W \in \mathcal{W} : t \in W\}$$

Figure 4 shows the distribution of credit with why-provenance-based DS for tuple o_1 . The credit is first equally divided between the two witnesses, so that both receive credit $1/2$. The credit is then further divided among the tuples in each witness. Since each witness has three tuples, each tuple in a witness receives $1/6$ of credit. At the end of the distribution, f_1 receives a total credit of $1/3$, and the other tuples receive $1/6$ each. This distribution better reflects the role of f_1 in the generation of o_1 since, as discussed earlier,

$$\begin{aligned}
\mathcal{H} &= \underbrace{3f_1 \cdot c2f_1 \cdot c_1}_{M_1} + \underbrace{f_1 \cdot c2f_2^3 \cdot c_2^3}_{M_2} \\
c(\mathcal{H}) &= 4 & e(M_2) &= 7 \\
mc(M_1) &= 3 & mc(M_2) &= 1 \\
te(c_2, M_2) &= 3 & \gamma(c_1, \mathcal{H}) &= \{M_1\} \\
\gamma(f_1, \mathcal{H}) &= \{M_1, M_2\}
\end{aligned}$$

Figure 5: Illustration of notation used to define the how-provenance based DS

687 it is the only mandatory tuple for o_1 to appear in the output; only one of the
688 two other pairs of tuples are necessary for o_1 to appear in the result.

689 This example illustrates that why-provenance can better reward input
690 tuples depending on their role. Tuples that appear in more than one witness
691 are rewarded more than others.

692 5.4. A How-Provenance Based Distribution Strategy

693 The how-provenance-based DS first distributes the credit to the mono-
694 mials of the polynomial accordingly to the weight represented by their co-
695 efficients, then to the tuples of each monomial accordingly to the weights
696 represented by their exponents.

697 To define the DS more formally, we introduce some notation and illustrate
698 it using the provenance polynomial \mathcal{H} shown in Figure 5. This notation is
699 also shown in Table 2 for easy reference.

700 We call c the function that, given a polynomial, returns the sum of its
701 coefficients; thus $c(\mathcal{H}) = 3 + 1 = 4$. We call e the function that, given a
702 monomial, returns the sum of its exponents, thus $e(M_2) = 1 + 3 + 3 = 7$.
703 mc is the function that takes as input a monomial and returns its coeffi-
704 cient; thus $mc(M_1) = 3$. te is a function that takes as input a tuple and a
705 monomial, and returns the exponent of the tuple in the monomial, if present;
706 thus $te(c_2, M_2) = 3$. Finally, γ takes as input a tuple and the whole poly-
707 nomial, and returns a set of monomials containing that tuple, if present in the
708 polynomial; thus $\gamma(f_1, \mathcal{H}) = \{M_1, M_2\}$, $\gamma(c_2, \mathcal{H}) = \{M_2\}$.

Definition 5.4. How-Provenance-Based Distribution Strategy

Let I be a database instance, Q a query over I , $o \in Q(I)$ an output tuple, \mathcal{H} be the provenance polynomial for o , and k_o the credit given to o . The credit

id	name
<i>oxs₁</i>	Dopamine Receptors

lineage	why-provenance	how-provenance
$\{f_1, c2f_1, c_1, c2f_2, c_2\}$	$\{\{f_1, c2f_1, c_1\}, \{f_1, c2f_2, c_2\}\}$	$f_1^2 c2f_1 c_1 + f_1^2 c2f_2 c_2$

Table 8: Result of query Q2 applied on the database of Table 1 and its different provenances. The reported numbers are the credit distributed through the process.

given to tuple t in I is:

$$f_{I,Q}(t, k_o) = \frac{k_o}{c(\mathcal{H})} \sum_{M \in \gamma(t, \mathcal{H})} mc(M) \frac{te(t, M)}{e(M)}$$

709 Going back to the example of Table 5, consider o_1 with provenance poly-
710 nomial $f_1 c2f_1 c_1 + f_1 c2f_2 c_2$. The how-provenance-based DS firstly divides
711 the credit between the two monomials. Since the coefficients of each mono-
712 mial are 1, the credit is split in half. If they were, for example, 1 and 2
713 respectively, 1/3 of the credit would go to the first monomial, and 2/3 to
714 the second. Since in our example each variable has exponent 1, the credit is
715 further divided equally among the three variables. Thus, at the end of the
716 computation, f_1 receives 1/3, and the other tuples receive 1/6.

717 In this specific example, the how-provenance-based DS has the same out-
718 come as the one based on why-provenance. We therefore consider another
719 query over GtoPdb, Q2, that asks for the families of type **gpcr** that have as
720 contributor a researcher located in the UK:

```

721 Q2: SELECT DISTINCT F.name
722 FROM family as F JOIN
723 (SELECT DISTINCT f.name AS name
724 FROM family AS f JOIN contributor2family AS c2f ON f.id = c2f.family_id
725 JOIN contributor AS c ON c2f.contributor_id = c.id
726 WHERE c.country = "UK") AS R ON F.name = R.name
727 WHERE F.type = "gpcr"

```

728 The result of Q2 is shown in Table 8, and consists of one tuple, *oxs₁*,
729 annotated with each of the three provenances. As can be seen, lineage and
730 why-provenance are identical to those of the tuple o_1 in the previous example.
731 The how-provenance, however, is different since tuple f_1 is used twice: first
732 in the join of the inner query, and second in the join of the outer query. This

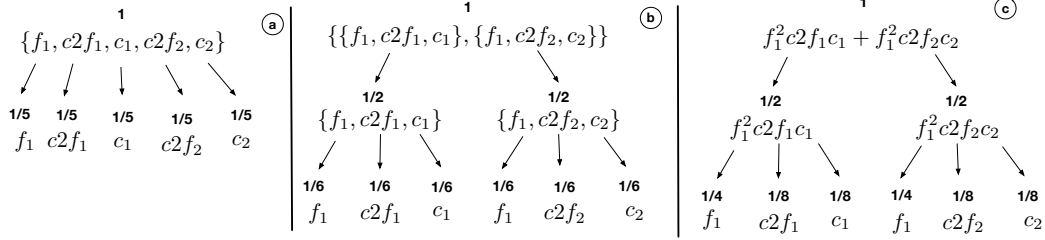


Figure 6: Comparison of different distributions strategies for tuple o_1 produced by query Q2.

information is lost in the first two forms of provenances since they are sets, but it is captured in how-provenance through the use of the operator ‘.’.

Figure 6 shows the differences between the three DS for the tuple o_1 of Table 8. Subfigure 8.a uses lineage, sub-figure 8.b uses why-provenance, and sub-figure 8.c uses how-provenance. The DS based on the provenance polynomial gives credit $1/2$ to f_1 , and $1/8$ to the other tuples. This is reasonable since Q2 relies on f_1 even more than Q1 does. The distribution based on how-provenance rewards f_1 more, showing that how-provenance is even more sensitive to the tuples’ role in a query than why-provenance. This is a direct consequence of the fact that, as proven in [32], how-provenance is more general than why-provenance and lineage, in the sense that it contains more information.

5.5. Responsibility-based Distribution Strategy

As described in Section 4.3, causality and responsibility is new information that is added to lineage. One possible option for defining a distribution strategy using responsibility is to simply assign the responsibility of each tuple in the lineage of an output tuple as its credit. In this way, responsibility is both a way to compute credit and to distribute it. Using the example of Table 6, in the case of output tuple o_1 , f_1 receives credit 1 and the other tuples receive credit 0.5.

However, we want a DS that is also a function of the input credit value k in order to be comparable with the other three strategies proposed so far. We define a new DS based on responsibility that is a function of the quantity of credit k_o that assigns to each tuple of the lineage a portion of this credit weighted by its normalized quantity of responsibility. This will give a bigger

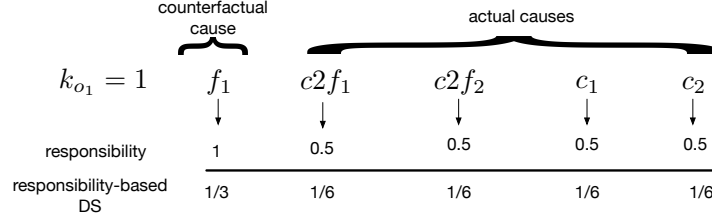


Figure 7: Example of distribution of credit using the responsibility-based DS, assuming $k_o = 1$.

portion of credit to tuples that are higher in the responsibility ranking. Formally:

761

Definition 5.5. *Responsibility-based Distribution Strategy*

Let I be a database instance, Q a query over I , $o \in Q(I)$ an output tuple, L the lineage of o , and k_o the credit given to o . The credit given to tuple t in I is:

$$f_{I,Q}(t, k_o) = k_o \frac{\rho_t}{\sum_{t' \in L} \rho_{t'}}$$

762 where ρ_j is the responsibility of tuple j as in Definition 4.1.

763 Note that only the tuples that belong to the lineage will receive a quantity
 764 of credit > 0 . Furthermore, the more important the tuple is, i.e., the higher
 765 its responsibility, the larger the quantity of credit received.

766 Figure 7 shows the responsibility and credit assigned to the tuples of the
 767 lineage of the output tuple o_1 of Table 6. Assuming that $k_{o_1} = 1$, f_1 receives
 768 credit $1/3$, while the others receive credit $1/6$. As we see, the DS in this
 769 case returns the same distribution as that obtained using why-provenance as
 770 shown in Figure 6. This is not always the case though, as we show in the
 771 example of Section 6.2.

772

773 *5.6. Shapley value-based Distribution Strategy*

774 Similarly to Responsibility, the Shapley value can be seen both as a
 775 method to generate and distribute credit. Moreover, it can be seen that,
 776 using the definition of Shapley value for Boolean queries given in Section 4.3,
 777 the sum of the Shapley values of all the tuples of the lineage L of an out-
 778 put tuple o is 1. Thus, the definition of a Shapley value-based distribution
 779 strategy is straightforward:

Definition 5.6. *Shapley Value-Based Distribution Strategy*

Let I be a database instance, Q a query over I , $o \in Q(I)$ an output tuple, L the lineage of o and k_o the credit given to o . The credit given to tuple t in I is:

$$f_{I,Q}(t, k_o) = k_o \cdot \text{Shapley}(\bar{Q}_o, L, I \setminus L, t)$$

780 Where \bar{Q}_o is the Boolean query such that $\bar{Q}_o(I) = 1$ if and only if o is in the
781 output of Q on I .

782 As shown in Table 7, tuple f_1 in o_1 's lineage takes credit 7/15 when
783 $k_{o_1} = 1$, while the other tuples of the lineage take credit 2/15. This DS still
784 rewards f_1 more than the other tuples, since it is more important than the
785 other tuples of the lineage. This DS thus behaves differently from all the
786 other four previous strategies. In particular, f_1 is rewarded more with this
787 DS than with the others.

788 In the case of o_2 there is only one witness set, thus this DS behaves like
789 all the others, distributing 1/3 of credit to each tuple in the lineage.

790 **6. Experimental Evaluation**

791 To understand the trade-offs between these Distribution Strategies (DSs),
792 we perform four sets of experiments using queries over target families pre-
793 sented on the GtoPdb website. The first set of experiments use real queries
794 extracted from citations to GtoPdb published in the British Journal of Phar-
795 macology. The second set uses synthetically produced provenance polyno-
796 mials, corresponding to more complex queries, in order to better highlight
797 the differences between the DSs. The third set of experiments considers
798 the accrual of credit over time by the three strategies, again using synthetic
799 queries. The fourth set of experiments shows how the DSs compare to tradi-
800 tional citations in giving credit to data curators using both real and synthetic
801 queries.

802 The source code for the experiments is written in Java and supported by
803 a PostgreSQL database. For purposes of reproducibility, the code we used
804 for our experiments and all queries are available here: [https://bitbucket.](https://bitbucket.org/dennis_dosso/credit_distribution_project)
805 [org/dennis_dosso/credit_distribution_project](https://bitbucket.org/dennis_dosso/credit_distribution_project).

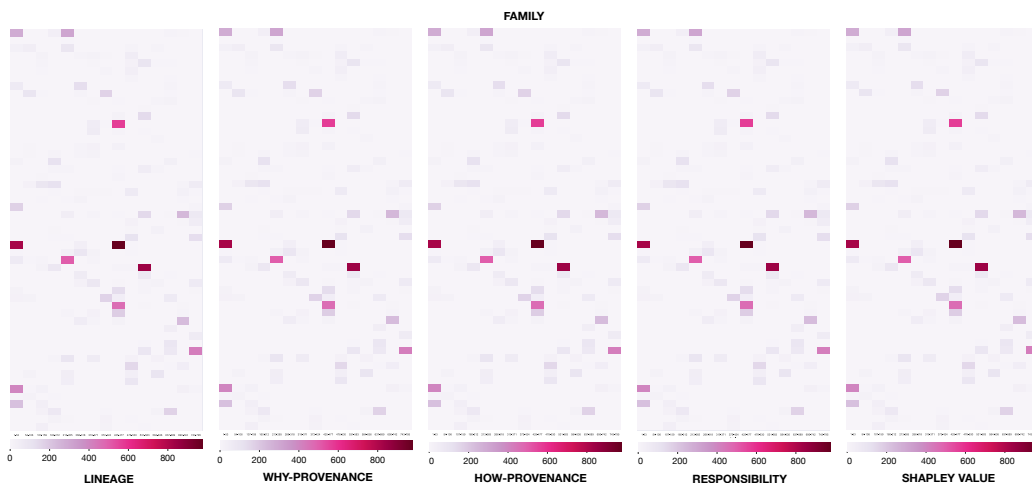


Figure 8: Comparison of four DS on the same table **family** using the distribution given by the queries retrieved from papers. Each cell is a tuple.

6.1. Real-world queries

Examples of real queries are drawn from papers published in the British Journal of Pharmacology (BJP)¹². Each time a paper in this journal cites a webpage from GtoPdb, it reports the URL of the page. From this URL, the query used to obtain the webpage data can be determined. We considered all 889 papers in BJCP citing the IUPHAR/BPS Guide to pharmacology [34] as of October 2020, and extracted all webpage URLs to GtoPdb contained within the paper.¹³

The queries that we inferred are those used to build target family webpages within GtoPdb. An example was given in Figure 3, where we show how the structure of the “Adenosine receptors” family can be mapped into queries over the underlying database. In GtoPdb, all target family pages share a similar structure; the only difference is that individual sections, such as “contributors” or “further readings”, may be missing. Therefore, the same queries can be used to build all of the target family pages by changing the family id used in the query (for example, in Figure 3, it is 3). Note that

¹²<https://bpspubs.onlinelibrary.wiley.com>

¹³The IUPHAR/BPS Guide is a journal that describes the structure and evolution of GtoPdb. At the time of writing, it had received more than 1200 citations on Google Scholar.

the queries are fairly simple SQL queries, and fall into a class called “select-project-join” or “SPJ” queries. A total of more than 12K different queries were built in this way. Without loss of generality, we give each tuple in the output of a query a credit of 1.

Results. Figure 8 shows the heat-maps obtained by the distribution of credit according to the **five** DS on one of the tables in the underlying database, **family**, which is often joined with other tables in the database to build the webpages. Each cell in a heat-map represents a tuple of the **family** table and the color indicates the amount of credit attributed to such tuple. It can be seen that the result of credit distribution over **family** is the same for all **five** strategies. The same result is also obtained with the other tables of the database used by the queries shown in Figure 3.

The reason why credit distribution is the same for all **four** strategies is that the queries are all simple SPJ queries, which use each table only once and do joins on key attributes. Under these conditions, each tuple of the output presents: (i) a how-provenance that is a single monomial with coefficient one and exponent one in each variable; (ii) a why-provenance with only one witness; (iii) a lineage that is the same of the witness in the basis, (iv) **all tuples are counterfactual causes when considering responsibility**, and (v) **they all have the same importance in the production of the output tuples according to their Shapley value**. Hence, for these queries, the **five** DSs behave in the same way: credit is uniformly distributed among the tuples of the lineage.

To illustrate this, consider one of the queries in Figure 3 which is used to build the output webpage:

```
Q3: SELECT c.first_names, c.surname
FROM contributor2family AS cf JOIN contributor AS c ON
cf.contributor_id = c.contributor_id
WHERE f.family_id = 3
```

Q3 returned 10 tuples from the version of GtoPdb used. The first tuple, <Bertil B., Fredholm>, has $c_{939} \cdot c_{2f_{496}}$ as its provenance polynomial. c_{939} represents the provenance token of a tuple in **contributor**, and $c_{2f_{496}}$ the provenance token of a tuple in table **contributor2family**. The why-provenance of this tuple is $\{\{c_{939}, c_{2f_{496}}\}\}$, its lineage is $\{c_{939}, c_{2f_{496}}\}$, **both these tuples are counterfactual causes and have a responsibility of one**. Therefore, the credit assigned to these tuples is 1/2 using all five DS. This happens for all the tuples in the output of each query of GtoPdb, thus making the distributions equivalent over all outputs.

859 However, this is not the case with more complex queries. As we showed
 860 in the previous section, when two or more tuples are merged as a result of a
 861 projection or union, the credit distributions will differ between the strategies.

862 6.2. Synthetic queries

863 To see what happens with more complex queries, we synthetically gener-
 864 ated provenance polynomials in which the coefficients and exponents could
 865 be greater than one, and picked them at random from a uniform distribution.
 866 The queries involve three GtoPdb tables: **family**, **contributor2family**, and
 867 **contributor**. The polynomials were generated as follows: first, the number
 868 of monomials was decided by randomly choosing a number between one and
 869 six. Then, we randomly chose a tuple from the **family** table, one from the
 870 **contributor2family** table and one from the **contributor** table; these are
 871 the variables of the monomial. We then chose a coefficient for the monomial
 872 (between one and three) and an exponent for each tuple (between one and
 873 four). For the next monomial, we decided if we wanted to keep the same
 874 tuple from the table **family** as first tuple of the new monomial. To do so, we
 875 generated a random float number between zero and one. If the number was
 876 above 0.2, we changed the family tuple.

877 An example can be found in Figure 9, which shows a sample synthetic
 878 provenance polynomial (the how-provenance), the corresponding why-provenance,
 879 lineage, the causality of the tuples of the lineage, together with their respon-
 880 sibility, and, finally, the Shapley values of the lineage tuples. The resulting
 881 credit distribution for each DS is also shown (except for the Shapley values,
 882 that coincide with the distribution of credit).

883 As an example of how the distribution strategies behave with these syn-
 884 thetic queries, consider tuple f_5 in Figure 9. This tuple receives the highest
 885 quantity of credit using responsibility-based distribution and less credit us-
 886 ing, in order, lineage, the Shapley value, why- and how-provenance. On the
 887 other hand, tuple f_1 is rewarded more by the Shapley value, then, in order,
 888 by why-provenance, how-provenance, responsibility, and finally lineage. This
 889 difference is explained considering the different role of the tuples in the gen-
 890 eration of the output and the characteristics of the distributions. Generally
 891 speaking, the more complex the distribution (e.g., the how-provenance), the
 892 more credit is given to tuples that are more frequently used or more crit-
 893 ical in the production of the output. Depending on the situation, i.e. on
 894 the syntax of the query, the distributions may differ among them. Respon-
 895 sibility creates a ranking among lineage’s tuples describing the importance

How-provenance: $3f_1^3c_2f_1^2c_1^2 + 2f_1c_2f_2^3c_2^3 + 4f_5c_2f_{17}^4c_{18}^3$

Credit distribution:

$$f_1 = \frac{59}{315}, f_5 = \frac{1}{18}, c_2f_1 = \frac{2}{21}, c_2f_2 = \frac{2}{15}, c_2f_{17} = \frac{2}{9}, c_1 = \frac{2}{21}, c_2 = \frac{2}{15}, c_{18} = \frac{1}{6}$$

Why-provenance: $\{\{f_1, c_2f_1, c_1\}, \{f_1, c_2f_2, c_2\}, \{f_5, c_2f_{17}, c_{18}\}\}$

Credit distribution:

$$f_1 = \frac{2}{9}, f_5 = \frac{1}{9}, c_2f_1 = \frac{1}{9}, c_2f_2 = \frac{1}{9}, c_2f_{17} = \frac{1}{9}, c_1 = \frac{1}{9}, c_2 = \frac{1}{9}, c_{18} = \frac{1}{9}$$

Lineage: $\{f_1, f_5, c_2f_1, c_2f_2, c_2f_{17}, c_1, c_2, c_{18}\}$

Credit distribution:

$$f_1 = \frac{1}{8}, f_5 = \frac{1}{8}, c_2f_1 = \frac{1}{8}, c_2f_2 = \frac{1}{8}, c_2f_{17} = \frac{1}{8}, c_1 = \frac{1}{8}, c_2 = \frac{1}{8}, c_{18} = \frac{1}{8}$$

Causality: counterfactual causes: \emptyset ,

actual causes: $\{f_1, f_5, c_2f_1, c_2f_2, c_2f_{17}, c_1, c_2, c_{18}\}$

Responsibility:

$$f_1 = \frac{1}{2}, f_5 = \frac{1}{2}, c_2f_1 = \frac{1}{3}, c_2f_2 = \frac{1}{3}, c_2f_{17} = \frac{1}{2}, c_1 = \frac{1}{3}, c_2 = \frac{1}{3}, c_{18} = \frac{1}{2}$$

Credit distribution:

$$f_1 = \frac{3}{20}, f_5 = \frac{3}{20}, c_2f_1 = \frac{1}{10}, c_2f_2 = \frac{1}{10}, c_2f_{17} = \frac{3}{20}, c_1 = \frac{1}{10}, c_2 = \frac{1}{10}, c_{18} = \frac{3}{20}$$

Shapley value:

$$f_1 = 0.258\bar{3}, f_5 = \frac{1}{8}, c_2f_1 = 0.091\bar{6}, c_2f_2 = 0.091\bar{6}, c_2f_{17} = \frac{1}{8}, c_1 = 0.091\bar{6}, c_2 = 0.091\bar{6}, c_{18} = \frac{1}{8}$$

Figure 9: Sample synthetic provenance polynomial (how-provenance) and corresponding why-provenance, lineage, responsibility, and Shapley values, together with the corresponding credit distributions. In the case of the Shapley value, the value is equivalent to the quantity of credit being distributed (assuming that the input credit is equal to 1).

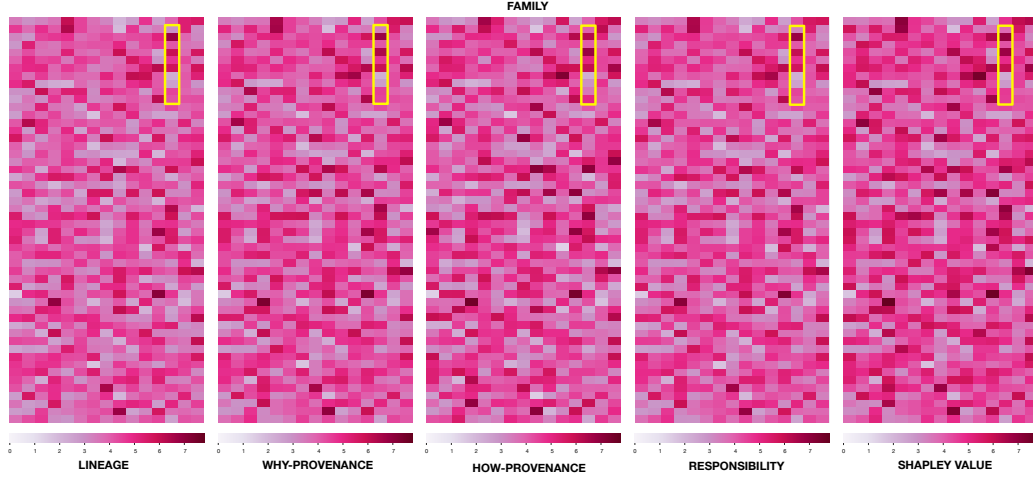


Figure 10: Comparison of three DS on the same table **family** after the distribution computed using 10K synthetic and randomly generated provenance polynomials. The tuples in the blue rectangles are used as example in the discussion connected to Figure 11.

of their role in generating the output. As such, the responsibility-based DS gives more credit to f_1, f_5, c_2f_17 and c_18 due to their higher responsibility values. “Importance” is connected to their corresponding minimal contingency sets. For example, f_1 has a minimal contingency set (one of the many) $\{f_5\}$, with cardinality 1. On the other hand, c_1 has, as minimal contingency set (one of the many) $\{f_5, c_2\}$, with cardinality two. This means that c_1 is the “least important” amongst the tuples with minimal contingency sets of lower cardinality, and this is reflected in the different quantities of credit being distributed.

The Shapley value behaves similarly, but it rewards the most tuple f_1 , then come in the ranking tuples f_5, c_2f_17 and c_18 , and the other tuples of the lineage with the lowest value. Although both Responsibility and the Shapley value create a ranking of the tuples based on their role in the generation of the output, the corresponding functions behave differently due to the syntax of the query, highlighting different aspects of the term “important”.

Despite being synthetic, these provenance polynomials are realistic: they can be obtained by any nested query with join and union operations that use the same tuple multiple times (in which case the exponents are larger than one), and the same combination of operations more than once (in which case the coefficients of monomials are larger than one).

Table 9: Quantities of credit given using the 5 DSs on the first five tuples of table `family` (tuples ordered by the `family_id` attribute of the table).

lineage	why	how	responsibility	Shapley
3.3603537	3.416667	3.5928571	3.3611114	3.425758
4.4893217	5.111111	4.8620114	5.1752524	5.788059
3.1333337	3.7888894	2.9106944	3.5000005	4.200001
2.7972224	3.1111116	3.5601408	3.0055559	3.3305562
3.4670746	3.8944445	3.7216337	3.8992426	4.31758

916 *Results.* The results of credit distribution on the `family` table using 10K
 917 randomly generated synthetic provenance polynomials are shown in Figure
 918 10. We set the maximum value in the heat maps to the highest value reached
 919 by a tuple in all `five` distributions (i.e., 7.7, with the Shapley value-based DS).

920 *** DD: is the table described in this paragraph below helpful? If**
 921 **not, please delete. *** As can be seen, the five strategies generate different
 922 credit distributions, indicated by the varying hues. We reported in Table
 923 9 the values of credit assigned to the first five tuples of the table to show
 924 how these values actually differ between the five strategies. As can be seen,
 925 the strategies in these cases all behave differently. It is not even possible
 926 to identify a strategy that consistently rewards tuples more than the others,
 927 since this changes depending on the cases, reflecting the syntaxes of the
 928 polynomials being used.

929 However, there is a certain amount of consistency between the strategies
 930 in that tuples which are highly rewarded by one strategy are also highly
 931 rewarded by the others. This shows that the four DSs consistently reward
 932 certain tuples more than others.

933 Note that lineage-based DS gives the least credit to tuples in the `family`
 934 table, indicated by an overall lighter hue. This is because the DS distributes
 935 credit equally to all tuples appearing in the lineage. Since these queries also
 936 use two other tables, credit is distributed to tuples in those tables.

937 Moving to why-provenance-based DS, we see that more credit is given to
 938 tuples in the `family` table than with the previous strategy. This is because
 939 the DS considers the different ways that a tuple is used, e.g. in joins with
 940 other tuples. If the same tuple is present in more than one witness, it will
 941 draw more credit and take it from other tuples in the witness basis. In this
 942 case, tuples in `family` drew more credit, taking it from tuples in the other
 943 two tables, due to the role that `family` tuples played in the queries that were

944 executed.

945 Consider the how-provenance-based DS heat-map. As with why-provenance,
946 more credit is typically given to tuples in **family** compared to lineage-based
947 DS, since it recognizes the role of these tuples in the queries, and the over-
948 all hue is deeper. The two distributions appear similar, although on closer
949 inspection, slight differences can be seen. This is because how-provenance
950 also considers the frequency with which tuples are used, not only the ways in
951 which they are used. Therefore, although the overall distribution is similar,
952 there are small differences due to the presence of exponents and coefficients
953 in the provenance polynomials, influencing the distribution of credit.

954 The responsibility-based distribution strategy has a distribution that is
955 also quite similar to the one provided by why-provenance (which is also visible
956 from Table 9, where the values of the two distributions are different but
957 very close). It is often the case, for example when the witnesses of the
958 why provenance share one common tuple, that the two distributions behave
959 similarly.

960 Finally, the heat-map reporting the distribution produced by the Shapley
961 value is the one that, at a closer inspection, shows the biggest differences.
962 Although the tuples that receive the biggest quantities of credit are the same,
963 the hue of this tuple is different. The Shapley value in certain circumstances
964 differs greatly from the other DSs, thus showing its ability to weight differ-
965 ently the roles of the tuples.

966 We note that the lineage-based DS gives an average credit of 3.92 to each
967 tuple in the table, while the DS based on why-provenance assigns 4.19, how-
968 provenance 4.18, the one based on responsibility 4.13, and the one based on
969 the Shapley value 4.40. Moreover, lineage distributed a total of about 3121
970 units of credit to the **family** table, why-provenance 3333, how-provenance
971 3331, while responsibility assigned 3290, and the Shapley value 3505. Thus,
972 the Shapley value is the method that accumulates the highest quantity of
973 credit in this table.

974 To better understand the differences between DSs, in the next subsection
975 we consider the accrual of credit over time. In doing so, we will focus on the
976 ten tuples shown within the large yellow rectangles in Figure 11. Each small
977 rectangle within a large yellow rectangle is a tuple, and we number them
978 from 1 (top) to 10 (bottom). These ten tuples were cherry-picked because
979 they allow us to see the evolution of the distribution of credit through time.
980 There are other tuple sets that could have been selected driving us to the
981 same considerations.

6.3. Credit accrual over time

Since credit accrues over time, we simulate the passage of time by varying the number of queries executed, and look at the “snapshots” of credit for each of the strategies using synthetic queries. The results are shown in Figure 11.

In this figure, four groups of heat-maps are shown. Each group represents a “snapshot” taken after 1K, 2K, 5K and 10K provenance polynomials have been considered for credit distribution. The ten tuples in each heat-map are those highlighted in the yellow boxes of Figure 10 from the family table.

The polynomials used are the same as the experiment of the previous section. The range of credit in each map goes from 0 (no credit) to 7 (the maximum quantity of credit reached – using how-provenance – on one of the tuples of the considered window at the “snapshot” with 10K queries). The color hue of the legend, as can be seen, still ranges from 0 to 7.7.

By the end of 1K queries, credit differentials between tuples as well as between strategies can be seen. For example, tuple 3 is usually rewarded the most credit by all five strategies. Moreover, it can be seen that tuple 1 receives a higher quantity of credit when how-provenance is adopted, showing how this form of provenance behaves differently from the others in this context. Moving to 2K queries, it is possible to see that tuple 3 and 7 are still the most rewarded by the strategies.

By the end of 5K queries, tuple 7 emerges with the highest value of credit with all five DSs, a position which is strengthened with 10K queries. Moreover, with the passing of time, tuple 3 ceases to be one of the most rewarded ones and new tuples, such as 6 and 9, emerge as being particularly rewarded at 5K, while at 10K tuples 6 and 7 are the most rewarded from the distributions. This is because tuple 7 is used several times within queries being executed, which is rewarded strongly by why- and how-provenance. We also note that the responsibility-based distribution confirms its trend of being similar to why-provenance, although not identical. This is more evident at step 10K, where tuple 7 is slightly less rewarded using responsibility (6.12) with respect to why-provenance (6.24). The responsibility that rewards the more tuple 7 is the one based on how-provenance (credit 7.03), followed by the Shapley value (credit 6.64). This is due to the fact that tuple 7 had, among some of the polynomials being used for the experiments, a high responsibility but it did not appear in all witnesses. This changed slightly the distribution.

While the relative value of credit “positions” of tuples within a DS strategy depends on what queries are being executed, the important thing to

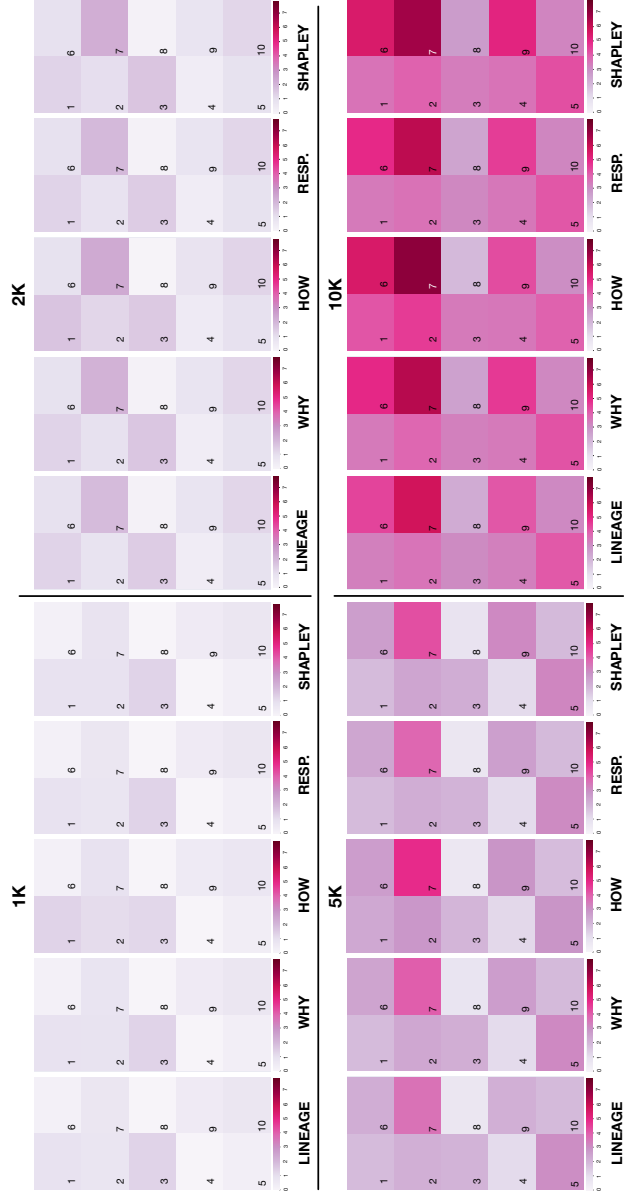


Figure 11: Comparison of the distribution of credit performed by the **five** DSs on a subset of 10 tuples taken from the **family** table, simulating the passing of time. The number at the top of each group of heat-maps represents the number of polynomials whose credit has been distributed.

notice is the difference between the DSs over time: overall, lineage gives less credit to tuples in the **family** table than the other strategies since credit is shared with tuples in other tables. The other strategies recognize the more important role being played by the **family** tuples than those in the other tables. The differences between why- and responsibility-based DS are, for the most times, negligible. The differences between the why- and how-provenance-based DSs are also relatively minor in most cases. However, there are certain situations in which the role of a tuple is particularly critical in a query, and in this case the difference in the value of credit assigned is notably higher for how-provenance and the Shapley value, as we saw with tuple 7 in the example of Figure 11.

To sum up, the DS based on lineage is sufficient to highlight which tuples in the database are used by a query, and distributes credit equally to these tuples. The resulting distribution rewards tuples that are used by more queries, but does not reward how many times tuples are used in the same query. However, a DS based on why-provenance, responsibility, Shapley value or how-provenance may be better if the queries are complex, since they reward more tuples that have a critical role in generating the output. In particular, these four DSs may be useful for finding “hotspots” in the database based on the role of tuples, with the how-provenance-based and Shapley value-based DSs being preferable if a higher sensitivity to the role of a tuple in queries is required.

6.4. Credit vs Citations

In the last set of experiments, we compare traditional citations to the proposed credit distribution strategies to see the difference in reward for data authors and curators. Using both real-world and synthetic queries, we distribute credit to the authors responsible for the data under the different strategies. Our results show that credit rewards authors of data that is cited fewer times, but that has a higher impact on the query results.

To do so, we need to identify a set of authors and queries that cite data curated by them. Considering GtoPdb, each target family page has a list of curators, representing the people who are co-creators and curators of the data comprising the page. This list can be obtained using the last query shown in Figure 3. Each time a target family page is cited, we assign one *citation* to each author associated with the page. The authors also receive *credit* in the amount assigned to the data used by the query to construct the webpage, equally divided between the authors of the webpage.

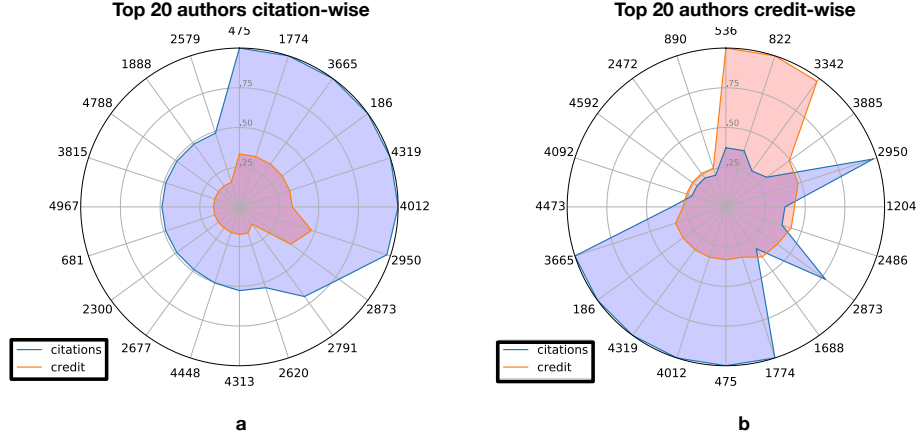


Figure 12: Radars presenting the top 20 authors citation-wise and credit wise, together with their (normalized between 0 and 1) values of citations and credit.

Results: Real-world queries. As described in Section 6.1, we consider real-world queries taken from papers published in the BJP which reference web-pages in GtoPdb. Since for these queries there is no difference in the distribution of credit between the DSs, only one value for credit is used.

The results are shown in the radar plots of Figure 12, in which each number on the outer circle (e.g. 475, 1774 and 3665) represents an author (id) and the blue (red) line represents the normalized value of credit generated by citations (credit), respectively. The first radar plot, Figure 12.a, shows the top 20 authors in terms of *citations*, ordered in a clockwise direction, whereas Figure 12.b orders the authors based on *credit*. Comparing the author ids used in the outer circles of these two plots, it can immediately be seen that the “top authors” are very different using these two metrics, although there is some overlap (for example, authors 1774, 475, and 4012).

Diving a bit deeper to focus on the red and blue areas in each of the plots reveals that there is a significance difference between citations and credit: The top 20 authors in terms of citations do not have the highest values of credit (Figure 12.a). Conversely, the authors with the highest values of credit do not necessarily have a large number of citations (Figure 12.b). For example, author 536 has the highest value of credit, but is not even in the top 20 authors in terms of citations. This means that authors like 536, 822, and 3342 in Figure 12.b receive much more credit from their relatively few citations than authors like 475, who receives the largest number of citations.

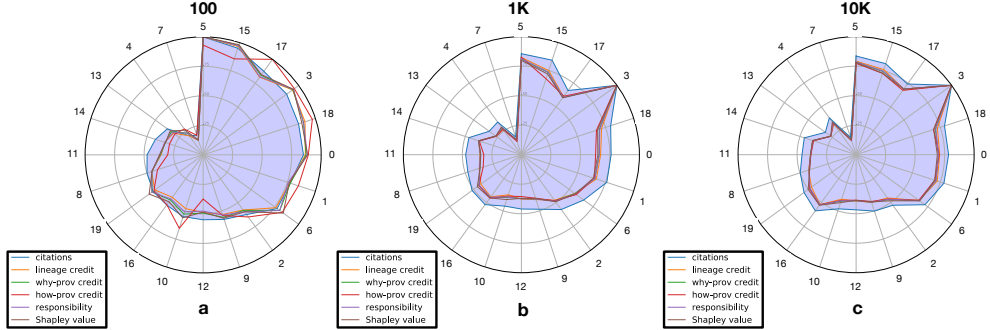


Figure 13: Radars presenting the 20 synthetic authors with corresponding citation and quantities of credit distributed through the 4 DSs (all values normalized between 0 and 1) through different numbers of polynomials (respectively, 100, 1K and 10K). The order is the one defined by figure a, i.e. descending order of citations obtained from 100 polynomials.

That is, the data underlying certain webpages is more “valuable” in terms of credit than a citation to the webpage.

The reason for the difference between citation and credit is partly due to the experimental setup: each output tuple carries a credit of 1, and there can be many tuples used to generate a webpage. Thus a webpage that is created from more tuples will have a higher credit value than one created from fewer tuples. Furthermore, authors who collaborated with fewer people will receive a biggest share of the equally divided credit. However, all authors will receive a citation of one.

Credit distribution therefore rewards authors differently than traditional citations: an author who has curated larger quantities of cited data and collaborated with fewer co-authors, will receive larger quantities of credit. Thus, credit rewards them for their larger contribution to the database.

Results: Synthetic queries. We used the same synthetic polynomials described in Section 6.2, and we distributed credit with the first 100, 1K, and 10K of them. Since these polynomials are created by randomly selecting tuples from three tables, they usually correspond to a set of data curated by authors who, in reality, did not collaborate. To make the size of the author set more realistic, we therefore created 20 synthetic authors, and randomly assigned one author to blocks of consecutive tuples in the database, with the size of each block varying between 10 and 40, to simulate different quantities of work performed by an author. Every time an author appears as curator of one or more tuples used in a polynomial, we assign them one citation. They

1101 also receive four kinds of credit, each one using a different DS.

1102 Figure 13 shows three radar plots, one for the results obtained with 100
1103 polynomials, one with 1K polynomials, one with 10K polynomials. Each
1104 plot shows the top 20 authors in terms of citations (hence the authors and
1105 clockwise ordering is the same in each of the plots), and additionally shows
1106 the the normalized values of citation (blue line), lineage-based credit (yellow
1107 line), why-provenance-based credit (green line), how-provenance-based credit
1108 (red line), responsibility-based credit (violet line), and the Shapley value-
1109 based credit (brown line).

1110 As can be seen, given the synthetic nature of the queries, the correlation
1111 between the number of citations and the quantity of credit assigned to the
1112 authors appears to be a much stronger than with the real-world queries of
1113 Figure 12. In fact, for Figure 13.a the linear correlation between the citation
1114 number and all four types of credit is always above 0.94 with p values in the
1115 order of $3e-8$. The credit distributed via lineage is closest to the number of
1116 citations (a linear correlation of 0.99, p value of $2e-16$ in Figure 13.a), while
1117 the other three types of credit behave slightly differently (a linear correlation
1118 of around 0.95 or above in all other four cases in Figure 13.a). Similar
1119 observations can be made for Figure 13.b and 13.c.

1120 What these figures show is that, in certain cases, authors who do not have
1121 a large number of citations receive more credit than others, as for example au-
1122 thors 17, 18 and 10 in Figure 13.a, and especially when credit is distributed
1123 using how-provenance. This again shows how credit gives a different per-
1124 spective on the role of data and authors by going beyond the limitations of
1125 traditional citations.

1126 It is worth noting that, when scaling up to 1K and 10K polynomials, the
1127 credit distributions become almost identical (the linear correlation for the
1128 values of Figure 13.c is more than 0.99 with a p-value of $1.32e-32$). This is
1129 consistent with what we observed in Figure 10.

1130 7. Discussion

1131 Before concluding, we discuss some design decisions: the focus on Credit
1132 Distribution (as opposed to Credit Generation), and the choice of Distribu-
1133 tion Strategies.

1134 7.1. Credit Generation

1135 In this paper we focused on Credit Distribution, the problem of distribut-
1136 ing credit generated by a citation to the parts of the database referenced by
1137 the query. A different problem is Credit Generation, the task of generating
1138 credit which is then distributed. Credit Generation presents a series of issues
1139 which are shared by traditional citation practices. For instance, defining the
1140 quantity of credit to be generated for a given citation is still an open prob-
1141 lem. Different types of citations may generate different quantities of credit.
1142 Data cited as previous work or as useful for previous work may generate less
1143 credit than other data extensively used to produce the results presented in
1144 a paper. The computation of credit could be done manually (although we
1145 must consider the complexity of the task, human biases and the resources
1146 required to carry it out) or automatically, but it must be based on a shared
1147 definition of impact which is still not agreed upon for data or for traditional
1148 citation. For this reason, we used a uniform credit assignment.

1149 There is also the problem of *transitive credit distribution*, i.e., how to
1150 transitively propagate credit from one cited unit to another unit that was
1151 used to produce the one being cited. For this, a graph of cited units that
1152 propagate credit between the units depending on influence could be used.
1153 How to propagate credit is an open and non-trivial problem that needs to
1154 consider the importance and impact of a citation in a work, be it a paper or
1155 data, and how to eventually compute the quantity of credit to be propagated.

1156 Finally, in our experiments we assumed that the credit carried by an
1157 output tuple is one. Thus, each tuple in the output has equal importance.
1158 As described above, this assumption may be revised and different credit to
1159 different output tuples could be assigned. Note that from the distribution
1160 model viewpoint no change is required since the DCD is defined for a generic
1161 value k .

1162 7.2. Choice of Distribution Strategies

1163 In this paper we presented four different DSs, so the natural question is
1164 which one to use. This depends on the task at hand. When we want to
1165 highlight the tuples being used in the database by a workload, the lineage-
1166 based DS may be sufficient. When we also want to know the relative impact
1167 of tuples in the context of the query, the other DSs should be used since they
1168 give a better understanding of the importance of data.

1169 In the real-world based experiments, the four DSs behaved the same,
1170 which was due to the specific nature of the data and the queries being used.

1171 However, the why-provenance of a query will differ from the lineage of the
1172 same query whenever the output tuples can be computed in more than one
1173 way by the query, i.e., if there is more than one witness. This is usually true
1174 when join and projection operators are used in the query.

1175 To address the question of what types of queries are likely to extract
1176 cited data, we turn to the results of published studies on the characteristics
1177 of query workloads and the complexity of their queries [38, 53, 58]. These
1178 studies show that operations such as inner-/outer-joins and projections occur
1179 in a significant number of queries. Therefore why- and how-provenances may
1180 become quite complex in certain cases and provide a distribution of credit
1181 that is significantly different from the one obtained with lineage.

1182 *** Is there more to say here? What are the general queries**
1183 **for which responsibility is hard to compute, and can the various**
1184 **provenances handle them at all? I know that provenance semi-rings**
1185 **has been extended to SPJU and aggregate queries, so imagine this**
1186 **means the others can be extended since it is a general framework. ***
1187 From a complexity standpoint, all four DS are similar since we focused on SPJ
1188 queries. However, responsibility is hard to compute for general queries. In
1189 terms of implementation, lineage is the simplest to compute since it only cares
1190 about a tuple being used, while the other provenances also need additional
1191 information to be taken into consideration.

1192 8. Conclusions and Future Work

1193 This paper defines four new distribution strategies based on why-provenance,
1194 how-provenance, responsibility, and the Shapley Value, and it compares them
1195 against the lineage-based distribution strategy defined in [26]. The first, why-
1196 provenance-based DS, uses the concept of a witness, and gives more credit
1197 to tuples that appear in more than one witness. In this way, tuples that are
1198 more important to the query and are used in different ways are rewarded
1199 more. The second, how-provenance-based DS, considers the frequency with
1200 which a tuple or combination of tuples is used in the query through the
1201 information contained in a provenance polynomial. In this case, the how-
1202 provenance-based DS is more sensitive than the why-provenance-based DS
1203 to the role and importance of tuples. The third DS exploits the notion of
1204 responsibility, a real value which ranks the lineage tuples based on their de-
1205 gree of causality in generating the output. The responsibility-based DS was
1206 shown to behave similarly to the why-provenance based DS. The fourth DS

uses the Shapley value function, used to rank the facts of the database, seen as players, in producing the required result. To do so, the wealth function in the Shapley value’s definition was adapted for general free-variable queries on the database.

To show the differences between the five DSs, we performed extensive experiments based on GtoPdb, a curated scientific relational database, using both real and synthetic queries. In the first set of experiments, we used select-project-join (SPJ) queries extracted from citations to webpages in GtoPdb found in papers published in the British Journal of Pharmacology. Using these “real” queries, we distributed credit to tuples in different tables of the database, highlighting tuples that were more frequently used. We showed that, with these queries, the four strategies produce the same distribution. This is because the SPJ queries were fairly simple, and did not use self-joins. Therefore the formulas underlying the different DSs had the same output.

In the second set of experiments, we synthetically produced more complex provenance polynomials, corresponding to more complex queries, that resulted in exponents and coefficients in the provenance polynomials that were greater than (or equal to) 1. These experiments highlighted the differences between the four DSs. While the DS based on lineage rewards all the tuples used by a query equally, the strategies based on why-provenance and responsibility give more credit to tuples that are more critical to the query. In particular, why-provenance considers the different ways in which a tuple is used in a query, while responsibility considers the relative importance of a tuple in the generation of the output. The DS based on the Shapley value similarly rewards the tuples based on their participation. The more impactful the role of a tuple, the higher its reward in credit. This distribution proved to be different from the previous two and to reward even more tuples that are used in more than one witness. How-provenance is even more sensitive to the tuple’s role: it also considers the frequency with which a tuple or a set of tuples is used.

In the third set of experiments, we showed how the differences between the DS are compounded over time, i.e. when more and more queries are processed by the system.

In the fourth set of experiments we compared traditional citations to authors to the credit accrued to them via the DSs. We showed how, in both real-world and synthetic scenarios, credit rewards authors who contribute/curate data that has the highest impact, and therefore receives the biggest quantity of credit, and not necessarily the data with the highest ci-

1245 tation count. In this sense, credit appears to be an useful new measure to
1246 discover data and their corresponding curators that have a high impact in
1247 the research world, even when they are cited few times or do not appear at
1248 all in the data that are cited (i.e. the case of data used to build the output
1249 of a query but that is not visualized in the output itself).

1250 In future work, we plan to explore different strategies to generate and
1251 distribute credit. In this paper we assumed that each output tuple carries
1252 credit 1. In more sophisticated scenarios we can employ different strategies
1253 to compute credit, that reflect the importance of cited data. Other, more
1254 sophisticated, strategies could also be used to decide how credit is distributed
1255 between the authors, beyond the uniform distribution used here, in a way
1256 to reflect the work performed by them on the cited data. There are also a
1257 number of other intriguing applications for credit over relational databases.
1258 One such application is *data pricing*, which gives a price to a query submitted
1259 by a user who wants to buy the produced information. Currently, a common
1260 strategy used for data pricing is based on query rewriting: A database stores a
1261 set of views with their price. When a new query arrives, the system rewrites
1262 it using the stored views to obtain a query price, a process that can be
1263 computationally expensive. We plan to distribute credit through carefully
1264 planned and representative queries, and use credit information to define a
1265 new, faster, and potentially more flexible pricing function.

1266 Another application is *data reduction* [47], which addresses the problem of
1267 reducing the vast – and rapidly expanding – amount of data that is being pro-
1268 duced. Data credit can help address this problem by identifying “hotspots”
1269 and “coldspots” of data. A hotspot is data in a database (e.g. a tuple) with
1270 a high quantity of credit, which is therefore valuable for the set of queries
1271 that execute frequently over the data and distribute the credit. A coldspot is
1272 data with a low quantity of credit which can therefore be considered as less
1273 important, and could be deleted, summarized, or moved to cheaper and/or
1274 less efficient memory.

1275 Acknowledgement

1276 The work was partially supported by the ExaMode project, as part of the
1277 European Union H2020 program under Grant Agreement no. 825292.

References

- [1] Abadi, D., Ailamaki, A., Andersen, D., Bailis, P., Balazinska, M., Bernstein, P., Boncz, P., Chaudhuri, S., Cheung, A., Doan, A., Dong, L., Franklin, M. J., Freire, J., Halevy, A., Hellerstein, J. M., Idreos, S., Kossmann, D., Kraska, T., Krishnamurthy, S., Markl, V., Melnik, S., Milo, T., Mohan, C., Neumann, T., Chin Ooi, B., Ozcan, F., Patel, J., Pavlo, A., Popa, R., Ramakrishnan, R., Ré, C., Stonebraker, M., and Suciu, D. (2020). The seattle report on database research. *SIGMOD Rec.*, 48(4):44–53.
- [2] Alawini, A., Davidson, S. B., Hu, W., and Wu, Y. (2017). Automating data citation in citedb. *PVLDB*, 10(12):1881–1884.
- [3] Alawini, A., Davidson, S. B., Silvello, G., Tannen, V., and Wu, Y. (2018). Data citation: A new provenance challenge. *IEEE Data Eng. Bull.*, 41(1):27–38.
- [4] Altman, M., Borgman, C. L., Crosas, M., and Martone, M. (2015). An Introduction to the Joint Principles for Data Citation. *Bulletin of the Association for Information Science and Technology*, 41(3):43–45.
- [5] Baggerly, K. (2010). Disclose all data in publications. *Nature*, 467(7314):401–401.
- [6] Bechhofer, S., Buchan, I. E., De Roure, D., Missier, P., Ainsworth, J. D., Bhagat, J., Couch, P. A., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D. T., Owen, S., Newman, D. R., Sufi, S., and Goble, C. A. (2013). Why linked data is not enough for scientists. *Future Gener. Comput. Syst.*, 29(2):599–611.
- [7] Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE*, 9(3):e92590.
- [8] Bertin-Mahieux, T., Ellis, D., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 591–596.
- [9] Borgman, C. L. (2016). Data Citation as a Bibliometric Oxymoron. In Sugimoto, C. R., editor, *Theories of Informetrics and Scholarly Communication*, pages 93–116. De Gruyter Mouton.

- [10] Buneman, P. (2006). How to cite curated databases and how to make them citable. In *18th International Conference on Scientific and Statistical Database Management, SSDBM*, pages 195–203. IEEE Computer Society.
- [11] Buneman, P., Christie, G., Davies, J. A., Dimitrellou, R., Harding, S. D., Pawson, A. J., Sharman, J. L., and Wu, Y. (2020). Why data citation isn’t working, and what to do about it. *Database J. Biol. Databases Curation*, 2020.
- [12] Buneman, P., Davidson, S. B., and Frew, J. (2016). Why data citation is a computational problem. *Commun. ACM*, 59(9):50–57.
- [13] Buneman, P., Khanna, S., and Tan, W. C. (2001). Why and where: A characterization of data provenance. In *Database Theory - ICDT 2001, 8th International Conference*, pages 316–330.
- [14] Buneman, P. and Silvello, G. (2010). A rule-based citation system for structured and evolving datasets. *IEEE Data Eng. Bull.*, 33(3):33–41.
- [15] Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A. M., Lowry, R. K., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, a., and Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC’s Environmental Data Centres. *International Journal of Digital Curation*, 7(1):107–113.
- [16] Candela, L., Castelli, D., Manghi, P., and Tani, A. (2015). Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, 66(9):1747–1762.
- [17] Cheney, J., Chiticariu, L., and Tan, W. (2009). Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474.
- [18] Chockler, H. and Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115.
- [19] CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013). *Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data*, volume 12.

- [20] Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., and Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, 18(1).
- [21] Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.
- [22] Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *JASIST*, 52(7):558–569.
- [23] Cui, Y., Widom, J., and Wiener, J. L. (2000). Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227.
- [24] Davidson, S. B., Deutch, D., Milo, T., and Silvello, G. (2017). A model for fine-grained data citation. In *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research*. www.cidrdb.org.
- [25] Deutch, D., Frost, N., Kimelfeld, B., and Monet, M. (2022). Computing the Shapley Value of Facts in Query Answering. In Bonifati, A. and Abbadi, A. E., editors, *SIGMOD '22: International Conference on Management of Data, Philadelphia, June 12-17, 2022*. ACM.
- [26] Dosso, D. and Silvello, G. (2020). Data credit distribution: A new method to estimate databases impact. *Journal of Informetrics*, 14(4):101080.
- [27] Dubernet, M. L., Antony, B. K., Ba, Y. A., et al. (2016). The virtual atomic and molecular data centre (VAMDC) consortium. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(7):074003.
- [28] Eiter, T. and Lukasiewicz, T. (2002). Complexity results for structure-based causality. *Artif. Intell.*, 142(1):53–89.
- [29] Fang, H. (2018). A discussion of citations from the perspective of the contribution of the cited paper to the citing paper. *JASIST*, 69(12):1513–1520.
- [30] Garfield, E. (1999). Journal impact factor: a brief review. *Can. Med. Assoc.*, 979-980.

- [31] Gößwein, B., Miksa, T., Rauber, A., and Wagner, W. (2019). Data identification and process monitoring for reproducible earth observation research. In *2019 15th International Conference on eScience (eScience)*, pages 28–38. IEEE.
- [32] Green, T. J., Karvounarakis, G., and Tannen, V. (2007). Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40. ACM.
- [33] Halpern, J. Y. and Pearl, J. (2013). Causes and explanations: A structural-model approach — part 1: Causes. *CoRR*, abs/1301.2275.
- [34] Harding, S. D., Sharman, J. L., Faccenda, E., Southan, C., Pawson, A. J., Ireland, S., Gray, A. J. G., Bruce, L., Alexander, S. P. H., Anderton, S., Bryant, C., Davenport, A. P., Doerig, C., Fabbro, D., Levi-Schaffer, F., Spedding, M., Davies, J. A., and Nc-Iuphar (2018). The IUPHAR/BPS guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Research*, 46(Database-Issue):D1091–D1106.
- [35] Hartley, J. (2017). Authors and their citations: a point of view. *Scientometrics*, 110(2):1081–1084.
- [36] Hey, T., Tansley, S., and Tolle, K. M. (2009). Jim Gray on eScience: a transformed scientific method.
- [37] Honor, L. B., Haselgrove, C., Frazier, J. A., and Kennedy, D. N. (2016). Data citation in neuroimaging: proposed best practices for data identification and attribution. *Frontiers in neuroinformatics*, 10:34.
- [38] Jain, S., Moritz, D., Halperin, D., Howe, B., and Lazowska, E. (2016). Sqlshare: Results from a multi-year sql-as-a-service experiment. In *Proceedings of the 2016 International Conference on Management of Data*, pages 281–293.
- [39] Joshi-Tope, G., Gillespie, M., Vastrik, I., D’Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database-Issue):428–432.

- [40] Katz, D. (2014). Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2(1).
- [41] Kosten, J. (2016). A classification of the use of research indicators. *Scientometrics*, 108(1):457–464.
- [42] Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, 6(2):4–37.
- [43] Livshits, E., Bertossi, L. E., Kimelfeld, B., and Sebag, M. (2020). The shapley value of tuples in query answering. In Lutz, C. and Jung, J. C., editors, *23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark*, volume 155 of *LIPICs*, pages 20:1–20:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.
- [44] Martone, M. (2014). Joint declaration of data citation principles. *FORCE11. San Diego CA. Data Citation Synthesis Group*. <https://www.force11.org/datacitationprinciples>, online September 2020.
- [45] Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the american society for information science and technology*, 58(13):2105–2125.
- [46] Meliou, A., Gatterbauer, W., Moore, K. F., and Suciu, D. (2010). The complexity of causality and responsibility for query answers and non-answers. *Proc. VLDB Endow.*, 4(1):34–45.
- [47] Milo, T. (2019). Getting rid of data. *Journal of Data and Information Quality (JDIQ)*, 12(1):1–7.
- [48] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, M., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers,

- 1435 E. J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research
1436 culture. *Science*, 348(6242):1422–1425.
- 1437 [49] Peters, I., Kraker, P., Lex, E., Gumpenberger, C., and Gorraiz, J.
1438 (2016). Research data explored: An extended analysis of citations and
1439 altmetrics. *Scientometrics*, 107(2):723–744.
- 1440 [50] Pröll, S. and Rauber, A. (2013). Scalable data citation in dynamic,
1441 large databases: Model and reference implementation. In *Proceedings of*
1442 *the 2013 IEEE International Conference on Big Data, 6-9 October 2013,*
1443 *Santa Clara, CA, USA*, pages 307–312.
- 1444 [51] Rauber, A., Ari, A., van Uytvanck, D., and Pröll, S. (2016). Identifi-
1445 cation of Reproducible Subsets for Data Citation, Sharing and Re-Use.
1446 *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue*
1447 *on Data Citation*, 12(1):6–15.
- 1448 [52] Rauber, A., Asmi, A., van Uytvanck, D., and Proell, S. (2015). Data
1449 citation of evolving data: Recommendations of the working group on data
1450 citation (wgdc). *Result of the RDA Data Citation WG*, 20.
- 1451 [53] Remil, Y., Bendimerad, A., Mathonat, R., Chaleat, P., and Kaytoue,
1452 M. (2021). ” what makes my queries slow?”: Subgroup discovery for sql
1453 workload analysis. *arXiv preprint arXiv:2108.03906*.
- 1454 [54] Shapley, L. S. (1954). A value for n-person games. In Kuhn, H. W. and
1455 Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages
1456 307–317. Princeton University Press, Princeton.
- 1457 [55] Silvello, G. (2018). Theory and practice of data citation. *J. Assoc. Inf.*
1458 *Sci. Technol.*, 69(1):6–20.
- 1459 [56] Simmhan, Y., Plale, B., and Gannon, D. (2005). A survey of data
1460 provenance in e-science. *SIGMOD Record*, 34(3):31–36.
- 1461 [57] Spengler, S. (2012). Data Citation and Attribution: A Funder’s Per-
1462 spective. In of Sciences’ Board on Research Data, N. A. and Information,
1463 editors, *Report from Developing Data Attribution and Citation Practices*
1464 *and Standards: An International Symposium and Workshop*, pages 177–
1465 178. National Academies Press: Washington DC.

- 1466 [58] Vogelsgesang, A., Haubenschild, M., Finis, J., Kemper, A., Leis, V.,
 1467 Mühlbauer, T., Neumann, T., and Then, M. (2018). Get real: How bench-
 1468 marks fail to represent the real world. In *Proceedings of the Workshop on*
 1469 *Testing Database Systems*, pages 1–6.
- 1470 [59] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G.,
 1471 Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B.,
 1472 Bourne, P. E., et al. (2016). The fair guiding principles for scientific data
 1473 management and stewardship. *Scientific data*, 3.
- 1474 [60] Wu, Y., Alawini, A., Davidson, S. B., and Silvello, G. (2018). Data
 1475 citation: Giving credit where credit is due. In *Proceedings of the 2018*
 1476 *International Conference on Management of Data, SIGMOD*, pages 99–
 1477 114.
- 1478 [61] Zeng, T., Wu, L., Bratt, S., and Acuna, D. E. (2020). Assigning credit to
 1479 scientific datasets using article citation networks. *Journal of Informetrics*,
 1480 14(2).
- 1481 [62] Zou, C. and Peterson, J. B. (2016). Quantifying the scientific output of
 1482 new researchers using the zp-index. *Scientometrics*, 106(3):901–916.
- 1483 [63] Zwölf, C. M., Moreau, N., and Dubernet, M.-L. (2016). New Model for
 1484 Datasets Citation and Extraction Reproducibility in VADMC. *Journal of*
 1485 *Molecular Spectroscopy*, 327:122–137.