
# Analyses of Open Academic Graph Challenge @ KDD Cup 2024

Fanjin Zhang

Tsinghua University

2024.04.26



智谱·AI

# Academic Knowledge Graph

Knowledge graph

Google Scholar

Microsoft Academic

SEMANTIC SCHOLAR

WEB OF SCIENCE™

Academic search
and mining system

AMiner

63,986,555
RESEARCHERS

282,537,875
PUBLICATIONS

8,912,399
CONCEPTS

4,208,422,145
CITATIONS

Large-scale
heterogeneous entities

https://www.aminer.cn,
accessed on April 25, 2024.

# Background

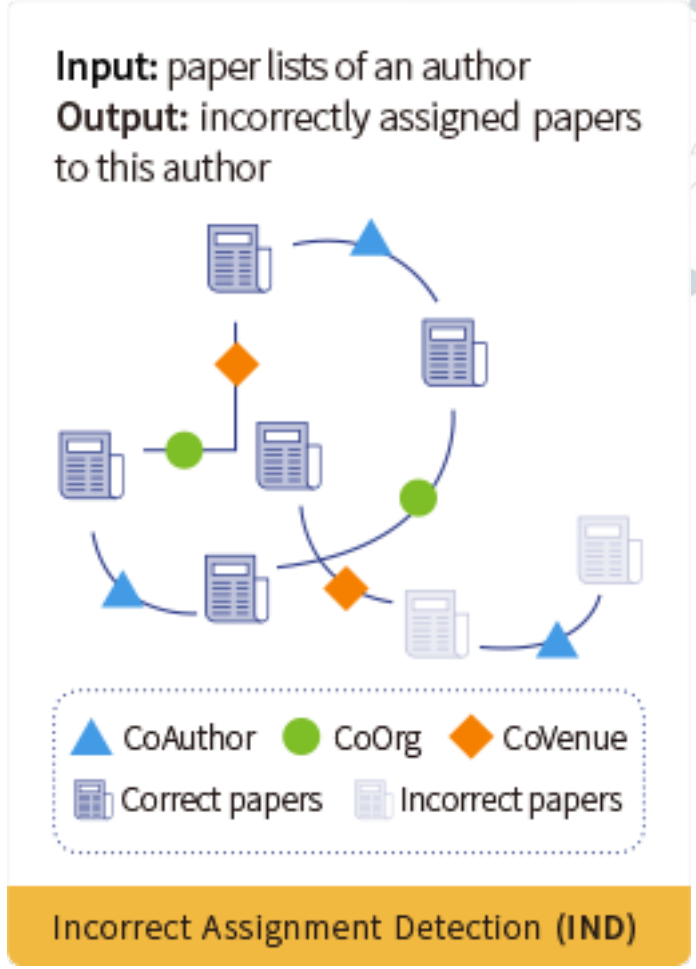
- Academic graph mining offers the potential to unlock enormous scientific, technological, and educational value

- Assist governments in making scientific policies
- Support companies in talent discovery
- Help researchers acquire new knowledge more efficiently

- Community efforts to advance academic graph mining have been severely limited by the lack of a suitable public benchmark.

# Task Description

- Track 1: Incorrect Assignment Detection (WhoIsWho-IND)
  - Given the paper assignments of each author and paper metadata,
  - the goal is to detect paper assignment errors for each author.




[1] Bo Chen, Jing Zhang, Fanjin Zhang, Tianyi Han, Yuqing Cheng, Xiaoyan Li, Yuxiao Dong, and Jie Tang. “Web-Scale Academic Name Disambiguation: the WhoIsWho Benchmark, Leaderboard, and Toolkit.” KDD 2023.
[2] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. “OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining”. arXiv:2402.15810.

# WhoIsWho-IND: Datasets

- Track 1: Incorrect Assignment Detection (WhoIsWho-IND)
  - train_author.json: [Dictionary] mapping author id to author name, correct paper assignments (“normal”), and incorrect paper assignments (“outliers”)
  - pid_to_info_all.json: [Dictionary] mapping paper id to paper attributes
  - ind_valid_author.json
  - ind_ valid_author_submit.json
- Total: 1600+ authors, 300K+ papers

```
{
    "HoH18DsE": { # Author IDs,
        "name": "xxx", # Name of the author,
        "normal_data":[ # Papers belong to the author,
        "VMYs96sn",
        "YT4XzThC",
        "S3RARClD",
        "wM8dXlKT"
        ],
        "outliers":[ # Papers wrongly assigned to the
 author,
        "OTvYjfnt",
        "EzzruFin"
        ]
    },
}
```

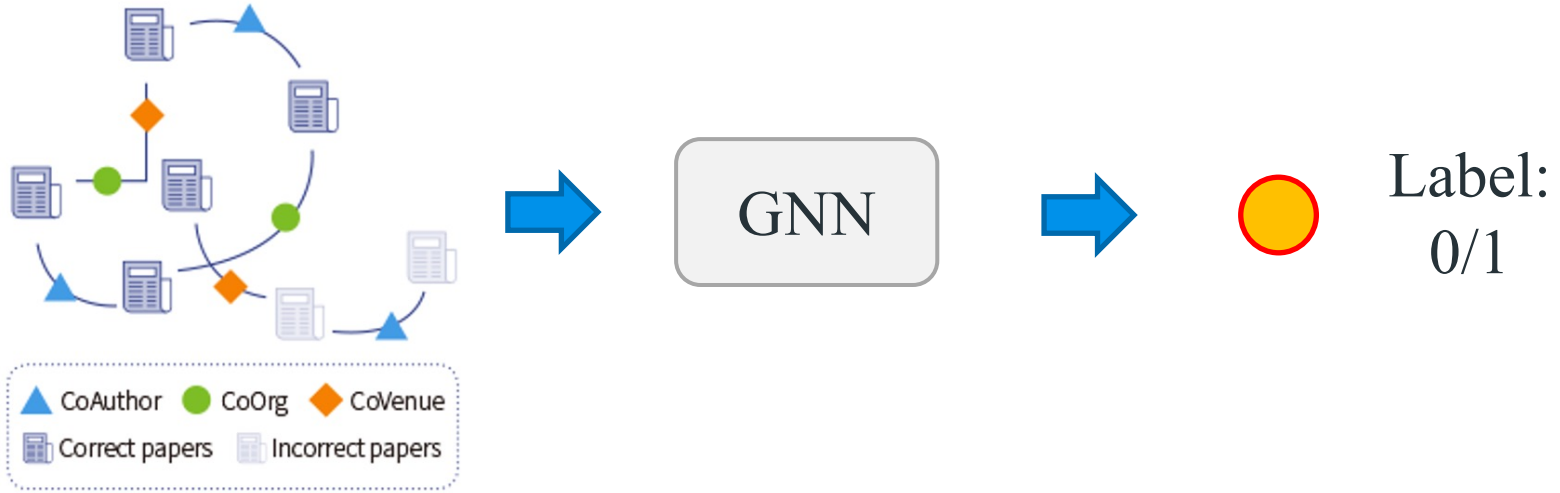
# WhoIsWho-IND: Use of external data and APIs

- The use of any externally labeled data is **NOT allowed** in this competition.
- This competition allows the use of APIs.

- Evaluation
  - Incorrect assignments weighted AUC

# Baselines

- Graph-based anomaly detection methods
  1. Graph Construction: for a specific author, nodes are papers, and edges are co-author/co-organization/co-venue relations between papers
  2. Graph Neural Networks (GNNs): GCN or GCCAD
  3. Node Classification




[1] Thomas N. Kipf, Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. ICLR 2017.
[2] Bo Chen, Jing Zhang, Xiaokang Zhang, Yuxiao Dong, Jian Song, Peng Zhang, Kaibo Xu, Evgeny Kharlamov, and Jie Tang. "Gccad: Graph contrastive learning for anomaly detection.” TKDE 2022.

# Baselines

- Large language model-based method
  - E.g., LLM: ChatGLM3-6B-32K; instruction tuning method: LoRA; GPU environment: 8 * A100

**Human instruct:** Identify the abnormal text from the text collection according to the following rules:\n Here is a paper collections: ' **{Context Papers}**' \n Does the paper ' **{Target Paper}** ' belong to the main part of these papers, give me an answer between 'yes' or 'no'.
**ChatGLM:** yes.

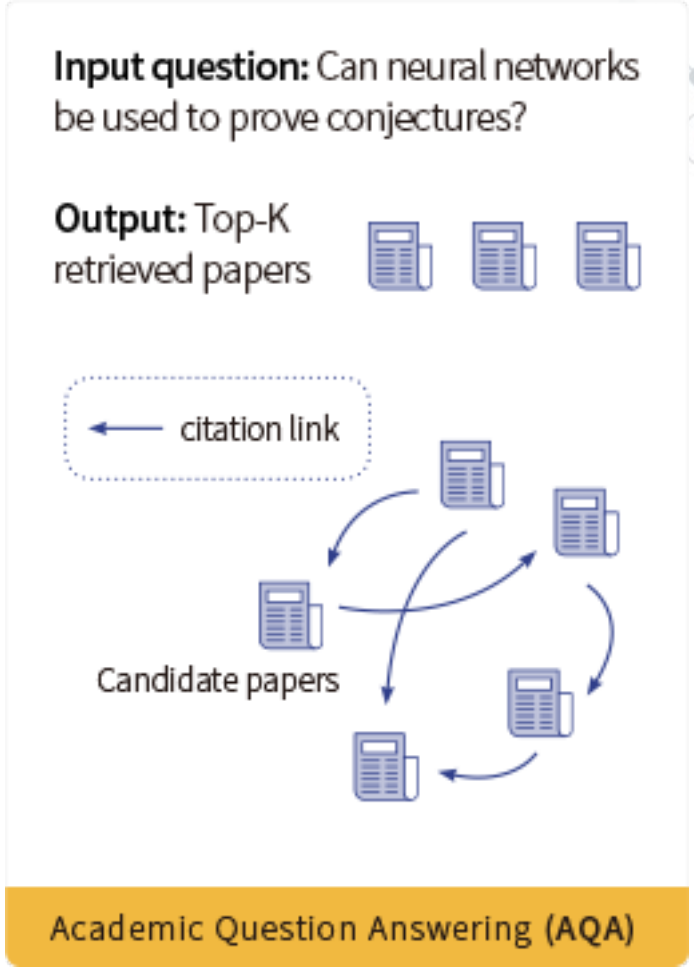
[1] https://github.com/THUDM/ChatGLM3

# Results on Validation Set

| Method | AUC |
| --- | --- |
| GCN | 0.58625 |
| GCCAD | 0.63451 |
| ChatGLM | 0.71385 |

# Task Description

- Track 2: Academic question answering (AQA)
  - Given professional questions and a pool of candidate papers,
  - The objective is to retrieve the most relevant papers to answer these questions.




[1] Weng Tam and Xiao Liu and Kaixuan Ji and Lilong Xue and Jiahua Liu and Tao Li and Yuxiao Dong and Jie Tang. “Parameter-Efficient Prompt Tuning Makes Generalized and Calibrated Neural Text Retrievers.” EMNLP Findings 2023.
[2] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. “OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining”. arXiv:2402.15810.

# Dataset

- OAG-AQA
  - pid_to_title_abs_new.json: [Dictionary] mapping paper id to paper title and abstract
  - qa_train.txt: [JSON lines] question description and ground-truth paper ids
  - qa_valid_wo_ans.txt: questions for prediction
- Total: 10K+ question-paper pairs

| Disc. | #Topic | Example Topic | #Query | Example query-paper pairs |
|---|---|---|---|---|
| Neural Network | 2 | Artificial Neural Network | 488 | **Q: Can neural networks be used to prove conjectures?** Paper: *Generating Correctness Proofs with Neural Networks* |
| Quantum Mechanics | 12 | Photon | 125 | **Q: What is the effective potential for photons in $X$-ray diffraction?** Paper: *Introduction to the theory of x-ray matter interaction* |

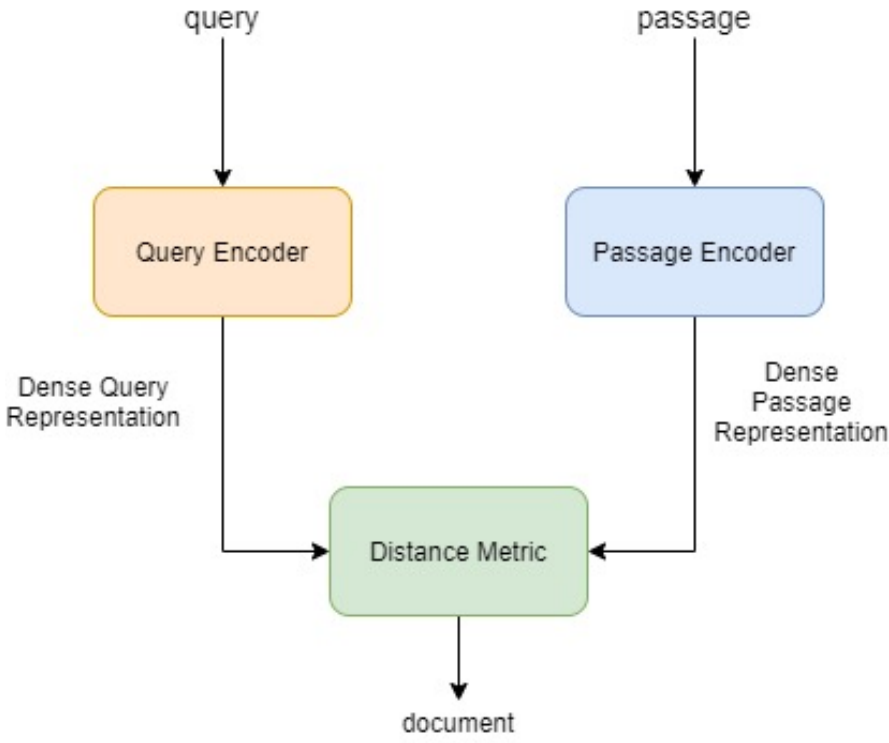
# AQA: Use of external data and APIs

- Participants are allowed to use the large-scale open academic graph dataset OAG and the paper citation dataset DBLP Citation.
- Participants are **NOT allowed** to use data other than those mentioned above.
- The use of any APIs is **NOT allowed**.

- Evaluation
  - Average MAP for ranked top-20 papers

# Baselines

- Dense retrieval
  - Constructing positive and negative QA pairs
  - Dense retrieval models: e.g. DPR
  - Training objective function: cross-entropy loss
  - Inference: similarity search vis FAISS
  - Results on valuation set: 0.16909

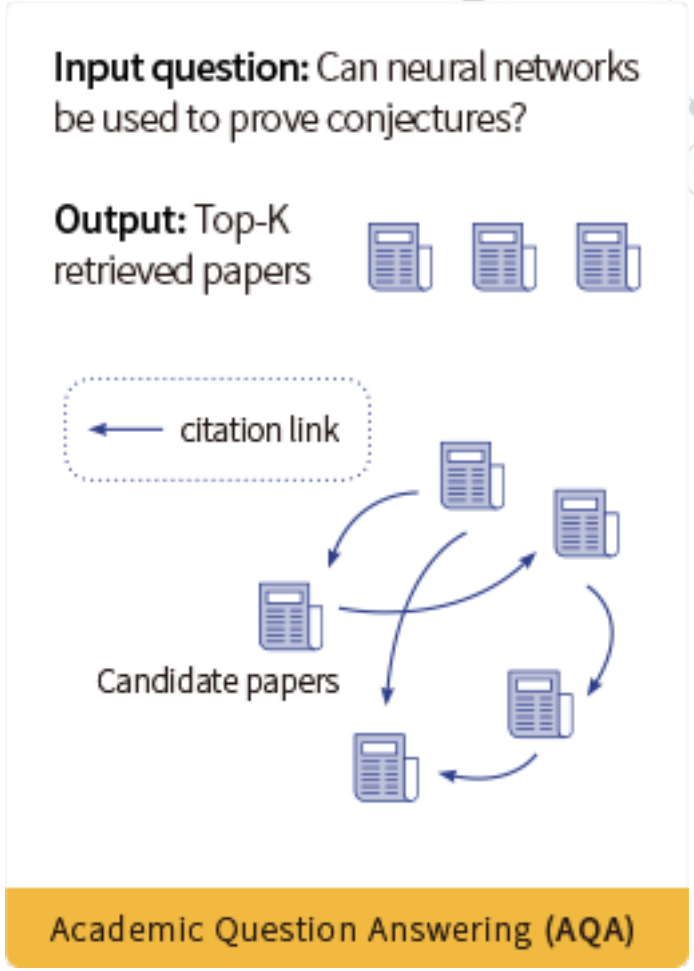



- Sparse retrieval
  - Preliminary experiments of BM25 is ineffective. (<5% MAP score)

[1] Vladimir Karpukhin and Barlas Oguz and Sewon Min and Patrick S. H. Lewis and Ledell Wu and Sergey Edunov and Danqi Chen and Wen-tau Yih. “Dense Passage Retrieval for Open-Domain Question Answering.” EMNLP 2020.

# Task Description

- Track 3: Paper source tracing (PST)
  - Given the full texts of each paper,
  - The goal is to automatically trace the most significant references that have inspired a given paper.




[1] Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, Jie Tang. “PST-Bench: Tracing and Benchmarking the Source of Publications.” arXiv:2402.16009.
[2] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. “OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining”. arXiv:2402.15810.

# Definition of Source Papers

- The following points define whether a reference is a source paper:
  1. Is the main idea of paper p inspired by the reference?
  2. Is the core method of paper p derived from the reference?
  3. Is the reference essential for paper p? Without the work of this reference, paper p cannot be completed.

# PST: Datasets

- Track 3: Paper source tracing (PST)
  - paper_source_trace_train_ans.json: [Dictionary] mapping paper id to paper attributes, using `` referenced_serial_number`` in “refs_trace” field as supervision.
  - paper_source_trace_valid_wo_ans.json: similar to above
  - paper-xml: main text source for prediction
  - paper_source_gen_by_rule.json: additional rule-generated data

- Total: 1.1K+ professionally labeled papers

# FAQ

- How to determine the number of references (n_refs) for the papers in the validation set?
  - Parsing the paper XML files by obtaining the maximum index of bibliography entries
  - If index bib_i (less than n_refs) is an invalid entry in the XML file, the corresponding output score can be set to zero.
- Paper IDs in the papers' "references" field
  - Generally smaller than n_refs
  - Used as supplementary information

# PST: Use of external data and APIs

- Participants are allowed to use the large-scale open academic graph dataset OAG and the paper citation dataset DBLP Citation.
- Participants are **NOT allowed** to use data other than those mentioned above.
- This competition allows the use of APIs.

- Evaluation
  - Average MAP

# Baselines

- Random Forest (RF)
  - Define features including citing count, citing position, author overlap, text similarity, etc.
  - Employ RF to classify the importance of each reference.

- Graph-based method
  - Learn paper embeddings on the paper citation graph
  - Measure the importance of references to the target paper by calculating the cosine similarity between embeddings

# Baselines

- Pretrained Language Model (PLM)-based methods
  - Extract the contextual texts where each reference appears in the full text
  - Encode these texts with the pre-trained models
  - Feed into an MLP classifier for binary prediction

**Target Paper 1**: ProteinBERT: A universal deep-learning model of protein sequence and function
**Ref-source 1**: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences
**Contexts**: … loss continues to improve on the training set (i.e., does not saturate), even after multiple epochs (Fig. 2), in accordance with other studies [20].

| Method | MAP |
|---|---|
| Random Forest | 0.21420 |
| ProNE | 0.21668 |
| SciBERT | 0.29489 |

# Conclusion

- The use of pre-trained language models boosts the performance of the three tasks
- Graph structure also matters
  - Generate negative samples
  - Measure structural high-order similarity
- How to combine PLM-based methods and graph-based methods is an open question

# Resources

- Competition website: https://www.biendata.xyz/kdd2024/

- WhoIsWho-IND Baseline Codes: https://github.com/THUDM/whoiswho-top-solutions/tree/main/incorrect_assignment_detection
- AQA Baseline Codes: https://github.com/THUDM/OAG-AQA
- PST Baseline Codes: https://github.com/THUDM/paper-source-trace/tree/main