DATA DESCRIPTION

Schema for patent citations to science (PCS) output files

The main output file, available at **http://relianceonscience.org**, is called *pcs.tsv* and is a tab-separated file containing the patent number, the unique identifier in the MAG database, confidence score, and whether the reference was filed by the applicant, an examiner, or other (if known). It contains PCS links of confidence score 3 or higher. Those using this data are asked to cite this paper. The schema is as follows:

Table A1.1: Contents of *pcs.tsv*.

Variable	Type	Notes
reftype	string	App = from applicant Exm=from examiner (Note: all references from non-USPTO patents are labeled as examiner unless otherwise indicated in the unstructured reference.) Unk = if unspecified in the unstructured reference (Note: almost every USPTO reference before 2006 is of unknown origin.)
confscore	numeric	Assigned confidence score to the match. Note that only matches with a confidence score of 3 or above are included in the distribution.
magid	numeric	Unique identifier for each paper in the Microsoft Academic Graph.
patent	string	Only patents for which our algorithm established a PCS linkage are included. Non-USPTO patents have a country prefix followed by a dash at the beginning of the patent number.

As described in the body of the paper, PCS are established via a probabilistic algorithm. Users of the data should consult Tables 2 and 3 as well as Figure 1 to determine their desired confidence-score cutoff. Matches for confidence scores 2 and 1 are not included in the distribution as there are very few correct matches at those levels. Even at confidence score 3, about half of the matches are incorrect. Most users will want to only use matches with a score of 4 or higher.

Files for Microsoft Academic Graph metadata

Also available is a series of files with metadata regarding not just the references reported in Appendix 1 but *all* papers in the 1 January 2019 release of the Microsoft Academic Graph (MAG). They are compressed using the 'zip' utility under Unix CentOS5. Reposting of these data is facilitated by the ODC-By license (https://opendatacommons.org/licenses/by/1-0/index.html), under which MAG is provided and under which these data are also provided. Those using these data should cite the following paper: *Sinha, Arnab, 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.*

Researchers who prefer to download the original MAG data directly from Microsoft can do so by signing up for an Azure account and billing plan, contacting Microsoft for access to MAG, selecting the 2019-1-1 release, and downloading the desired files. Instructions are at https://docs.microsoft.com/en-us/academic-services/graph/. Note however that some of the original MAG files are several dozen gigabytes in size; for example, the Papers.txt file from which several of these files are derived, is 56 gigabytes.

All files are in tab-separated format, compressed as .zip files. The first set of files contain direct metadata for papers in MAG.

Filename	Variables	MAG file (fields)	Notes
paperyear	paperid, paperyear	Papers.txt (1,8)	
papervolisspages	paperid, papervolume, paperissue, paper1stpage, paperlastpage	Papers.txt (1,14,15,16,17)	Issue and pages are sometimes blank. First page is available more often than last page.
papertitle	paperid, papertitle	Papers.txt (1,5)	Titles are often blank for conference papers.
papercitations	citingpaperid, citedpaperid	PaperReferences.txt (1,2)	Adds headings to PaperReferences.txt.
paperdoi	paperid, doi	Papers.txt (1,3)	DOI is not available for every paper in MAG
paperauthororder	paperid, authorid, authororder	PaperAuthorAffiliations.txt (1,2,4)	Author order not available for every author
paperauthoraffiliationname	paperid, authorid, affiliationname	PaperAuthorAffiliations.txt (1,2,5)	Affiliation not available for many authors

The next set of files contain indirect metadata, i.e. identifiers that need to be matched to dictionaries in the next set of files. One could provide the full strings of the authors, journals, etc., directly but the files would be much larger and unnecessarily redundant.

Filename	Variables	MAG file (fields)	Notes
paperconferenceid	paperid, conferenceid	Papers.txt (1,13)	
paperfieldid	paperid, fieldid	PaperFieldsOfStudy.txt (1,2)	ID for field of paper.
paperjournalid	paperid, journalid	Papers.txt (1,11)	

The third set of files contains the string values for indirect metadata identifiers:

Filename	Variables	MAG source (fields)	Notes
authoridname_normalized	authorid, authorname_normalized	Authors.txt (1,3)	Lowercase name w/o punctuation.
authoridname_raw	authorid, authorname_raw	Authors.txt (1,4)	As originally appeared.
conferenceidname	conferenceid conferencename	ConferenceInstances.txt (1,2)	Name of conference
fieldidname	fieldid fieldname	FieldsOfStudy.txt (1,3)	Paper field, inferred from title+abstract.
journalidname	journalid journalname journalissn	Journals.txt (1,3,5)	ISSN is often unavailable.

Schema for extensions to the Microsoft Academic Graph (MAG) data

In addition to the redistribution of the MAG data, we provide two extensions for fields not present in the MAG data. First, we calculate Journal Impact Factor for all journals in MAG. The schema is as follows:

Contents of *jif.tsv*.

Variable	Type	Notes
journalid	numeric	
journalname	String	
jif	numeric	Journal impact factor. A journal's impact factor is a popular measure of its quality, calculated for year t as the number of times articles from years t-1 and t-2 were cited *by other articles* during year t, divided by the number of articles published during years t-1 and t-2.

In addition, we provide a new measure of journal impact: Journal Commercial Impact Factor (JCIF). Just like JIF is a journal-level measure of quality, it is possible to build a journal-level measure of appliedness or commercial relevance by replacing paper-to-paper citations by patent-to-paper citations. Bikard and Marx (2019) introduced this concept and calculated it for the Web of Science; here, we calculate JCIF for MAG. That paper should be cited if the JCIF data available here are used.

Contents of *jcif.tsv*.

Variable	Type	Notes
journalid	numeric	
journalname	String	
jcif	numeric	Journal commercial impact factor. A journal's commercial impact factor is calculated for year t as the number of times articles from years t-1 and t-2 were cited *by patents* during year t, divided by the number of articles published during years t-1 and t-2.

Finally, we provide an aggregation of the more than 200,000 fields automatically extracted from the papers themselves. We mapped the MAG subjects to 6 OECD fields and 39 subfields, defined here: http://www.oecd.org/science/inno/38235147.pdf. Clarivate provides a crosswalk between the OECD classifications and Web of Science fields, so we include WoS fields as well. This file is magfield_oecd_wos_crosswalk.zip.

Contents of *magfield_oecd_wos_crosswalk.tsv*.

Variable	Type	Notes
paperid	numeric	Unique identifier for each paper in the Microsoft Academic Graph.
paperfieldid	paperid, fieldid	PaperFieldsOfStudy.txt (1,2)
oecd_field	String	One of six top-level OECD fields.
oecd_subfield	String	One of 39 OECD subfields.
wosfield	String	One of 251 Web of Science fields.