DATA DESCRIPTION

Schema for patent citations to science (PCS) output files

The main output file, available at **http://relianceonscience.org**, is called *_pcs_mag_doi_pmid.tsv* and is tab-separated. Each record contains a patent-to-article citation established by our algorithm.

Contents of *_pcs_mag_doi_pmid.tsv*.

Variable	Type	Notes
reftype	string	**App** = from applicant **Exm**=from examiner (Note: non-USPTO refs are examiner unless otherwise indicated in the reference.) **Unk** = if unspecified in the unstructured reference (Note: most pre-2006 USPTO references are unkown.)
confscore	numeric	Assigned confidence score to the match.
magid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
doi	string	Digital Object Identifier as provided by Microsoft
pmid	numeric	PubMed ID as provided by Microsoft
patent	string	Only patents for which our algorithm established a PCS linkage are included.
wherefound	string	frontonly, bodyonly, or both (i.e., both on the front page of the patent, and also in the body text)
uspto	binary	Indicates whether the patent is from the USPTO. International patents start with a two-letter code for the granting office. These can be matched up by patent family using *intlpatfamily.tsv*.

The set of known-good patent-to-article citations is called *bodytextknowngood.tsv* and is tab-separated. Each record is a true patent-to-article citations that was verified by at least two research assistants.

Contents of *bodytextknowngood.tsv*.

Variable	Type	Notes
patent	string	Patent in which the in-text reference was found. Each reference
magid	numeric	Unique identifier for the paper cited in the Microsoft Academic Graph
doi	string	Digital Object Identifier as provided by Microsoft, if available
pmid	numeric	PubMed ID as provided by Microsoft, if available

The set of in-text patent-to-patent citations is called *bodytextpatrefstopatents.tsv* and is tab-separated. Each record is a patent-to-patent established by our algorithm.

Contents of *bodytextknowngood.tsv*.

Variable	Type	Notes
Citingpatent	string	Patent containing the citation in its body text
Citedpatent	String	Patent cited in the body text of Citingpatent
Citingpatentyear	Int	Year of Citingpatent
Citedpatentyear	Int	Year of Citedpatent

Contents of *intlpatfamily.tsv*.

Variable	Type	Notes
patent	string	All non-USPTO patents available in DOCDB with a family ID are included.
Family	numeric	Patent family number.

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 2020 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 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, whereas we have partitioned the files into smaller pieces for convenience. 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 a categorization of scientific fields per paper at a high level. Microsoft automatically extracts more than 200,000 fields from the abstracts and titles of 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.