

CoCon: A Data Set on Combined Contextualized Research Artifact Use

Tarek Saier
tarek.saier@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Youxiang Dong
youxiang.dong@student.kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Michael Färber
michael.farber@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

In the wake of information overload in academia, methodologies and systems for search, recommendation, and prediction to aid researchers in identifying relevant research are actively studied and developed. Existing work, however, is limited in terms of granularity, focusing only on the level of papers or a single type of artifact, such as data sets. To enable more holistic analyses and systems dealing with academic publications and their content, we propose CoCon, a large scholarly data set reflecting the combined use of research artifacts, contextualized in academic publications' full-text. Our data set comprises 35 k artifacts (data sets, methods, models, and tasks) and 340 k publications. We additionally formalize a link prediction task for "combined research artifact use prediction" and provide code to utilize analyses of and the development of ML applications on our data. All data and code is publicly available at <https://github.com/III/Depence/contextgraph>.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Information extraction**; *Language resources*; **Knowledge representation and reasoning**; • **Mathematics of computing** → **Graph theory**.

KEYWORDS

scholarly data, research artifacts, link prediction, information extraction

ACM Reference Format:

Tarek Saier, Youxiang Dong, and Michael Färber. 2023. CoCon: A Data Set on Combined Contextualized Research Artifact Use. In *Proceedings of* . ACM, New York, NY, USA, 4 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The rate of scholarly publication is consistently rising [2], confronting researchers with an ever greater challenge of identifying research that is relevant to their endeavours and interests. To assist researchers in this task, several systems such as academic search engines and recommender systems, as well as prediction approaches have been developed. However, existing approaches are limited in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

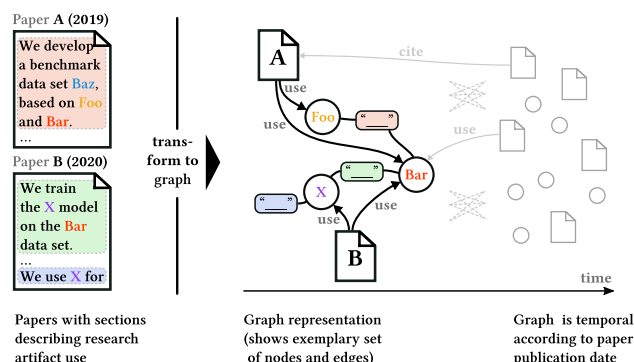


Figure 1: Schematic of our data set.

Created from a large corpus of paper full-texts (unarXive) and metadata on research artifacts (Papers With Code), our data set is a graph representing papers being published over time, and the research artifacts they use. Usage is contextualized through text sections within the papers.

terms of the granularity and variety of entities that are subject to the search, recommendation, or prediction offered. Many systems and methods focus on papers, for example, the recommendation of papers for reading based on a user profile. Recent approaches operate on a more fine grained level—for example data set recommendation based on a research task description. These are, however, still constrained to a single type of entity (e.g. data sets only). This is problematic because research endeavours commonly involve multiple "research artifacts" such as *methodologies* and *data sets*, which means single entity type methods only reflect them partially. Furthermore can the *combined* use of existing ideas and artifacts be seen to reflect innovation, in academia [4, 5] as well as other collaborative areas [16].

To enable the study of combined use of research artifacts, and to overcome aforementioned limitations, we propose CoCon, a data set capturing the combined use of research artifacts, contextualized through academic publication text. The proposed data set is a graph containing 35 k research artifacts as well as 340 k publications (graph nodes). As shown in Figure 1, these are connected (graph edges) according to publications' use of artifacts, as well as through a citation network. In addition to our data set, we provide a code base for easy extension to machine learning (ML) applications, as well as a formalization of a link prediction task in a dynamic graph for "combined research artifact use prediction".

Overall, we make the following contributions.

- We propose CoCon, a graph data set reflecting 35 k research artifacts and their contextualized use in 340 k academic publications. By providing CoCon we, enable the study of and

development of ML applications for a more holistic picture of research endeavours than was possible before.

- We formalize a link prediction task for “combined research artifact use prediction”.
- We provide code to facilitate easy updates of our data set as well as easy extension to additional ML tasks.

2 DATA SET

We first discuss our process of data source selection and data scheme definition in accordance to each other. After that, we describe the resulting preprocessing steps as well as our final data set.

2.1 Scope and Data Sources

The term “research artifact” is used in the literature to refer to research outputs which can be re-used or further developed in subsequent work. Typical examples are datasets, software prototypes, ontologies, and methodologies [12, 17]. We considered various data sources providing information on, content of, as well as associations between (1) research artifacts and (2) publications. In particular, the data sets SCIERC [10], SciREX [6], and Papers With Code¹ are candidates regarding research artifacts, while S2ORC [9] and unarXive [15] are viable data sources for publications.

For research artifacts, we decide to use Papers With Code in favor of SCIERC and SciREX. This choice is motivated by the fact SCIERC is rather small in comparison to the other two alternatives, and SciREX is built from Papers With Code. Directly using Papers With Code allows us to obtain a larger amount and also more recent data. Furthermore, it will allow us to update our data in the future, as Papers With Code periodically update their shared data.

Using Papers With Code as a data source for research artifacts has two important implications. First, the classes of research artifacts we are able to model are (1) data set, (2) method, (3) model, and (4) task. Secondly, the field of publications covered is mainly from the area of machine learning. Because a majority of the publications in Papers With Code are linked to author copies on arXiv.org, both unarXive and S2ORC are a good fit, since both cover the entirety of arXiv papers. Between the two, we chose to use unarXive, which is motivated by two reasons. First, unarXive includes a more complete citation network (43% compared to 31%), and second, newer S2ORC releases do not include arXiv papers parsed from L^AT_EX sources anymore²—which are less prone to noise than PDF parses. Our data sources therefore are *Papers With Code* and *unarXive*.

2.2 Preprocessing

The data provided by Papers With Code contains metadata on data sets, methods, models, and tasks, as well as the papers that use the aforementioned. However, the data model is not suitable for transformation into a graph *as is*. In particular, not all entities are readily available as “first class citizen” records, but only described in relation to others. Models, for example, are only available as attributes of tasks. We transform the data accordingly, persisting (1) for each artifact its set of metadata attributes provided by Papers

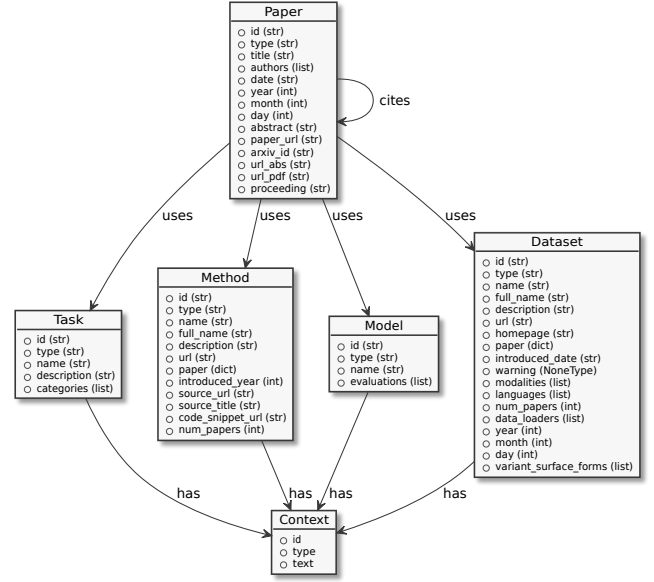


Figure 2: Graph schema of CoCon.



Figure 3: Simplified graph schema of CoCon_{simple}.

With Code, and (2) links from each artifact to the papers it is used by.

To extend our data with paper full-texts and subsequently usage contexts (see Figure 1), we need to connect the paper metadata records in Papers With Code to unarXive. This process is straightforward by use of paper’s arXiv ID. In addition to the paper full-texts, this also adds unarXive’s citation network to our data.

For the extraction of paper text segments, in which the usage of an artifact is described, we match each occurrence of the artifact’s name in all papers which Papers With Code lists as using the artifact. In most cases this is unproblematic, as artifact names are unique, capitalized proper names or acronyms. We note some exceptions of artifacts named ambiguously, such as the data sets “seeds”³ and “iris”,⁴ which we filter heuristically to prevent noise.

2.3 Data Set Description

Through the process described above, we get data on papers, artifacts, and their interconnections according to the schema shown in Figure 2. We additionally define a simplified graph schema, CoCon_{simple}, shown in Figure 3, where artifacts are connected by weighted edges according to the number of papers that use them together. Contexts in the simplified graph schema are represented as artifact nodes’ attributes.

Our data set contains 340,965 paper nodes, 22,189 model nodes, 7,592 data set nodes, 3,389 task nodes, and 2,026 method nodes.

¹See <https://github.com/paperswithcode/paperswithcode-data>.

²“Release: 2020-07-05 [...] omitted LaTeX parses from this release.”, see <https://github.com/allenai/s2orc> (accessed 2023/02/12).

³See <https://paperswithcode.com/dataset/seeds>.

⁴See <https://paperswithcode.com/dataset/iris-1>.

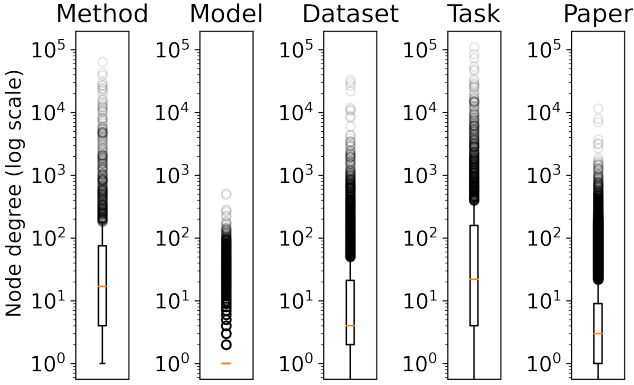


Figure 4: Node degrees.

Furthermore, it contains 2,413,664 contexts. Figure 4 shows the distribution of degrees across the different node types. We note a clear long tail distribution for all types—that is, most research artifacts and papers have few connections, while a few have a large number of connections. A distinct case are models, most of which (72%) have a degree of 1. This can be explained by the fact that authors use experiment specific model identifiers in tables describing evaluation results, which are the source of model names.⁵

To make our data set easily usable as a graph, we provide code to load it as a `networkx`⁶, and `PyTorch Geometric`⁷ graph, as shown in Listing 1.

```
from contextgraph.util.graph import
    load_full_graph
G = load_full_graph(with_contexts=True)
print(type(G))
# output: networkx.classes.digraph.DiGraph

from contextgraph.util.torch import
    load_entity_combi_graph
G_combi = load_entity_combi_graph()
print(type(G_combi))
# output: torch_geometric.data.data.Data
```

Listing 1: Python usage example.

3 LINK PREDICTION TASK

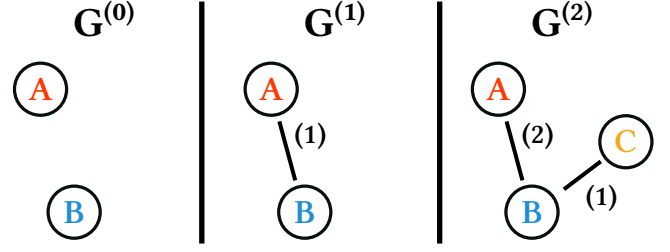
In line with existing work using the Papers With Code data, our data set can be used for various types of analyses and applications, such as the analysis of research artifact usage patterns [8], the evaluation of information extraction approaches [18], and content based paper recommendation [13].

As a novel application made possible by CoCon specifically, we formalize the link prediction task “combined research artifact

⁵See <https://github.com/paperswithcode/sota-extractor>.

⁶See <https://networkx.org/>.

⁷See <https://pytorch-geometric.readthedocs.io/>.

Figure 5: Visualisation of graph snapshots for CoCon_{simple}.

At $G^{(0)}$, two papers are published, one of them using artifact A, the other one using artifact B. At $G^{(1)}$ a paper is published that uses artifacts A and B in combination. At $G^{(2)}$ a paper is published that uses artifacts A, B, and C in combination. **Note:** edge weights are indicated by numbers in brackets.

use prediction” (CRAUP). The reasons for choosing this task are twofold. First, while commonly used scholarly data sets for link prediction tasks exist, they typically consist only of paper nodes connected through citation edges and furthermore are comparatively small (see Cora, CiteSeer, and PubMed Diabetes in the CTU Prague Relational Learning Repository [11] used in, for example, [1, 3, 14]). We therefore see a benefit in formalizing a link prediction task with a focus on research artifacts insofar, as a new type of real world phenomenon (use of research artifacts) is reflected. Secondly does the real world application scenario bear potential for the identification of beneficial research endeavours—for example, if reliable predictions can be made that a certain method is suited for application on a task, or that a certain data set is suited as training data for a model. In the following, we present a brief description and formalization of the task.

3.1 Task Description

The goal of CRAUP, as the name suggests, is to predict whether or not two research artifacts will be used together in the future. The basis for the prediction is the evolution of the graph reflecting publications and research artifacts. Papers within CoCon naturally have a publication date which, by extension, can be used to also date research artifacts—i.e., they enter the graph with the paper that first uses them. We can therefore view both our CoCon and CoCon_{simple} as a time discrete dynamic graph [7], that is, a graph that changes in discrete time steps.

Formally, we can then define the prediction problem on the series of graph snapshots $G^{(0)}, \dots, G^{(T)}$ with time steps $t \in T$. Each graph $G^{(t)} = (V^{(t)}, E^{(t)})$ consists of vertices $v \in V^{(t)}$ and edges $(u, v) \in E^{(t)}$. The types of possible changes from any $G^{(t)}$ to the subsequent $G^{(t+1)}$ are as follows.

- Adding of nodes (papers published in $t + 1$, and artifacts used for the first time in a papers published in $t + 1$)
- Adding of edges (edges associated with above nodes)
- Increase of edge weights in CoCon_{simple} (the weight of an edge (u, v) increases by the number of papers published in $t + 1$ that use both u and v)

In Figure 5 we show a simple visualisation, illustrating all possible changes over three graph snapshots in CoCon_{simple}.

We define the most basic form of the prediction task, CRAUP_{step}, on CoCon_{simple}. The task is for every time step to predict which

edges will be newly added to the graph. Formally, for each point in time $t \in T$ and each pair of nodes u, v with $u \in V^{(t)}$, $v \in V^{(t)}$, and $(u, v) \notin E^{(t)}$, predict whether or not $(u, v) \in E^{(t+1)}$.

Because CRAUP_{step} is very specific in requiring an edge to be added at the immediate subsequent point in time $t + 1$, we also define a more general task CRAUP_{future}. Here, the prediction is not made for $(u, v) \in E^{(t+1)}$ but rather $(u, v) \in E^{(t+1, \dots, T)}$, that is, predict whether or not the two artifacts u and v will be used together at any point in the future, not necessarily already at the immediate subsequent point in time $t + 1$.

As a further extension of CRAUP_{future}, we define CRAUP_{gap}, where for an additional time gap parameter Δ the prediction is made for $(u, v) \in E^{(t+1+\Delta, \dots, T)}$. In other words, predict whether or not the two artifacts will be used together in the future, but not immediately. Our reason for defining CRAUP_{gap} is to introduce a sense of “early prediction”. The higher Δ can be set for a model without compromising on its evaluation performance, the earlier it can predict whether or not two artifacts will be used together in the future.

We see investigations of the three tasks CRAUP_{step}, CRAUP_{future}, and CRAUP_{app} as a first step into exploring regularities in combined contextualized research artifact use over time. Further extensions to, for example, the inclusion of predicting edge weight changes in CoCon_{simple}, link prediction tasks in CoCon, or other types of prediction tasks, are potential following steps.

4 CONCLUSION

We propose CoCon, a large scholarly data set reflecting the combined use of research artifacts, contextualized in academic publications’ full-text. The data set is a graph comprising 35 k artifacts (data sets, methods, models, and tasks) and 340 k publications, primarily from the area of machine learning. By focusing on artifact usage as opposed to mentions or references in general, and by contextualizing usage in publications’ full-text, we enable novel types of analyses and ML applications.

In addition to making CoCon freely available, we formalize a link prediction task “combined research artifact use prediction” (CRAUP) and provide code to utilize analyses of and the development of ML applications with CoCon. We plan to investigate CRAUP potential application scenarios of models developed in the process. This paves the way toward a new generation of systems aiding researchers in effectively identifying publications and research artifacts to use in their future endeavors.

AUTHOR CONTRIBUTIONS

Tarek Saier: Conceptualization (lead), Data curation, Formal analysis (lead), Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. Youxiang Dong: Conceptualization (support), Formal analysis (support). Michael Färber: Writing – review & editing.

ACKNOWLEDGMENTS

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) via [KOM,BI], a Software Campus project (01IS17042)..

REFERENCES

- [1] Seong Jin Ahn and MyoungHo Kim. 2021. Variational Graph Normalized AutoEncoders. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2827–2831. <https://doi.org/10.1145/3459637.3482215>
- [2] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (2021), 224. <https://doi.org/10.1057/s41599-021-00903-w>
- [3] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. 2018. Hyperspherical variational auto-encoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018 (34th Conference on Uncertainty in Artificial Intelligence (AUAI))*, 856–865.
- [4] Sitaram Devarakonda, James R. Bradley, Dmitriy Korobskiy, Tandy Warnow, and George Chacko. 2020. Frequently cocited publications: Features and kinetics. *Quantitative Science Studies* 1, 3 (Aug. 2020), 1223–1241. https://doi.org/10.1162/qss_a_00075
- [5] Hugo Hedlund. 2020. Mapping the evolving intellectual structure of digital innovation research on the public sector : a document co-citation analysis. *IEEE Computer Society*, 2042–2051. <http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-167165>
- [6] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, 7506–7516. <https://doi.org/10.18653/v1/2020.acl-main.670>
- [7] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. 2022. Representation Learning for Dynamic Graphs: A Survey. *J. Mach. Learn. Res.* 21, 1 (June 2022).
- [8] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/3b8a614226a953a8cd9526fca6fe9ba5-Paper-round2.pdf>
- [9] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4969–4983.
- [10] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.
- [11] Jan Motl and Oliver Schulte. 2015. The CTU Prague Relational Learning Repository. <https://doi.org/10.48550/ARXIV.1511.03086>
- [12] Viet Bach Nguyen and Vojtěch Svátek. 2021. Ontology for Informatics Research Artifacts. In *The Semantic Web: ESWC 2021 Satellite Events*, Ruben Verborgh, Anastasia Dimou, Aidan Hogan, Claudia d’Amato, Ilaria Tiddi, Arne Bröring, Simon Mayer, Femke Ongena, Riccardo Tommasini, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 126–130.
- [13] Malte Ostendorf, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized Document Embeddings for Aspect-Based Similarity of Research Papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22)*. Association for Computing Machinery, Article 7, 12 pages. <https://doi.org/10.1145/3529372.3530912>
- [14] Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. 2018. Adversarially Regularized Graph Autoencoder for Graph Embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (Stockholm, Sweden) (IJCAI '18)*. AAAI Press, 2609–2615.
- [15] Tarek Saier and Michael Färber. 2020. unarXive: A Large Scholarly Data Set with Publications’ Full-Text, Annotated In-Text Citations, and Links to Metadata. *Scientometrics* 125, 3 (Dec. 2020), 3085–3108. <https://doi.org/10.1007/s11192-020-03382-z>
- [16] Juan Tan, Dongqing Miao, and Qiong Tan. 2020. Empirical Study on Influencing Factors of Knowledge Product Remixing in OIC. *IEEE Access* 8 (2020), 34215–34224. <https://doi.org/10.1109/ACCESS.2020.2974693>
- [17] Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL Citations in Scholarly Papers for Promoting Utilization of Research Artifacts. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics, Online, 8–19. <https://aclanthology.org/2022.wiesp-1.2>
- [18] Raquib Bin Yousuf, Subhodip Biswas, Kulendra Kumar Kaushal, James Dunham, Rebecca Gelles, Sathappan Muthiah, Nathan Self, Patrick Butler, and Naren Ramakrishnan. 2022. Lessons from Deep Learning applied to Scholarly Information Extraction: What Works, What Doesn’t, and Future Directions. <https://doi.org/10.48550/ARXIV.2207.04029>