README FILE for Journal Pair Metrics dataset
Jennifer D'Souza and Neil R. Smalheiser, August 5 2014

This dataset comes in the form of 3 separate compressed (.gz) files named journalPair_metrics.txt, journal_features.txt, and journal_features.xlsx. Uncompressed sizes of the tab-delimited files should be ~4GB, ~1.0MB, and ~1.1MB, respectively. These data are being released under the terms of the Creative Commons **Attribution-NonCommercial-ShareAlike CC BY-NC-SA** International Public License 4.0. (https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode).

How was the dataset created?

This dataset describes various characteristics, features and metrics applied to journals and pairs of journals. The criteria for a journal to be included in this dataset is that it is indexed in PubMed, has at least 100 articles and 10 predicted author-individuals in the 2009 Author-ity author name disambiguation dataset(see below), and is indexed with at least 50 major MeSH terms in MEDLINE (2012 baseline data release). We then consider pairs of these journals; the criteria for a pair to be included is that they share at least one author as predicted by the 2009 Author-ity dataset. See References 7 and 8 for details on the Author-ity dataset.

What are the metrics and how are they defined?

The 2009 Author-ity dataset is based on a snapshot of PubMed (which includes both MEDLINE and PubMed-not-MEDLINE records) taken in July 2009, including a total of 19,011,985 Article records, 61,658,514 author name instances and 20,074 unique journal names. Each instance of an author name is uniquely represented by the PMID and the position on the paper (e.g., 10786286_3 is the third author name on PMID 10786286). Thus, each predicted author-individual is associated with a list of predicted PMIDs written by that individual.

For all metrics, we filtered the set of journals to retain only the 9,284 journals that have at least 100 articles within the Author-ity dataset, in order to ensure statistical robustness and to exclude journals that are very old or very new. This generated 13,129,909 pairs of journals which had at least one predicted author-individual in common. To characterize the author metric, this set was filtered further to retain only those journals which have at least 10 distinct authors in Author-ity, giving 9,281 unique journals and 13,129,781 journal pairs. To characterize the MeSH metric, the original set was filtered to retain only journals having at least 50 major MeSH terms within MEDLINE (2012 release), giving 8,900 journals and 12,569,613 journal pairs. Finally, to compare the metrics, we created a single filtered set of journals, having at least 100 articles and 10 authors within Author-ity, and at least 50 major MeSH terms in MEDLINE, giving 8,897 journals and 12,569,485 journal pairs.

MeSH based metric. For each journal, we identified all PMIDs published in that journal that were included in the 2012 baseline version of MEDLINE, and made a list of all major Medical Subject Headings (MeSH) associated with the articles. To compare two journals pairwise with regard to MeSH term similarity, we scored the number of MeSH terms in common between the two journals, giving larger weight to terms that occurred in multiple articles. For example, if a given MeSH term occurred in 3 articles in Journal 1 (J1) and in 10 articles in Journal 2 (J2), we would score this term as having weight 3. The final MeSH normalized co-occurrence score (Co) is the weighted sum over all MeSH terms, normalized by journal size (geometric mean of the total number of unique MeSH terms in each journal) (Figure 1A). In order to ensure that the metric would be robust and meaningful, comparisons were only made for journals that had at least a total of 50 major MeSH headings.

$$\mathrm{MeSH}_{\mathrm{Cooccurrence}_{ob}}(J_i, J_j) = \mid MeSH_T(J_i) \cap MeSH_T(J_j) \mid$$

$$MeSH_{size}(J_i, J_j) = \sqrt{\mid MeSH_{UT}(J_i) \mid \times \mid MeSH_{UT}(J_j)}$$

$$MeSH_{Co}(J_i, J_j) = \frac{MeSH_{Cooccurrence_{ob}}(J_i, J_j)}{MeSH_{size}(J_i, J_j)}$$

where, $MeSH_T(J)$ are all major MeSH terms in journal J, and $MeSH_{UT}(J)$ are the unique major MeSH terms in journal J.

Next, we computed the Co score that would be expected simply by chance (for two journals of their size). This was done by ranking all journal pairs by journal size, dividing into bins of 5,000 pairs (each having roughly the same journal size), and calculating the average Co score across all journal pairs in the same bin. Finally, we calculated the MeSH odds ratio for each pair of journals present in that bin, by dividing the observed Co score divided by the Co score expected by chance.

$$Bin_j = \left\{ \left(J_{m_j}, J_{n_j}\right) \mid \forall \left(J_{mi}, J_{n_i}\right) \in Bin_i \wedge \left(i < j \wedge MeSH_{size}\left(J_{m_i}, J_{n_i}\right) < MeSH_{size}\left(J_{m_j}, J_{n_j}\right)\right) \wedge \right.$$
$$\left. \forall \left(J_{m_k}, J_{n_k}\right) \in Bin_k \wedge \left(j < k \wedge MeSH_{size}\left(J_{m_j}, J_{n_j}\right) < MeSH_{size}\left(J_{m_k}, J_{n_k}\right)\right) \right\}$$

$$MeSH_{Cooccurrence_{\exp}}\left(Bin_j\right) = \frac{\sum_{\left(J_{m_j}, J_{n_j}\right) \in Bin_j} MeSH_{Cooccurrence_{ob}}\left(J_{m_j}, J_{n_j}\right)}{\mid Bin_j \mid}$$

where, $\mid Bin \mid = 5000$ and $MeSH_{Cooccurrence_{\exp}}\left(Bin_j\right)$ is the major MeSH term co-occurrence expected by chance for all $\left(J_{m_j}, J_{n_j}\right) \in Bin_j$.

$$MeSH_{odds}\left(J_i, J_j\right) = \frac{MeSH_{Cooccurrence_{ob}}\left(J_i, J_j\right)}{MeSH_{Cooccurrence_{\exp}}\left(J_i, J_j\right)}$$

Author based metric. Using the Author-ity 2009 dataset as a gold standard, we scored the author Co score as the number of predicted author-individuals in common between each pair of journals, normalized by journal size ((geometric mean of the total number of unique author-individuals publishing in each journal).

$$\text{Author}_{\text{Cooccurrence}_{ob}}(J_i, J_j) = |\, Author_I(J_i) \cap Author_I(J_j)\,|$$

$$Author_{size}(J_i, J_j) = \sqrt{|\, Author_I(J_i)\,| \times |\, Author_I(J_j)}$$

$$Author_{Co}(J_i, J_j) = \frac{Author_{Cooccurrence_{ob}}(J_i, J_j)}{Author_{size}(J_i, J_j)}$$

where, $Author_I(J)$ are the unique author-individuals journal J.

Next, we computed the author Co score that would be expected simply by chance (for two journals of their size). This was done by dividing all journal pairs into bins of 5,000 pairs, each having roughly the same journal size, and calculating the average number of author-individual co-occurrences across all journal pairs in the same bin. Finally, we calculated the author odds ratio for each pair of journals, by dividing the observed Co score divided by the Co score expected by chance. In order to ensure that the metric would be robust and meaningful, comparisons were only made for journals that had at least a total of 10 or more predicted author-individuals.

$$Bin_j = \left\{ \begin{array}{l} \left(J_{m_j}, J_{n_j}\right) |\, \forall \left(J_{mi}, J_{n_i}\right) \in Bin_i \wedge \left(i < j \wedge Author_{size}\left(J_{m_i}, J_{n_i}\right) < Author_{size}\left(J_{m_j}, J_{n_j}\right)\right) \wedge \\ \forall \left(J_{m_k}, J_{n_k}\right) \in Bin_k \wedge \left(j < k \wedge Author_{size}\left(J_{m_j}, J_{n_j}\right) < Author_{size}\left(J_{m_k}, J_{n_k}\right)\right) \end{array} \right\}$$

$$Author_{Cooccurrence_{\exp}}\left(Bin_j\right) = \frac{\sum_{\left(J_{m_j}, J_{n_j}\right) \in Bin_j} Author_{Cooccurrence_{ob}}\left(J_{m_j}, J_{n_j}\right)}{|\, Bin_j\,|}$$

where, $|\, Bin\,| = 5000$ and $Author_{Cooccurrence_{\exp}}\left(Bin_j\right)$ is the author co-occurrence expected by chance for all $\left(J_{m_j}, J_{n_j}\right) \in Bin_j$.

$$Author_{odds}\left(J_i, J_j\right) = \frac{Author_{Cooccurrence_{ob}}\left(J_i, J_j\right)}{Author_{Cooccurrence_{\exp}}\left(J_i, J_j\right)}$$

Article pair based metric. Using the Author-ity 2009 dataset, we compiled all article pairs that co-occur within each author-individual cluster, i.e., that are predicted to be written by the same author-individual, and compiled a list of all such article pairs across all individuals. After mapping the PMIDs of these co-occurring article pairs to their journals, we obtained 13,129,909 unique journal pairs -- including pairs in which both articles were published in the same journal, which allowed us to assess the relative tendency of individuals to publish repeatedly in the same journal over time. The article pair Co score equals the total number of co-occurrences for that journal pair divided by the geometric mean of the journal sizes (i.e., total number of articles that map to each journal).

$$Article\_Pair_{Cooccurrence_{Ob}}\left(J_i, J_j\right) = |\, Article\left(Author_I\left(J_i, J_j\right)\right)|^2 - |\, Article\left(Author_I\left(J_i, J_j\right)\right)|$$

$$Article\_Pair_{size}(J_i, J_j) = \sqrt{|\, Article(J_i)\,| \times |\, Article(J_j)}$$

$$Article\_Pair_{Co}(J_i, J_j) = \frac{Article\_Pair_{Cooccurrence_{ob}}(J_i, J_j)}{Article\_Pair_{size}(J_i, J_j)}$$

where, $Author_I(J_i, J_j)$ is the set of author-individuals who have published articles in both journals J_i and J_j, and $Article(Author_I(J_i, J_j))$ is the list of all articles from either J_i or J_j by authors that the two journals share in common, and $Article(J)$ are all the articles in journal J. In the mathematical formula for observed article pair co-occurrence, $|Article(Author_I(J_i, J_j))|^2$ gives the total number of article pair combinations from the two journals where the articles are authored by author-individuals the two journals share in common. Since the count also includes an article paired with itself, subtracting by the total number of the articles i.e. $|Article(Author_I(J_i, J_j))|$ removes the added effect.

Next, we computed the article pair Co score that would be expected simply by chance (for two journals of their size). This was done by dividing all journal pairs into bins of 5,000 pairs, each having roughly the same journal size, and calculating the average Co score across all journal pairs in the same bin. Finally, we calculated the article pair odds ratio for each pair of journals, by dividing the observed Co score divided by the Co score expected by chance. In order to ensure that the metric would be robust and meaningful, comparisons were only made for journals that had either 10 or more observed article pair co-occurrences, or 10 or more co-occurrences expected by chance.

$$Bin_j = \left\{ \begin{array}{l} \left(J_{m_j}, J_{n_j}\right) | \forall \left(J_{mi}, J_{n_i}\right) \in Bin_i \wedge \left(i < j \wedge Article\_Pair_{size}\left(J_{m_i}, J_{n_i}\right) < Article\_Pair_{size}\left(J_{m_j}, J_{n_j}\right)\right) \wedge \\ \forall \left(J_{m_k}, J_{n_k}\right) \in Bin_k \wedge \left(j < k \wedge Article\_Pair_{size}\left(J_{m_j}, J_{n_j}\right) < Article\_Pair_{size}\left(J_{m_k}, J_{n_k}\right)\right) \end{array} \right\}$$

$$Article\_Pair_{Cooccurrence_{\exp}}\left(Bin_j\right) = \frac{\sum_{\left(J_{m_j}, J_{n_j}\right) \in Bin_j} Article\_Paie_{Cooccurrence_{ob}}\left(J_{m_j}, J_{n_j}\right)}{|Bin_j|}$$

where, $|Bin| = 5000$ and $Article\_Pair_{Cooccurrence_{\exp}}\left(Bin_j\right)$ is the article pair co-occurrence expected by chance for all $\left(J_{m_j}, J_{n_j}\right) \in Bin_j$.

$$Article\_Pair_{odds}\left(J_i, J_j\right) = \frac{Article\_Pair_{Cooccurrence_{ob}}\left(J_i, J_j\right)}{Article\_Pair_{Cooccurrence_{\exp}}\left(J_i, J_j\right)}$$

Disciplines. The 2011 release of the Text Categorization (TC) toolkit developed at NLM (http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/tc/2011/web/index.html) was used to annotate journals with their disciplines. For most journals, pre-computed journal descriptor annotations were available; 300 journals were not included, and for these we computed annotations

directly from the Journal Descriptor Indexing tool included in the TC package. The tool takes as input a journal's MeSH terms (both major and non-major terms were included), and based on a similarity measure, provides a ranked list of disciplines from most to least similar. We used the highest ranked discipline as the journal's discipline (though the pre-computed annotations sometimes consisted of two or three categories simultaneously). These discipline categories, which are thus MeSH terms, are selected from a pre-created list of 122 unique categories [3]. For the most part, the JDI categories were similar to those assigned in the NLM catalog and corresponded to common sense. However, JDI had limited coverage in non-biomedical fields. For example, Astrophysical Journal was indexed as Chemistry, Analytical. Physical Review Letters was indexed as Medicine, and Los Alamos Science as Environmental Health. We manually changed the latter two to Physics, which otherwise does not exist as a JDI category. Because of these and other errors, JDI discipline categories were simply used for display purposes and were not incorporated into any metrics. Nevertheless, we preferred use of JDI instead of NLM categories (which did not cover all journals in Author-ity) or Web of Science categories (which are manually assigned rather than computed according to a standard terminology and reproducible algorithm [9]).

What is in the dataset?

The dataset consists of two files: File 1 contains metrics related to journal pairs (data in file journalPairs.txt); and Part II contains metrics related to individual journals (data provided in two different file formats viz. as text file called journal_features.txt and as an excel sheet called journal_features.xls).

What is the format of the data?

File I: journalPair.txt

Each line in the file corresponds to a pair of journals J1 and J2, including in some cases J1 = J2.
The file includes 7 MeSH based metrics, 7 author based metrics, and 7 article pair based metrics, as indicated by three super-headers.

1. Data lines are tab-delimited for major fields and colon-delimited for paired values within a field.
2. Columns with paired information are "journal pair", "# MeSH terms", "# authors", and "# articles".
3. Details about values by column:
   a. journal pair - is a pair of journals (represented as J1:J2) that satisfy the criteria for inclusion stated above.
   b. # MeSH terms - is a paired value (represented as value1:value2), where value1 is equal to the total unique major MeSH terms in J1's articles from MEDLINE, and value2 are the total unique major MeSH terms in J2's articles from MEDLINE.

c. geomean # MeSH terms - is the geometric mean of the two values in column b.
d. # common MeSH terms - are the number of major MeSH terms in common between J1 and J2 where each MeSH term is weighted by its frequency of occurrence either in J1 or in J2 (whichever is smaller).
e. MeSH Co score - value in column d divided by value in column c.
f. # expected common MeSH terms - is an estimate of the expected number of common MeSH terms for the journal pair based on the geometric mean of the size of the journals. (see method section above).
g. expected MeSH Co score - value in column f divided by value in column c.
h. MeSH odds-ratio - value in column d divided by value in column f.
i. # authors - is a paired value (represented as value1:value2), where value1 is the number of unique predicted author-individualss who have published articles in J1, and value2 is the number of unique authors who have published articles in J2.
j. geomean # authors - is the geometric mean of the two values in column i.
k. # common authors - are the number of unique authors in common between J1 and J2.
l. author Co score - value in column k divided by value in column j.
m. # expected common authors - is an estimate of the expected number of common authors for the journal pair based on the geometric mean of the size of the journals. (see methods, above).
n. expected author Co score - value in column m divided by value in column j.
o. author odds-ratio - value in column k divided by value in column m.
p. # articles - is a paired value (represented as value1:value2), where value1 is the number of unique articles in J1 from Author-ity, and value2 is the number of unique articles in J2.
q. geomean # articles - is the geometric mean of the two values in column p.
r. # article pairs with common authors - is the number of unique article pairs formed by pairing J1's articles with J2's articles that have been published by common authors in J1 and J2. For example, if the same author-individual published 3 articles in J1 and 10 articles in J2, then the number of article pairs in J1:J1 is 3 (= 3 choose 2), the number of pairs in J2:J2 is 45 (= 10 choose 2) and the number of pairs in J1:J2 is 30 (= 3*10).
s. article pair Co score - value in column r divided by value in column q.
t. # expected article pairs with common authors - is an estimate of the expected number of common article pairs for the journal pair based on the geometric mean of the size of the journals. (see methods, above).

u. expected article pair Co score - value in column t divided by value in column q.
v. article pair odds-ratio - value in column r divided by value in column t.

File II: journal_features.txt and journal_features.xlsx

Each line in the tab-delimited file corresponds to a journal with subsequent numeric metrics characterizing the journal as per its: i) JDI-discipline annotation, ii) broadness, iii) MeSH cloud, iv) author cloud, and v) author cloud / MeSH cloud ratio.

MeSH cloud: For each journal, one can envision that there is a "cloud" of other journals which are topically related to it more than expected by chance. That is, for each journal J1 one can count the number of journals Jx for which the MeSH odds ratio for the journal pair J1:Jx is greater than 1. Author cloud: That is, for each journal J1 one can count the number of journals Jx for which the author odds ratio for the journal pair J1:Jx is greater than 1. This measures the tendency of authors who publish in J1 to publish in other specific journals Jx as well.

1. Data lines are tab-delimited for major fields.
2. Details about values column-wise:
   a. journal - is the ISO abbreviated journal name.
   b. discipline - is a paired value (represented as discipline(s):"discipline size"), where discipline(s) corresponds to the Journal Discriptor Indexing [3] annotation of a journal's discipline with possibly multiple disciplines separated by '|', and discipline size is the number of journals from the dataset with the same discipline. As size of multiple disciplines as a single value is not counted in this study, their size is assigned the value -1.
   c. # articles with MeSH - is the number of the journal's articles that are indexed with MeSH terms in MEDLINE (2012 release).
   d. # MeSH terms - is the count of unique major MeSH terms in the journal.
   e. broadness - value in column d divided by value in column c.
   f. broadness index - value in column e divided by mean of all values in column e.
   g. MeSH cloud - value is the count of journals in a journal's MeSH cloud.
   h. MeSH cloud vs expected by size - difference between the expected size of the MeSH cloud (based on the journal size, using linear regression) and actual observed MeSH cloud. Here journal size is measured as the number of the journal's articles in MEDLINE indexed with MeSH. A positive value means that the MeSH cloud is larger than expected.

i. # articles with authors - is the number of the journal's articles that have listed authors in Author-ity.
j. # authors - is the count of unique authors in the journal.
k. author cloud - value is the count of journals in a journal's author cloud.
l. author cloud vs expected by size - difference between the expected author cloud and actual author cloud. Here journal size is measured as the number of the journal's articles having listed authors in Author-ity. A positive value means the author cloud is larger than expected.
m. author cloud/MeSH cloud - ratio of value in column k and value in column g.

References Cited

1. Boyack KW, Klavans R, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3): 351-374.
2. Leydesdorff L, Goldstone RL (2014) Interdisciplinarity at the journal and specialty level: The changing knowledge bases of the journal Cognitive Science. J Assoc Inf Sci Technol 65(1): 164-177.
3. Humphrey SM, Lu CJ, Rogers WJ, Browne AC (2006) Journal Descriptor Indexing tool for categorizing text according to discipline or semantic type. AMIA Annu Symp Proc (Vol. 2006, p. 960). American Medical Informatics Association.
4. Åström F (2002, July). Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters. Emerging frameworks and methods: Proceedings of the fourth international conference on conceptions of Library and Information Science (CoLIS4) (pp. 185-197).
5. Ni C, Ding Y (2010) Journal clustering through interlocking editorship information. Proceedings of the American Society for Information Science and Technology 47(1): 1-10.
6. Cordier S (2012) A measure of similarity between scientific journals and of diversity of a list of publications. Version 1.0 - Oct. 2012. http://arxiv.org/pdf/1210.6510v1.pdf. Accessed Aug. 4, 2014.
7. Torvik VI, Smalheiser NR (2009) Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data (TKDD) 3(3): 11.
8. Torvik VI, Weeber M, Swanson DR, Smalheiser NR (2005) A probabilistic similarity metric for MEDLINE records: a model for author name disambiguation. J Assoc Inf Sci Technol 56(2): 140-158.
9. Leydesdorff L, Bornmann L (2014) The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies". http://arxiv.org/abs/1407.7849. Accessed Aug. 4, 2014.