A APPENDIX

A.1 LIMITATIONS

In this work, we mainly focus on learning multiplex text/node representations on text-rich networks and solving downstream tasks (*e.g.,* classification, regression, matching) with the learned embeddings. Because of the limited computational budgets, our PLM text encoder (bert-base-uncased) is medium-scale. In the future, we will explore applying a similar multiplex representation learning philosophy to large-scale language models. Other interesting future directions include designing more advanced graph-empowered language models to learn the multiplex embeddings and adopting the model into more real-world applications such as network generation.

A.2 ETHICAL CONSIDERATIONS

PLMs have been shown to be highly effective in encoding contextualized semantics and understanding documents, as evidenced by several studies Devlin et al. (2019); Liu et al. (2019b); Clark et al. (2020). However, some researchers have pointed out certain limitations associated with these models, including the presence of social bias Liang et al. (2021) and the propagation of misinformation Abid et al. (2021). In our work, we focus on utilizing the relation signals between texts from the multiplex text-rich network structure to facilitate the understanding of the semantics of the texts, which we believe could help to address issues related to bias and misinformation.

A.3 DATASETS

The statistics of the five datasets can be found in Table 7. In academic networks, nodes correspond to papers and there are five types of relations between papers: "`cited-by`" (cb), "`same-author`" (sa), "`same-venue`" (sv), "`co-reference`" (cr), and "`co-cited-by`" (ccb); while in e-commerce networks, nodes are items and there are four types of relations between items: "`co-purchased`" (cop), "`co-viewed`" (cov), "`bought-together`" (bt), and "`co-brand`" (cob).

Table 7: Dataset Statistics.

Dataset	#Nodes	#Relations (Edges)
Geology	431,834	cb (1,000,000), sa (1,000,000), sv (1,000,000) cr (1,000,000), ccb (1,000,000)
Mathematics	490,551	cb (1,000,000), sa (1,000,000), sv (1,000,000) cr (1,000,000), ccb (1,000,000)
Clothes	208,318	cop (100,000), cov (100,000) bt (100,000), cob (50,000)
Home	192,150	cop (100,000), cov (100,000) bt (50,000), cob (100,000)
Sports	189,526	cop (100,000), cov (100,000) bt (50,000), cob (100,000)

A.4 DISTRIBUTION SHIFT BETWEEN DIFFERENT RELATIONS

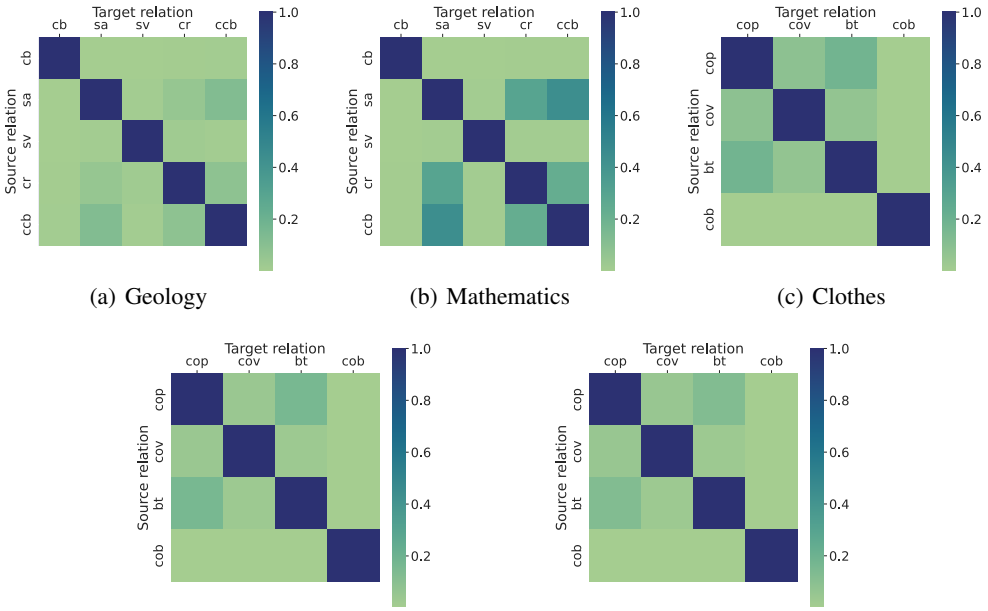
In Section 3, we show the learned embedding distribution shift between different relations on Geology in Figure 2. In this section, we calculate the raw data distribution shift between different relations' distribution $P_{r_k}(e_{ij}|v_i, v_j)$. The distribution shift is measured by Jaccard score:

$$\text{Jac}(r_k, r_l) = \frac{|P_{r_k}(e_{ij}) \cap P_{r_l}(e_{ij})|}{|P_{r_k}(e_{ij}) \cup P_{r_l}(e_{ij})|} \tag{11}$$

Since the whole networks are too large to calculate the Jaccard score, we randomly sample a sub-network from each network that contains 10,000 nodes and calculate Eq.(A.4). The results on Geology, Mathematics, Clothes, Home and Sports networks can be found in Figure 6. If the

assumption of analogous distributions (*i.e.*, $P_{r_k}(e_{ij}|v_i, v_j) \approx P_{r_l}(e_{ij}|v_i, v_j)$) holds, the values in each cell should be nearly one, which is not the case in Figure 6.

More empirical experiments on the learned embedding distribution shift between relations in Mathematics, Clothes, Home, and Sports networks can be found in Figure 7. We finetune BERT[4] Devlin et al. (2019) to generate embeddings on one source relation distribution $P_{r_k}(e_{ij}|v_i, v_j)$ (row) and test the embeddings on the same or another target relation distribution $P_{r_l}(e_{ij}|v_i, v_j)$ (column). If the assumption of analogous distributions (*i.e.*, $P_{r_k}(e_{ij}|v_i, v_j) \approx P_{r_l}(e_{ij}|v_i, v_j)$) holds, the values in each cell should be nearly the same, which is not the case in Figure 7.



(a) Geology (b) Mathematics (c) Clothes

(d) Home (e) Sports

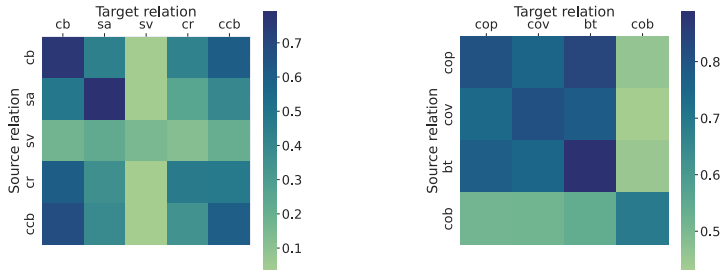
Figure 6: Raw data distribution shift between different relations on Geology, Mathematic, Clothes, Home, and Sports network. cb, sa, sv, cr, and ccb represent "`cited-by`", "`same-author`", "`same-venue`", "`co-reference`", and "`co-cited-by`" relation respectively. cop, cov, bt, and cob represent "`co-purchased`", "`co-viewed`", "`bought-together`", and "`co-brand`" relation respectively. Each entry is the Jaccard score between the corresponding two relation distributions.

A.5 Experimental Setting

A.5.1 Multiplex Representation Learning

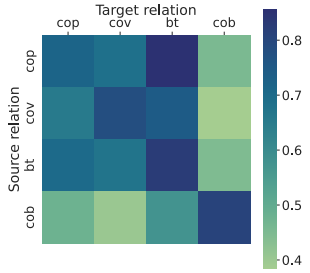
Hyperparameter setting. To facilitate the reproduction of our representation learning experiments, we provide the hyperparameter configuration in Table 8. Vanilla FT, MTDNN, and METERN use exactly the same set of hyperparameters for a fair comparison. The last layer `[CLS]` token hidden states are utilized to develop $\boldsymbol{h}_{v|r}$ for Vanilla FT, MTDNN, and METERN. Paper titles and item titles are used as text associated with the nodes in the two kinds of networks, respectively. (For some items, we concatenate the item title and description together since the title is too short.) The models are trained for 40 epochs on 4 Nvidia A6000 GPUs with a total batch size of 512. The total time cost is around 17 hours and 2 hours for networks in the academic domain and e-commerce domain respectively. Code is available at `https://anonymous.4open.science/r/METER-submit-2C7B`.

[4]We use the bert-base-uncased checkpoint.

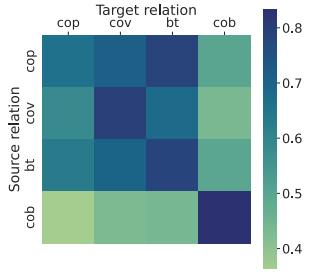


(a) Mathematics

(b) Clothes



(c) Home



(d) Sports

Figure 7: Distribution shift between different relations on Mathematic, Clothes, Home, and Sports network. cb, sa, sv, cr, and ccb represent "`cited-by`", "`same-author`", "`same-venue`", "`co-reference`", and "`co-cited-by`" relation respectively. cop, cov, bt, and cob represent "`co-purchased`", "`co-viewed`", "`bought-together`", and "`co-brand`" relation respectively. Each entry is the PREC@1 of BERT fine-tuned on the corresponding source relation distribution and tested on the corresponding target relation distribution.

Table 8: Hyper-parameter configuration for representation learning.

Parameter	Geology	Mathematics	Clothes	Home	Sports
Max Epochs	40	40	40	40	40
Peak Learning Rate	5e-5	5e-5	5e-5	5e-5	5e-5
Batch Size	512	512	512	512	512
# Prior Tokens m	5	5	5	5	5
Warm-Up Epochs	4	4	4	4	4
Sequence Length	32	32	32	32	32
Adam ϵ	1e-8	1e-8	1e-8	1e-8	1e-8
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Clip Norm	1.0	1.0	1.0	1.0	1.0
Dropout	0.1	0.1	0.1	0.1	0.1

A.5.2 Direct Inference with an Evident Source Relation

We provide problem definitions and experimental settings for tasks in Section 5.4. The tasks include venue recommendation, author identification, and brand prediction.

Venue Recommendation.

Problem Definition. Given a query paper node (with associated text) and a candidate venue list (each with its published papers), we aim to predict which venue in the candidate list should be recommended for the given query paper.

Experimental Settings. We adopt in-batch testing with a testing batch size of 256. We use PREC@1 as the metric. The max sequence length is 32 and 256 for the query paper and venue (concatenation of its 100 randomly selected published papers' titles) respectively.

Author Identification.

Problem Definition. Given a query paper node (with associated text) and a candidate author list (each with his/her published papers), we aim to predict which people in the candidate list is the author of the given query paper.

Experimental Settings. We adopt in-batch testing with a testing batch size of 256. We use PREC@1 as the metric. The max sequence length is 32 and 256 for the query paper and author (concatenation of his/her 100 randomly selected published papers' titles) respectively.

Brand Prediction.

Problem Definition. Given a query item node (with associated text) and a candidate brand list (each with items in the brand), we aim to predict which one in the candidate list is the brand for the given query item.

Experimental Settings. We adopt in-batch testing with a testing batch size of 256. We use PREC@1 as the metric. The max sequence length is 32 and 256 for query paper and brand (concatenation of its 100 randomly selected items' texts) respectively.

A.5.3 Learn to Select Source Relations

We provide problem definitions and experimental settings for tasks in Section 5.5. The tasks include paper recommendation, paper classification, citation prediction, year prediction, item classification, and price prediction.

Paper Recommendation.

Problem Definition. Given a query paper node (with associated text) and a candidate paper list (each with associated text), we aim to predict which paper in the candidate list should be recommended to the people who are interested in the query paper.

Experimental Settings. We have 1,000 samples in the train set to teach the models how to select source relations, 1,000 samples in the validation set to conduct the early stop, and 100,000 samples in the test set to evaluate the performance of the models. The learning rate is set as 1e-3, the training batch size is 128, and the testing batch size is 256. We conduct the in-batch evaluation with PREC@1 as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 4.

Paper Classification.

Problem Definition. Given a query paper node (with associated text), we aim to predict what is the category of the paper. The number of paper node categories in academic networks (Geology and Mathematics) is shown in Table 9.

Table 9: Number of paper node categories in academic networks.

Geology	Mathematics
18	17

Experimental Settings. We have 1,000 samples for each category in the train set to teach the models how to select source relations, 200 samples for each category in the validation set to conduct the early stop, and 200 samples for each category in the test set to evaluate the performance of the models. The learning rate is tuned in 5e-3 and 1e-3, the training batch size is 256, and the testing batch size is 256. We adopt Macro-F1 as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 4.

Citation Prediction.

Problem Definition. Given a query paper node (with associated text), we aim to predict its future number of citations.

Experimental Settings. For both Geology and Mathematics datasets, we extract papers the citation of which is in the range from 0 to 100. We randomly select 10,000 papers from the extracted papers to form the training set, 2,000 papers to form the validation set, and 2,000 papers to form the test set. The learning rate is set as 1e-2 for all compared methods, the training batch size is 256, and the

testing batch size is 256. We adopt RMSE as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 4.

Year Prediction.

Problem Definition. Given a query paper node (with associated text), we aim to predict the year when it was published.

Experimental Settings. For both Geology and Mathematics datasets, we conduct minus operations to make the smallest ground truth year to be 0 (we minus all year numbers by the earliest year, *i.e.,* 1981, in MAG). We randomly select 10,000 papers from the extracted papers to form the training set, 2,000 papers to form the validation set, and 2,000 papers to form the test set. The learning rate is set as 1e-2 for all compared methods, the training batch size is 256, and the testing batch size is 256. We adopt RMSE as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 4.

Item Classification.

Problem Definition. Given a query item node (with associated text), we aim to predict what is the category of the item. The number of item node categories in e-commerce networks (Clothes, Home, and Sports) is shown in Table 10.

Table 10: Number of item node categories in e-commerce networks.

Clothes	Home	Sports
7	9	16

Experimental Settings. We have 1,000 samples for each category in the train set to teach the models how to select source relations, 200 samples for each category in the validation set to conduct the early stop, and 200 samples for each category in the test set to evaluate the performance of the models. The learning rate is tuned in 5e-3 and 1e-3, the training batch size is 256, and the testing batch size is 256. We adopt Macro-F1 as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 5.

Price Prediction.

Problem Definition. Given a query item node (with associated text), we aim to predict its price.

Experimental Settings. For Clothes, Home, and Sports, we delete the long-tail items and keep items whose prices are under 100/1,000/100 respectively. We randomly select 10,000 items from the extracted items to form the training set, 2,000 items to form the validation set, and 2,000 items to form the test set. The learning rate is set as 1e-2 for all compared methods, the training batch size is 256, and the testing batch size is 256. We adopt RMSE as the metric. All experiments are done on one NVIDIA A6000. We repeat three runs for each model and show the mean and standard deviation in Table 5.

A.6 Relation Embedding Initialization Study.

We study how different relation embedding initialization affects the quality of multiplex representations learned by METERN. We explore three initialization settings: zero vectors initialization, normal distribution initialization, and word embedding initialization. The results of average PREC@1 on different networks are shown in Table 11. From the results, there is no significant difference between the representation learning quality of different initialized relation embeddings.

Table 11: Performance of different relation embedding initialization on different networks.

Model	Geology	Mathematics	Clothes	Home	Sports
METERN w/ zero init	36.31	**55.26**	79.99	79.68	77.59
METERN w/ randn init	**36.43**	55.20	**80.19**	79.69	77.56
METERN w/ word init	36.36	55.20	80.12	**79.83**	**77.71**

A.7 More Results on Learn to Select Source Relations.

In section 4.2, we propose to let the model learn to select source relations for different downstream tasks and show the learned source relation weight for citation prediction and paper recommendation in Figure 3. In this section, we show more results on the learned source relation weight on academic network downstream tasks in Figure 8 and on e-commerce network downstream tasks in Figure 9. There are four downstream tasks in the academic domain: year prediction, citation prediction, paper classification, and paper recommendation. There are two downstream tasks in the e-commerce domain: price prediction and item classification. From Figure 8 and Figure 9, we can find that different relations can benefit different downstream tasks.

On the year prediction task, the "`co-cited-by`" relation and "`same-author`" relation are more useful. This indicates that papers tend to cite recent papers and authors tend to be active within a short period (*e.g.,* active during the Ph.D. study and stop publishing papers after graduation).

On the citation prediction task, the "`same-author`" relation and "`same-venue`" relation are more beneficial. This implies that the impact of papers is more determined by the published venue and the author who writes them (people tend to follow works from famous researchers).

On the paper classification task, the "`cited-by`" relation is quite useful. This means that papers and their cited papers have a tendency to have the same fine-grained topics.

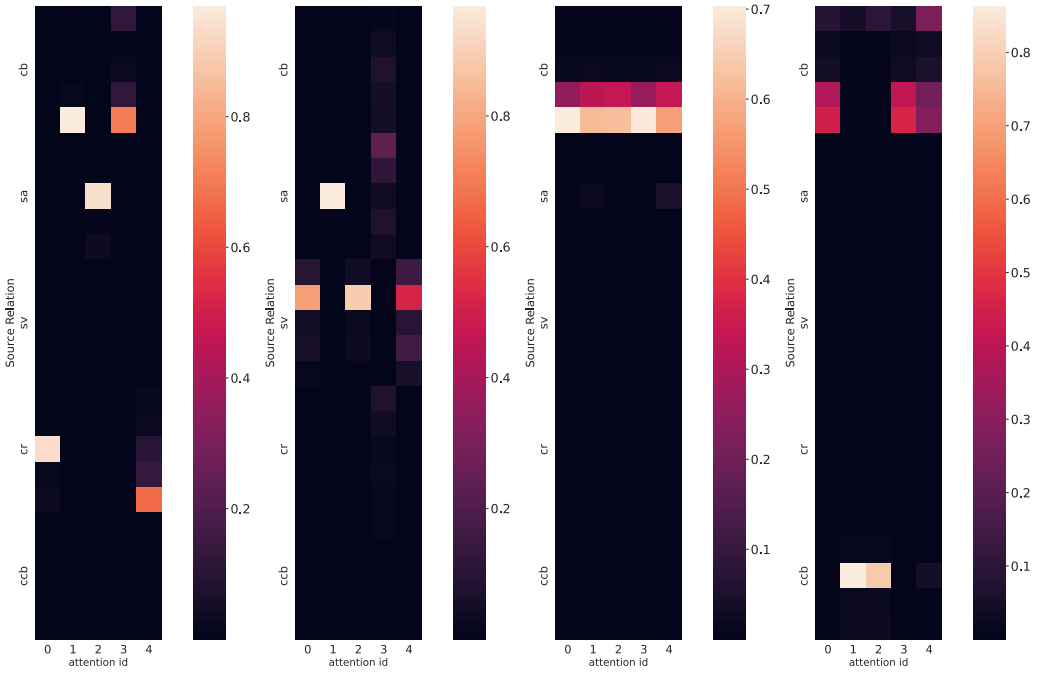
On the paper recommendation task, the "`cited-by`" relation and "`co-cited-by`" relation dominate. The goal of the paper recommendation task is to recommend similar papers to researchers which may contribute to their own research development. The result is interesting since papers and their cited papers have a tendency to have the same topic, and thus should be recommended to researchers together, and papers in the "`co-cited-by`" relation have already demonstrated that their ideas can be combined and result in a new paper (they are cited by same papers), and thus should be recommended together.

On the price prediction task, the "`co-viewed`" and "`bought-together`" relation matters a lot. This implies that the same user tends to view items and buy together items of a similar price range.

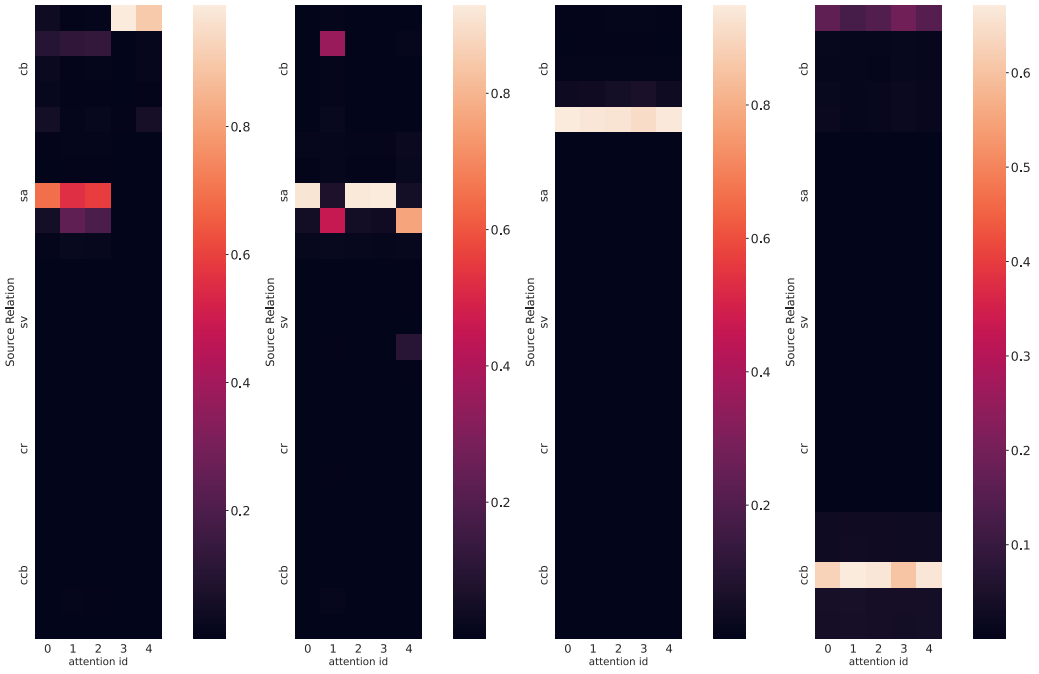
On the item classification task, the "`co-viewed`" relation dominates. This means that items co-viewed by the same user tend to have similar functions.

A.8 Relation Token Embedding Visualization.

We visualize the relation token embeddings $Z_{\mathcal{R}}$ learned by METERN with t-SNE Van der Maaten & Hinton (2008), projecting those embeddings into a 2-dimensional space. The results on the Geology network and Mathematics network are shown in Figure 10, where the embeddings belonging to the same relation are assigned the same color. From the results, we can find that embeddings belonging to the same relation are close to each other, while those belonging to different relations are discriminative. This demonstrates that METERN can learn to capture the different semantics of different relations by assigning different relations' embeddings to different areas in the latent space.

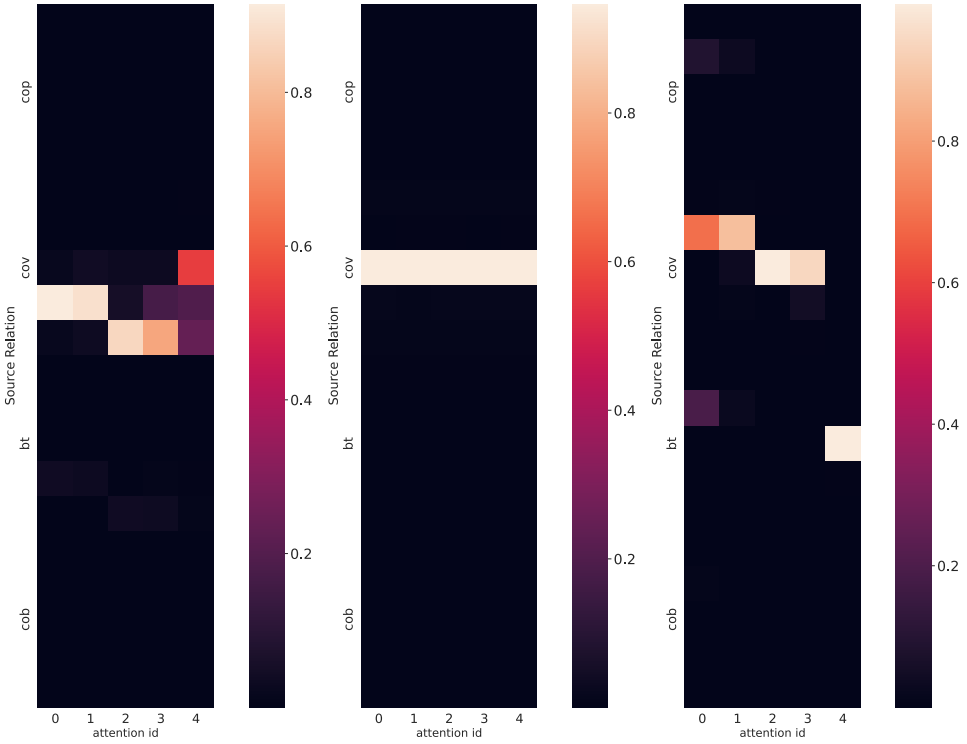


(a) Geology: Year (b) Geology: Citation (c) Geology: Class (d) Geology: Rec

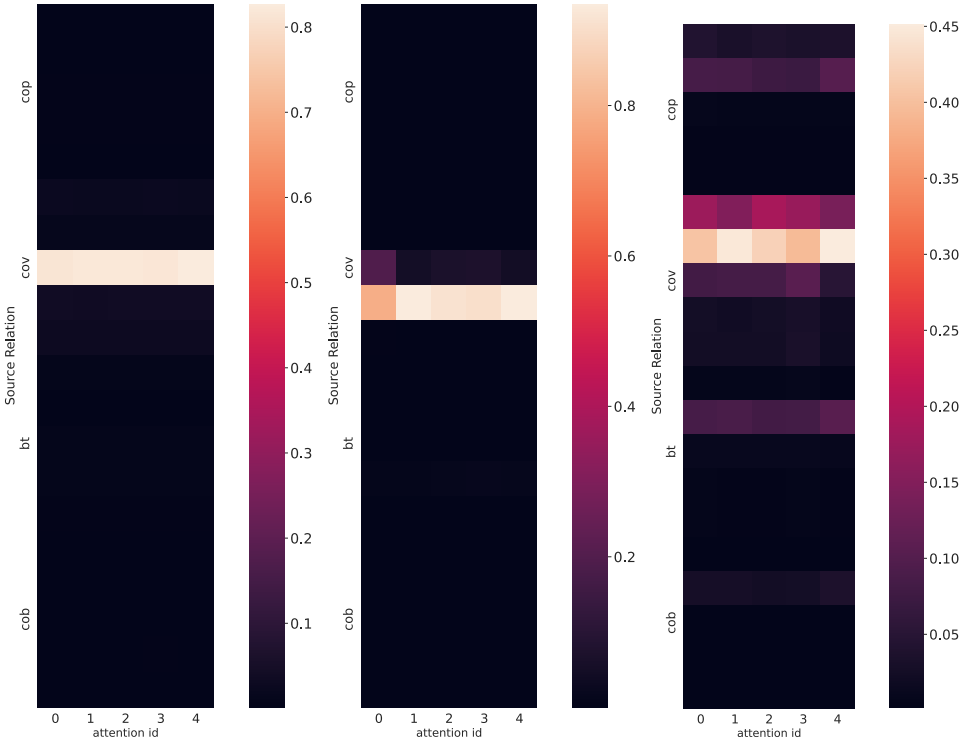


(e) Mathematics: Year (f) Mathematics: Citation (g) Mathematics: Class (h) Mathematics: Rec

Figure 8: Learnt source relation weights for tasks on academic networks (Geology and Mathematics). The x-axis is the learned weight vector id and the y-axis is the relation embedding id grouped by relation id. cb, sa, sv, cr, and ccb represent "`cited-by`", "`same-author`", "`same-venue`", "`co-reference`", and "`co-cited-by`" relation respectively.

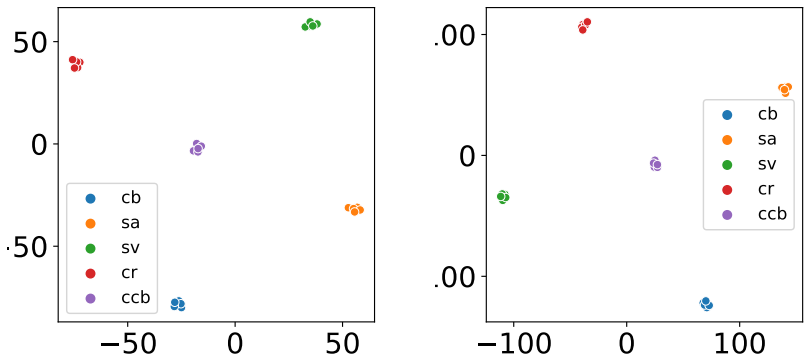


(a) Clothes: Price (b) Clothes: Class (c) Home: Price



(d) Home: Class (e) Sports: Price (f) Sports: Class

Figure 9: Learnt source relation weights for tasks on e-commerce networks (Clothes, Home, and Sports). The x-axis is the learned weight vector id and the y-axis is the relation embedding id grouped by relation id. cop, cov, bt, and cob represent "`co-purchased`", "`co-viewed`", "`bought-together`", and "`co-brand`" relation respectively.



(a) Geology (b) Mathematics

Figure 10: Visualization of relation embedding Z_R on Geology and Mathematics. cb, sa, sv, cr, and ccb represent "`cited-by`", "`same-author`", "`same-venue`", "`co-reference`", and "`co-cited-by`" relation respectively. We can find that embeddings belonging to the same relation are close to each other, while those belonging to different relations are discriminative.