Data Credit Distribution through Lineage*

Dennis Dosso[0000−0001−7307−4607] and Gianmaria Silvello[000−0003−4970−4554]

Department of Information Engineering, University of Padua,
{dosso, silvello}@dei.unipd.it

Abstract. Data are a fundamental asset in the current world of research. Data citation is becoming more common and supported by research databases, but it still presents many research challenges.
This paper describes *Data Credit*, a new measure of value for data derived from data citation, that enables us to annotate databases with real values representing their importance. Credit, computed through the citations, can be used alongside them to better understand the importance of data.
We introduce the task of *Data Credit Distribution*, the process by which credit produced by a citation is and assigned to the data in a database responsible for producing the output information being cited.
We describe how this process can be performed and, through experiments, we show that credit can serve, among other things, to highlight "hotspots" in the database.

Keywords: Data Citation · Data Credit · Data Provenance

1 Introduction

It is widely accepted that citations are the "currency" of the scientific world, a fundamental method to perform dissemination of knowledge and foster scientific development [22]. Scientific databases, "populated and updated with a great deal of human effort" [4], are numerous and at the core of the scientific research [5]. It is globally accepted that data must be cited and citable [18, 7, 10].

Data citations should be, among other things, counted alongside traditional citations and contribute to bibliometrics indicators to reward scientific database curators for their effort [1, 20]. Data citation is often considered in the current literature as a driving force to "facilitate giving scholars credit" [19]. One of its central aspects is how to attribute credit to data creators and curators [6]. Many data creators and curators still do not receive any form of reward for their work; this fosters the growth of detrimental phenomenons like the "reward dilemma", the fear from researchers to share their data, losing their competitive advantage without proper recognition of their work [14].

* The full paper was originally published in the Journal of Informetrics [12]

How to handle and count the credit generated by data citations and how it contributes to traditional and new bibliometrics are long-standing research issues [15, 2]. However, even when correctly applied, data citations and the related bibliometrics do not always accurately reward data. Indeed, a query often uses more data than the one present in its output result set. The data being used but not visualized do not receive a citation, nor do their contributors.

To overcome this limitation, in recent years, the idea of *crediting data* emerged in the academic discussion through the concept of *data credit*, a real positive value describing the importance of data in a given context. We argue that credit can be used to address some of the limitations highlighted above. Credit is not atomic like a citation. Once computed, it can be divided into portions and assigned to all the data used by a query. Credit can be used as an annotation set at different granularity levels within a database to describe their importance.

In this work, we discuss the problem of *data credit distribution*, the issuance of credit generated by some query Q on a relational database instance I to the data in I responsible for the generation of $Q(I)$. In particular, we discuss how the distribution is possible in relational databases through lineage, a form of data provenance [9]. While data citation and credit distribution are not limited to relational databases, they are a good test bed for this first approach. In Section 2, we report the related work; Section 3 presents the methods used and the experimental results carried on a real scientific database, GtoPdb; Section 4 contains the conclusions.

2 Related Work

Kats in [17] suggests the need for a *modified citation system* that includes the idea of *transient* and *fractional credit*. Credit is defined as a "quantity" representing the importance of a research entity (a paper, software or data) mentioned in a citation, but these ideas are proposed without any formalism.

Fang in [13] presents a framework to distribute credit generated by a paper to its authors and to the papers in its reference list in a transitive way. Each cited paper's quantity of credit depends on its impact/role in the citing paper. This theoretical framework works for a graph composed of only papers, but it can be extended to another graph model that includes data.

Zeng et al. in [21] proposed the first method designed to compute credit within a network of papers citing data. This is the first step towards an automatic credit computation procedure. However, it is limited to assigning credit to the whole dataset without considering variable data granularity. Therefore, this is not a way to assign credit to a single research entity within a dataset.

3 Methods and Experiments

Methods. Data Credit is a non-negative real value representing the importance of data in a specific context. It can be computed with different strategies and rationales. In this paper's context, we consider credit as the product of a data

citation; therefore, it is a quantity representing the importance of the data being cited in the citing paper. Ideally, the higher the impact of the cited data in the citing paper, the bigger the credit.

The task of *Data Credit Distribution* (DCD) consists of dividing this credit into portions and assigning it to the recipients in a database responsible for generating the cited data. Formally:

Definition 1. ***Data Credit Distribution at tuple level*** *(DCD) [12]*
Given a database instance I, a query Q over I and the value $k \in \mathbb{R}_{>0}$, DCD is defined as the computation of the function $f_{I,Q} : TupleLoc \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ such that $f_{I,Q}(t,k) = h$ where $0 \leq h \leq k$ and $\sum_{t \in TupleLoc} f_{I,Q}(t,k) = k$.

f is the *Distribution Strategy* (DS), it aims to annotate each tuple (thus we speak of DCD at *tuple* level) in I with a portion of the credit. Its only requirement is that it has to be *conservative*: no credit is generated or lost during the distribution. A DS can be defined in many different ways, but what we may prefer is a function that distributes credit coherently with the role of the input tuples as defined by Q. That is, only tuples that had some role in generating $Q(I)$ should receive credit.

To do so, we propose one definition of DS that exploits the concept of *lineage* [11]. Given a tuple $t \in Q(I)$, its lineage is the set of all and only the tuples that have a role, whatever it is, in the generation of t.

Definition 2. ***Lineage-based Distribution Strategy*** *[12]*
Let I be a database instance, Q a query over I, $o \in Q(I)$ an output tuple and k the credit associated to o. Let L be the lineage of o and t be a generic tuple in I. t receives a credit equal to:

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

As we see, this DS equally rewards the tuples of the lineage of a tuple. To perform the whole distribution on $Q(I)$, it is simply necessary to apply this DS to each tuple $o \in Q(I)$.

Evaluation. We considered the IUPHAR/BPS Guide to Pharmacology (GtoPdb) [16], a famous and highly cited medical database containing information about drugs, targets, and ligands. GtoPdb is maintained and curated by a consortium of 512 scientists collaborating with in-house curators, distributed in committees [3].

GtoPdb is relational in nature, and its information is also organized into webpages describing specific diseases, receptors, ligands, and families of these elements.

To gather data citations, we considered papers published in the British Journal of Pharmacology (BJP) that cite [16]. [16] is a recent version of a series of papers that the GtoPdb consortium releases every two years to describe the database and its evolutions. It works as a data journal that can be cited in place of the whole database [8]. The papers published in BJP that refer to specific

Fig. 1. Heat-map of the distribution of credit to the `family` table. Each cell represents a tuple in the table.

webpages of GtoPdb report the URL of the referenced page. It is possible from these URLs to reverse-engineer the SQL queries that compute the data contained in the webpages. A webpage is composed by different parts, each part created with data extracted from the GtoPdb through SQL queries. We use these queries to perform DCD. We focused only on queries referring to the so-called target families[1].

Without any loss of generality, we assumed that each tuple present in the output of these queries contains credit equal to 1, and we performed credit distribution through lineage using these queries that we inferred from the BJP papers. We used the ∼900 BJP papers citing [16] as of October 2020, and we extracted from them more than 1200 SQL queries to families of receptors.

The results of the distribution on the `family` table of GtoPdb, that contain information about the target families, are shown in the heat-map of Figure 1. Each cell in the map is a tuple, and the intensity of the color represents the assigned quantity of credit. Interestingly, few tuples receive almost all the credit, following a Pareto distribution. This shows how credit distribution can highlight "hotspots", elements in the database that receive high values of credit. These are tuples that are used frequently by queries. Interestingly, these may also be tuples that are used but not visualized in the final output. This means that credit allows to rewards parts of the database that are used but not visualized, overcoming a limitation of traditional citations.

To better see how credit differs from traditional citations, consider Figure 2. We reported two radar plots, presenting the top 10 authors citation-wise and

[1] `https://www.guidetopharmacology.org/targets.jsp`

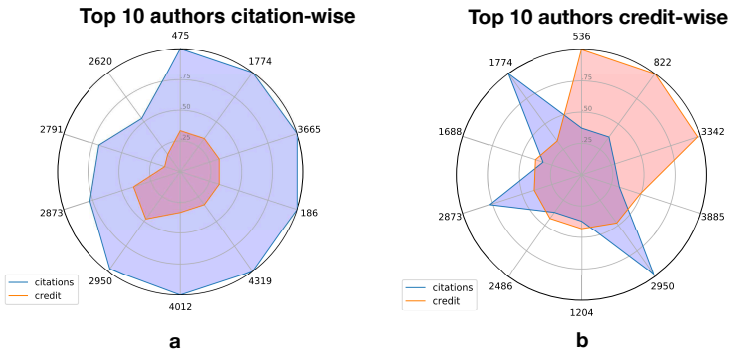

Fig. 2. Radar plots showing the top-10 authors of GtoPdb citation (a) and credit (b) with their normalized values of credit and citations.

credit-wise (values normalized between 0 and 1, and the authors were substituted with numbers for privacy reasons). To compute the citations, we proceeded as follows: each time a query identifies data curated by an author, that author receives one citation and equally shares the credit assigned to that data with the other co-authors of that data. As we see from Figure 2.a, the top 10 authors, citation-wise, do not have the highest values of credit. Similarly, in Figure 2.b, the authors with the higher values of credit do not also have the highest citation count.

This shows that credit can reward authors whose data have a high impact in the research community, i.e., those data generated the highest quantity of credit, even if they received fewer citations than other authors. That is, specific citations are "more valuable", credit-wise. Since we assumed that each output tuple carries credit 1, the queries that return outputs with more tuples also generate more credit. In more complex scenarios, where different and more sophisticated techniques may be used to decide how to generate quantities of credit, credit distribution can help to understand how data and their corresponding authors impact the scientific environment.

4 Conclusions

We showed how credit can highlight parts of the database that cover certain topics instead of others, as defined by queries. Credit and citations are correlated measures, but credit offers a new perspective to evaluate the impact of both data and curators. It can highlight parts of the database related to certain query topics, so-called "hotspots". It directly rewards the tuples, and corresponding authors, that contributed to the production of cited data, even those that are not in the output itself. Moreover, it proportionately rewards data and curators based on their impact in the context defined by the issued queries. This helps to reward authors that would otherwise remain unnoticed. In future works, credit can

become the basis for new bibliometrics and applications based on its presence. For example, *data pricing*, that is the identification of the price of certain data in a database based on how much they are used by queries.

Acknowledgments

This work is partially supported by the ExaMode project, as part of the European Union Horizon 2020 program under Grant Agreement no. 825292.

References

1. Belter, C.W.: Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLoS ONE **9**(3), e92590 (2014)
2. Borgman, C.L.: Data Citation as a Bibliometric Oxymoron. In: Sugimoto, C.R. (ed.) Theories of Informetrics and Scholarly Communication, pp. 93–116. De Gruyter Mouton (2016)
3. Buneman, P.: How to cite curated databases and how to make them citable. In: 18th International Conference on Scientific and Statistical Database Management, SSDBM. pp. 195–203. IEEE Computer Society (2006)
4. Buneman, P., Cheney, J., Tan, W.C., Vansummeren, S.: Curated Databases. In: Proc. of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008. pp. 1–12 (2008), `https://doi.org/10.1145/1376916.1376918`
5. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Commun. ACM **59**(9), 50–57 (2016)
6. Buneman, P., Christie, G., Davies, J.A., Dimitrellou, R., Harding, S.D., Pawson, A.J., Sharman, J.L., Wu, Y.: Why data citation isn't working, and what to do about it. Database J. Biol. Databases Curation **2020** (2020), `https://doi.org/10.1093/databa/baaa022`
7. 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., Wright, D.: 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 (2012), `http://dx.doi.org/10.2218/ijdc.v7i1.218`
8. Candela, L., Castelli, D., Manghi, P., Tani, A.: Data Journals: A Survey. Journal of the Association for Information Science and Technology **66**(9), 1747–1762 (2015), `http://dx.doi.org/10.1002/asi.23358`
9. Cheney, J., Chiticariu, L., Tan, W.: Provenance in databases: Why, how, and where. Foundations and Trends in Databases **1**(4), 379–474 (2009)
10. CODATA-ICSTI Task Group on Data Citation Standards and Practices: Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, vol. 12 (September 2013). https://doi.org/http://doi.org/10.2481/dsj.OSOM13-043
11. Cui, Y., Widom, J., Wiener, J.L.: Tracing the lineage of view data in a warehousing environment. ACM Trans. Database Syst. **25**(2), 179–227 (2000)
12. Dosso, D., Silvello, G.: Data credit distribution: A new method to estimate databases impact. Journal of Informetrics **14**(4), 101080 (2020)

13. Fang, H.: A discussion of citations from the perspective of the contribution of the cited paper to the citing paper. JASIST **69**(12), 1513–1520 (2018)
14. Fienberg, S.E., Martin, M.E., Straf, M.L.: Sharing research data. National Academy Press (1985)
15. Garfield, E.: Journal impact factor: a brief review (1999), *Can. Med. Assoc.*, 979-980
16. 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., Nc-Iuphar: 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 (2018)
17. Katz, D.: 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) (2014)
18. Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S.: Citation and Peer Review of Data: Moving Towards Formal Data Publication. International Journal of Digital Curation **6**(2), 4–37 (2011)
19. Martone, M.: Joint declaration of data citation principles. FORCE11. San Diego CA. Data Citation Synthesis Group. (2014). https://doi.org/10.25490/a97f-egyk, `https://www.force11.org/datacitationprinciples`, online September 2020
20. Peters, I., Kraker, P., Lex, E., Gumpenberger, C., Gorraiz, J.: Research data explored: An extended analysis of citations and altmetrics. Scientometrics **107**(2), 723–744 (2016)
21. Zeng, T., Wu, L., Bratt, S., Acuna, D.E.: Assigning credit to scientific datasets using article citation networks. Journal of Informetrics **14**(2) (2020)
22. Zou, C., Peterson, J.B.: Quantifying the scientific output of new researchers using the zp-index. Scientometrics **106**(3), 901–916 (2016)