A Appendix

The appendix includes supplementary information including links to the dataset and the code repository for repeated experiments in subsection A.1, as well as the detailed dataset documentation and intended uses in the form of a datasheets for datasets available in subsection A.3.

A.1 Supplementary information and links

The URL to access the dataset is provided below:

`https://doi.org/10.6084/m9.figshare.14709507`

The obtained persistent dereferencable identifier (DOI minted by the data repository) is therefore: 10.6084/m9.figshare.14709507.

Authors bear all responsibility in case of violation of rights. The data is made publicly available under the Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). The dataset should not be used for commercial purposes.

Hosting is performed by FigShare, the authors are responsible for maintaining the dataset.

All explanations on how to read the dataset, with examples, is provided via jupyter notebooks as part of the code repository for repeated experiments:

`https://github.com/mateuszjurewicz/procat`

Additionally, the best performing model is made available with the DOI 10.5281/zenodo.4896303 at this hosting address:

`https://zenodo.org/record/4896303#.YLnxgZMzb0Q`

The dataset is intended to be publicly available forever, hence it was uploaded to the FigShare data repository, which also handles its discoverability through structured metadata. For more information, see:

`https://knowledge.figshare.com/publisher/fair-figshare`

A.2 Further notes on dataset diversity

The diversity of the dataset is limited due to the offer text being in Danish. Our intention was to provide a valuable resource for an underrepresented language. One important aspect of the dataset is that the catalogues come from a wide variety of providers, including cross-border shops that have a significant following in neighboring Scandinavian countries, particularly Sweden and Norway, as well as Germany.

We also provide an overview of commercial categories that the catalogues belong to, following the Global Product Classification (GPC-GS1), with multiple categories per catalogue, in table 5.

Table 5: Global Product Classification of PROCAT Catalogues

Category	Number of Catalogues	%
Food (FBT)	7,456	67.40%
Electronic	5,231	47.28%
Personal Care	5,113	46.22%
Tools	3,311	29.93%
Sports Equipment	2,147	19.41%
Lawn/Garden Supplies	2,039	18.43%
Home Appliances	2,028	18.33%
Baby Care	1,986	17.95%
Household Furniture	1,672	15.11%
Pet Care	1,522	13.76%
Footwear	1,324	11.97%
Toys and Games	1,293	11.69%
Fuels	548	4.95%

Finally, the number of individual retailers that the catalogues belonged to is approximately 2,400 and the total number of unique users who have viewed the catalogues within the app is approximately 2.5 million. Our hope is to represent a broad array of product categories and providers.

A.3 Datasheets for Datasets

The following includes answers to all the questions from the suggested datasheets for datasets framework [Gebru et al., 2018].

1. **Motivation**
   (a) **For what purpose was the dataset created?**
   The dataset in its current form was created with the purpose of helping solve an industrial challenge of optimal catalogue structure prediction.
   (b) **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
   Original raw data collection was performed as part of the day-to-day operations of the company Tjek A/S, which aggregates product catalogues for viewing in a digital format. The curation and preprocessing was performed by the authors of this paper.
   (c) **Who funded the creation of the dataset?**
   The research is funded through an Innovation Fund Denmark research grant that Tjek A/S is a beneficiary of (grant number 9065-00017B).
2. **Composition**
   (a) **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
   The instances represent 3 types of entities. The most atomic entity is an *offer*, which represents a specific product with a text heading and description, which often includes its on-offer price. Individual product offers are then grouped into *sections*, which represent pages in a physical catalogue brochure. Finally, an ordered list of sections comprise a single *catalogue*, for which a prediction about its optimal structure is made. This takes the form of permuting the input set of offers into an ordered list, with section breaks marking the start and end of a section.
   (b) **How many instances are there in total (of each type, if appropriate)?**
   The dataset consists of just over 10 thousand catalogs (11063), almost a quarter of a million sections (238256) and over 1.5 million offers (1613686). These are further grouped into a suggested 80/20 train and test split, with 8850 catalogs in the train set and 2212 in the test set.
   (c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   The dataset is not a sample, it contains all catalogue instances from the years 2015 - 2019 available for viewing in the Tjek A/S app. No other selection filter was used.
   (d) **What data does each instance consist of?**
   Each instance consists of both raw data and pre-processed features.
   Each offer instance consists of its unique id, its related section and catalogue ids, a text heading and description in both raw form and as word tokens using the nltk tokenizer [Bird, 2006], the total token count, and finally the full offer text as a vector referencing a vocabulary of 300 thousand word tokens. Additionally, each offer is categorized into a priority class, representing how visually prominent it was in the original catalogue in terms of relative image size (on a 1-3 integer scale).
   Each catalogue instance consists of its unique id, an ordered list of associated section ids, and an ordered list of offer ids that comprise the catalogue in question, including section break markers. Additionally, each catalogue instance also includes information in the form of ordered lists of offers as vectors, grouped into sections, their corresponding priority class and the catalogue's total number of offers. Finally a shuffled x of offer vectors (with section breaks) is provided for each catalogue, along with the target y representing the permutation required to restore the original order.
   (e) **Is there a label or target associated with each instance?**

Yes, each catalogue instance is pre-processed into a shuffled x of offer vectors and section break markers, along with the target y representing the permutation required to restore the human-designed structure of the original catalogue.

(f) **Is any information missing from individual instances?**
No data is missing.

(g) **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
Yes, every offer instance is tied to its section and catalogue via their ids in the appropriate columns of the provided comma-separated files.

(h) **Are there recommended data splits (e.g., training, development/validation, testing)?**
Yes, the entire catalogue set is grouped into a suggested 80/20 train and test split, with 8850 catalogs in the train set and 2212 in the test set. Catalogues were assigned to each group randomly. A validation set can be extracted from the train set based on each researcher's individual preference.

(i) **Are there any errors, sources of noise, or redundancies in the dataset?**
There are no known errors, sources of noise or redundancies in the dataset, however there is a possibility of some degree of overlap between individual offers in terms of the underlying product.

(j) **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
The dataset is self-contained.

(k) **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)?**
The dataset does not contain data that might be considered confidential.

(l) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
The dataset does not contain data that the authors would consider offensive, insulting, threatening or causing anxiety.

(m) **Does the dataset relate to people?**
The dataset does not relate to people (thus skipping the remainder of this section's questions).

3. **Collection Process**

(a) **How was the data associated with each instance acquired?**
The data was acquired through a combination of feed readers and custom scraping scripts developed by Tjek A/S. For further details, see the answer to the next question.

(b) **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**
The scripts read the feeds and scrape a list of stores and PDF catalogs associated with said stores. This provides the basic tooling and processing of the data and communicates this to the company's core API, running the scrapers on a defined schedule as well as on-demand. Following that, a human curation step is performed by the operations department to make sure the obtained data is correct. The data is directly observable.

(c) **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
The dataset is not a sample.

(d) **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
The data collection process was done as part of the day-to-day operations of Tjek A/S, by properly compensated full-time employees.

(e) **Over what timeframe was the data collected?**
The data was collected within the full 4 year period between 2015 and 2019.

(f) **Were any ethical review processes conducted (e.g., by an institutional review board)?**
No.

(g) **Does the dataset relate to people?**
No, thus skipping the remainder of the questions in this section.

4. **Preprocessing / cleaning / labeling**

(a) **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
Yes, the raw text features of each offer instance were tokenized using the nltk tokenizer [Bird, 2006], a vocabulary of word tokens was limited to 300 thousand words and used to obtain offer vectors. Each offer instance was truncated or padded to 30 word tokens, with over 75% of offers consisting of fewer than 24 tokens. Each catalogue instance was truncated or padded to 200 offer instances, with over 75% of catalogues consisting of fewer than 163 offers.
Additionally, to obtain the prominence class per offer per section, signifying the relative size of the offer's image on the page, a proprietary algorithm was used.

(b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
Yes, raw data is also provided.

(c) **Is the software used to preprocess/clean/label the instances available?**
Yes, the nltk library is available under the Apache License 2.0.

5. **Uses**

(a) **Has the dataset been used for any tasks already?**
The dataset is actively being used to help predict the optimal structure of product catalogues given a provided set of offers, based on their textual description and to recommend complementary offers. It has not been used in prior research.

(b) **Is there a repository that links to any or all papers or systems that use the dataset?**
The repository containing the scripts for repeated experiments will include links to any and all papers using this dataset. For more information, see the appendix subsection A.1.

(c) **What (other) tasks could the dataset be used for?**
The dataset can be used for representation learning through the co-occurrence of offers within the same section, leading to a complementariness-based recommendation system. It can also be used for learning to cluster a set of offers into a variable number of sections, which is an implicit step in the main task of predicting the entire structure of a catalogue through permutation learning (as it includes the section break markers).

(d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
It is important to remember that the provided catalogues represent the Danish market between 2015-2019, and thus might not represent patterns that will hold in other societies. This, however, has no bearing on demonstrating a machine learning model's ability to learn structure through joint clustering and permutation learning, which is the intended use of the dataset.

(e) **Are there tasks for which the dataset should not be used?**
The dataset is not meant to be used as a representation of the market for any form of trend prediction.

6. **Distribution**

(a) **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
The dataset will be made publicly available under the chosen license to any and all parties. For more information see the appendix subsection A.1.

(b) **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?**
The dataset is distributed through a dataset hosting service and has a DOI, for details see the appendix subsection A.1.

(c) **When will the dataset be distributed?**
The dataset will be distributed by the time of the paper's submission.

(d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
The dataset will be distributed under the Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). The dataset should not be used for commercial purposes.

(e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

(f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
No.

7. **Maintenance**

(a) **Who is supporting/hosting/maintaining the dataset?**
The dataset is hosted by *figshare*, an open access repository where researchers can preserve and share their research outputs, including figures, datasets, images and videos. It is supported by Digital Science & Research Solutions Ltd.
It is maintained by the authors of this paper.

(b) **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
Via the emails provided in the contact information above the abstract, repeated here for convenience: maju@itu.dk; leod@itu.dk.

(c) **Is there an erratum?**
There is currently no erratum, it will be added to both the main sharing link and the github repository containing the code for repeated experiments should the need to create an erratum occur.

(d) **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
If labeling errors are found, they will be corrected. The dataset may be expanded with further instances, depending on the academic interest and number of downloads.

(e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
The dataset does not relate to people.

(f) **Will older versions of the dataset continue to be supported/hosted/maintained?**
Yes, all previous versions of the dataset will continue to be available.

(g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
Others are encouraged to extend the dataset and can choose to either do so in cooperation with the authors of this paper after contacting them via the provided email addresses or individually in accordance with the chosen license.