Open Academic Graph

Open Academic Graph (OAG) is a large knowledge graph unifying two billion-scale academic graphs: Microsoft Academic Graph (MAG) and AMiner. In mid 2017, we published OAG v1, which contains 166,192,182 papers from MAG and 154,771,162 papers from AMiner (see below) and generated 64,639,608 linking (matching) relations between the two graphs. This time, in OAG v2, author, venue and newer publication data and the corresponding matchings are available.

Overview of OAG v2

The statistics of OAG v2 is listed as the three tables below. The two large graphs are both evolving and we take MAG November 2018 snapshot and AMiner July 2018 or January 2019 snapshot for this version.

Data Set	#Pairs/Venues	Date
Linking relations	29,841	2018.12
AMiner venues	69,397	2018.07
MAG venues	52,678	2018.11

Table 1: statistics of OAG venue data

Data set	#Pairs/Papers	Date
Linking relations	91,137,597	2018.12
AMiner papers	172,209,563	2019.01

MAG papers	208,915,369	2018.11

Table 2: statistics of OAG paper data

Data set	#Pairs/Authors	Date
Linking relations	1,717,680	2019.01
AMiner authors	113,171,945	2018.07
MAG authors	253,144,301	2018.11

Table 3: statistics of OAG author data

Downloads

venue_linking_pairs.zip	paper_linking_pairs.zip	author_linking_pairs.zip

aminer_venues.zip	mag_venues.zip

mag_papers_0.zip	mag_papers_1.zip	mag_papers_2.zip

aminer_papers_0.zip	aminer_papers_1.zip
aminer_papers_2.zip	aminer_papers_3.zip

mag_authors_0.zip	mag_authors_1.zip	mag_authors_2.zip

aminer_authors_0.zip	aminer_authors_1.zip

aminer_authors_2.zip	aminer_authors_3.zip

Please note that for author matching, we only consider authors whose paper count is not less than 5. After filtering those authors with small paper count, there are 6,855,193 authors in AMiner and 13,173,936 authors in MAG.

Data description

For linking relations, each pair is an “ID to ID” pair. More specifically, its JSON schema is:

{
“mid”: “xxxx”,
“aid”: “yyyy”
}

where “mid” is MAG entity ID and “aid” is AMiner entity ID.

Other entity attributes are also provided, which can be used to do different types of research. The data schemas of venues, papers and authors are described as below:

Field Name	Field Type	Description	Example
id	string	venue id	5bf574641c5a1dcdd96f817b
JournalId	string	journal id	137773608
ConferenceId	string	conference id	
DisplayName	string	venue name	Nature
NormalizedName	string	normalized venue name	nature

Table 4: venue schema

Field Name	Field Type	Description	Example
id	string	paper ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han

author.org	string	author affiliation	Department of Computer Science, University of Illinois at Urbana-Champaign
author.id	string	author ID	53f42f36dabfaedce54dcd0c
venue.id	string	paper venue ID	53e17f5b20f7dfbc07e8ac6e
venue.raw	string	paper venue name	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	[“data mining”, “structured data”, “world wide web”, “social network”, “relational data”]
n_citation	int	citation number	40829
page_start	string	page start	11
page_end	string	page end	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	[“http://dx.doi.org/10.4114/ia.v10i29.873”, “http://polar.lsi.uned.es/revista/index.php/ia/article/view/479”]
abstract	string	abstract	Our ability to generate...

Table 5: venue schema

(Note: MAG paper attributes are incomplete. If you would like more attributes, please refer to this link to get MAG on Azure.)

Field Name	Field Type	Description	Example
id	string	author id	53f42f36dabfaedce54dcd0c
name	string	author name	Jiawei Han
normalized_name	string	normalized author name	jiawei han
orgs	list of strings	author affiliations	["Department of Computer Science, University of Illinois at Urbana-Champaign"]
org	string	last known affiliation	Department of Computer Science, University of Illinois at Urbana-Champaign
position	string	author position	professor
n_pubs	int	the number of author publications	1217
n_citation	int	author citation count	191526
h_index	int	author h-index	175
tags.t	string	research interests	"data mining"
tags.w	int	weight of interests	243
pubs.i	string	author paper id	53e9b9fbb7602d97045f7bb8
pubs.r	int	author order in the paper	0

Table 6: author schema

Evaluation

We evaluated a small subset of matchings (around one thousand venue/paper/author pairs). The estimated accuracy is shown in Table 7.

entity type	venue	paper (newly matched)	author
accuracy	99.26%	99.10%	97.41%

Table 7: accuracy of entity matching

Please continue to use the OAG group to discuss OAG related problems. If you have any issue or want to contribute to Open Academic Graph, feel free to share your ideas and thoughts at OAG Group.

Overview of OAG v1

This data set is generated by linking two large academic graphs: **Microsoft Academic Graph (MAG)** and **AMiner**, and it is used for research purpose only. This version includes **166,192,182** papers from MAG and **154,771,162** papers from AMiner. We generated **64,639,608** linking (matching) relations between the two graphs. In the future, more linking results, like authors, will be published. It can be used as a unified large academic graph for studying citation network, paper content, and others, and can be also used to study integration of multiple academic graphs.

The overall data set includes three parts, which are described in the table below:

Data Set	#Paper	#File	Total Size	Date
Linking relations (matching)	64,639,608	1	1.6GB	2017-06-22
MAG papers	166,192,182	9	104GB	2017-06-09
AMiner papers	154,771,162	3	39GB	2017-03-22

Downloads

AMiner Papers:

Data Set	#Paper	#File
aminer_papers_0.zip	aminer_papers_1.zip	aminer_papers_2.zip

MAG Papers:

Data Set	#Paper	#File

mag_papers_0.zip	mag_papers_1.zip	mag_papers_2.zip
mag_papers_3.zip	mag_papers_4.zip	mag_papers_5.zip
mag_papers_6.zip	mag_papers_7.zip	mag_papers_8.zip

Data Description

The detailed description of data is presented in this section.

For **Linking relations**, each linking pair is an "ID to ID" pair. More specifically, its JSON schema is:

```
{
  "mid": "xxxx",
  "aid": "yyyy"
}
```

where "mid" is MAG paper ID and "aid" is AMiner paper ID.

For data set **MAG papers** and **AMiner papers**, each paper is a JSON object. Its data schema is:

Field Name	Field Type	Description	Example
id	string	MAG or AMiner ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	department of computer science university of illinois at urbana champaign
venue	string	paper venue	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos	list of strings	fields of study	["relational database", "data model", "social network"]
n_citation	int	number of citation	29790

references	list of strings	citing papers' ID	["53e99ef4b7602d97027c2346", "53e9aa23b7602d970338fb5e", "53e99cf5b7602d97025aac75"]
page_stat	string	start of page	11
page_end	string	end of page	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.lsi.uned.es/revista/index.php/ia/article/view/479"]
abstract	string	abstract	Our ability to generate...

For example:


```
{
  "id": "53e9ab9eb7602d970354a97e",
  "title": "Data mining: concepts and techniques",
  "authors": [
    {
      "name": "jiawei han",
      "org": "department of computer science university of illinois at ur
    },
    {
```

```
      "name": "micheline kamber",
      "org": "department of computer science university of illinois at ur
    },
    {
      "name": "jian pei",
      "org": "department of computer science university of illinois at ur
    }
  ],
  "year": 2000,
  "keywords": [
    "data mining",
    "structured data",
    "world wide web",
    "social network",
    "relational data"
  ],
  "fos": [
    "relational database",
    "data model",
    "social network"
  ],
  "n_citation": 29790,
  "references": [
    "53e99ef4b7602d97027c2346",
    "53e9aa23b7602d970338fb5e",
    "53e99cf5b7602d97025aac75"
  ],
  "doc_type": "book",
  "lang": "en",
  "publisher": "Elsevier",
  "isbn": "1-55860-489-8",
  "doi": "10.4114/ia.v10i29.873",
  "pdf": "//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d97035
  "url": [
    "http://dx.doi.org/10.4114/ia.v10i29.873",
    "http://polar.lsi.uned.es/revista/index.php/ia/article/view/479"
  ],
  "abstract": "Our ability to generate and collect data has been increasi
}
```

Method and Evaluation

Method

We obtain linking relations of two publication graphs by two steps:

1. Use Microsoft Graph Search API to query each AMiner paper's title and obtain candidate matching papers for each AMiner paper.
2. We match two papers if they have
   - very similar titles
   - similar author names and
   - same published year

Evaluation

We random sampled **100,000** linking pairs and evaluated the matching accuracy. The number of truly matching pairs is **99,699** and the matching accuracy can achieve **99.70**%.

Reference

We kindly request that any published research that makes use of this data cites the following papers.

- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998. [PDF] [Slides] [System] [API]
- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246. [PDF][System][API]