Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database

Kevin W. Boyack[1]

[1] *kboyack@mapofscience.com*
SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122 (USA)

Abstract

A dataset containing 111,616 documents in astronomy and astrophysics has been created and is being partitioned by several research groups using different algorithms. In this paper, rather than partitioning the dataset directly, we locate the data in a previously created model in which the full Scopus database was partitioned. Given that the other research groups are partitioning the data directly, use of this method will allow comparisons between using local and global data for community detection. In other words, use of this method will allow us to start to answer the question of how much the rest of a large database affects the partitioning of a journal-based set of documents. We find that the astronomy document set, while spread across hundreds of partitions in the Scopus map, is located in only a few regions of the map. Thus, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. However, we do not expect local and global data to give as similar results for other fields, because most other fields are less insular than astronomy.

Conference Topic

Methods and techniques

Introduction

Partitioning of a dataset into groups of similar objects – alternatively known as clustering, community detection or topic detection – is a current research topic in a number of fields, including scientometrics and network science. A number of different algorithms are used to partition scientific literature into topics or clusters. While many of these are based on the property of modularity (cf., Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Newman & Girvan, 2004; Waltman & van Eck, 2013), others are based on graph layout and pruning (Martin, Brown, Klavans, & Boyack, 2011) or on complex network flows (Rosvall & Bergstrom, 2008). Despite the obvious differences between these algorithms, they are all based on a common principle – that of dividing a literature set into partitions where the within-partition signals are much stronger or denser than the between-partition signals.

It is well known that different topic detection algorithms give somewhat different results for the same data set. What is not known is the specifics of why particular algorithms give particular results, or exactly what operations of a particular algorithm lead it to give different results than those obtained by another algorithm. In general, we know very little about what types of features result from different algorithms, and how these affect the output structures. This can make it difficult to interpret the partitions and maps that are produced by an algorithm. Are the partitions produced by an algorithm representative of actual structures in science, are they merely artifacts resulting from the algorithm and its parameters, or are they something in between? This is a difficult question to which, we suspect, the answer is far beyond the scope of even a large study. Nevertheless, we are hopeful that a comparison of partitioning methods and their results using a single dataset might lead to some general understanding of the types of features that result from different methods and algorithms. This type of understanding has the potential to enable both researchers and decision makers to more clearly understand and interpret the results of a particular partitioning.

To this end, a number of researchers (see papers from this special session) have come together to explore this question. As a first step, each research group has created a partitioning of a single dataset using their own algorithms. The work-in-progress papers in this session describe the partitioning algorithms and results from each group. The multiple results will be combined and compared in a next phase of the project to determine similarities and differences in the features resulting from the different methods and algorithms. Beyond that, we collectively hope to learn more about both common and unique structural features that result from the different algorithms.

This paper details the method used by SciTech Strategies to partition an "astronomy and astrophysics" literature dataset. It differs from the other methods in one significant aspect – the other groups have all created local solutions (partitioning the dataset directly), while we have created a global model (partitioning the entire Scopus database) and have located the astronomy dataset within those partitions (Klavans & Boyack, 2011). Use of this method enables us to start to answer the question of how much the rest of the database affects the partitioning process.

Global Model

Our global model of science consists of 48,533,301 documents from Scopus. Of these, 24,615,844 documents are indexed source documents from Scopus 1996-2012, while the remaining 23,917,457 are non-source documents that were each cited at least twice by the set of source documents. The method used to generate the document set and citing-cited pairs list is very similar to that used for the recent "non-source" map of Boyack and Klavans (2014).

The model was created by taking the over 582 million citing-cited pairs within this set of 48.5 million documents, calculating similarity values between pairs of documents based on direct citation, and then partitioning the documents using the new CWTS smart local moving algorithm (Waltman & van Eck, 2013). The citing-cited pairs were provided by SciTech Strategies (STS) to Ludo Waltman at CWTS, who ran the similarity calculation and partitioning steps. The CWTS smart local moving algorithm was used to create a four-level hierarchical solution, with resolution values chosen to result in a solution with roughly 100k, 10k, 1000, and 100 clusters. Details of the partitioning are given in Table 1.

Table 1. Multi-level partitioning of the Scopus database using the CWTS smart local moving algorithm.

Level	*Partitions Desired*	*Resolution*	*Partition Min Size*	*# Partitions*	*Partitions > Min Size*	*# Pubs*	*% Pubs Lost*
1	100000	3e-5	50	114679	91726	48399235	0.28%
2	10000	3e-6	500	13157	10059	47323189	2.49%
3	1000	3e-7	5000	1048	849	46929303	3.30%
4	100	5e-8	50000	122	114	46705047	3.77%

Visual maps of the partition solutions at level 1 and level 2 were created using the following process. At each level, 1) pairwise similarity between partitions was calculated from the titles and abstracts of the documents in each partition using the BM25 textual similarity measure, 2) each resulting similarity list was filtered to retain the top-n (5-15) similarities per partition, and 3) layout of the partitions on the x,y plane was done using the DrL algorithm. These steps are ones we commonly use to create science maps, and are described in more detail in Boyack & Klavans (2014). In each case, only those partitions that were of the minimum size desired (91,726 for level 1, and 10,059 for level 2) were included in the map. Field counts for each cluster in each map were calculated using UCSD map of science journal-to-field assignments (Börner et al., 2012), and each cluster was assigned to its dominant field and correspondingly colored in the map. The two maps are similar in that they show that the 12 large fields (see legend at the bottom of Figure 1) occupy similar positions in both maps. The change in granularity of the partitions does not change the overall look and feel of the map.

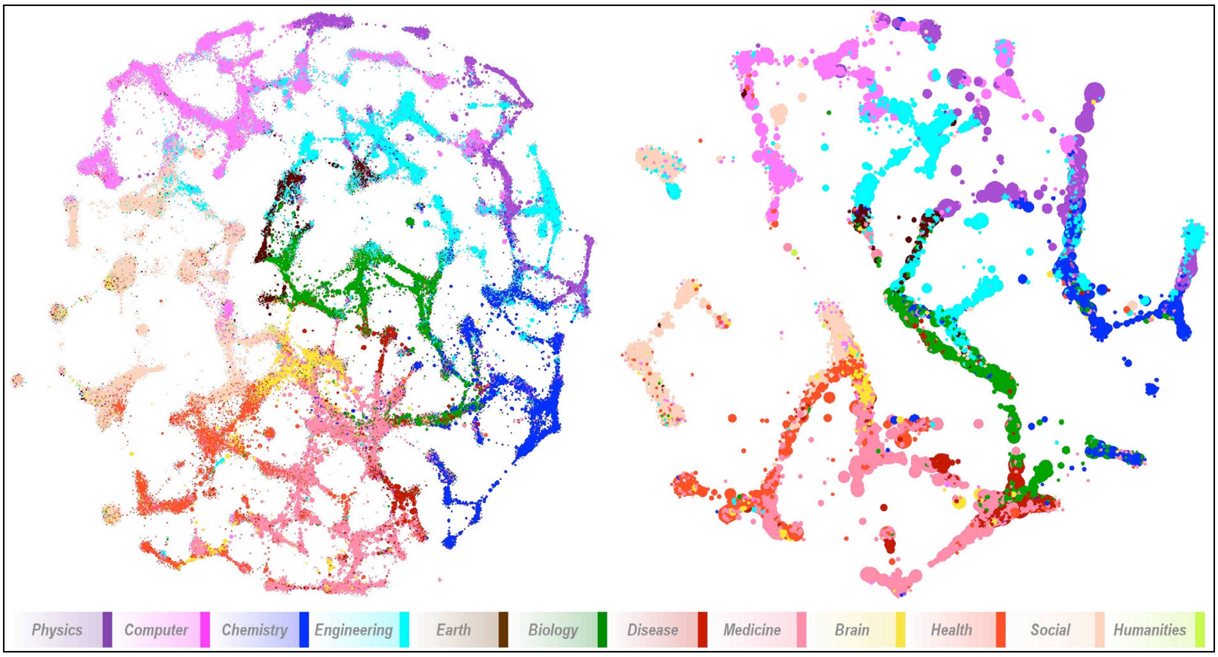

Figure 1. Visual maps of the Scopus database using level 1 (left) and level 2 (right) partitions.

Astronomy Dataset

The astronomy dataset used by each research group consists of 111,616 document records with accompanying data from the Web of Science. This dataset was created by researchers at Humboldt University for use by project participants, and is comprised of documents published from 2003-2010 in a set of 59 astronomy and astrophysical journals, limited to articles, letters, and proceedings papers. Over half of the documents were from four journals, as shown in Table 2.

Table 2. Dominant journals in the astronomy and astrophysics dataset.

Journal	Count
Astrophysical Journal	19582
Physical Review D	19061
Astronomy & Astrophysics	14668
Monthly Notices of the Royal Astronomical Society	11599

In order to use the Scopus-based global model and map, Scopus identifiers for the WoS records were identified to the extent possible by matching source data (journal, title, volume, page, year). Definitive matches were obtained for 107,888 (96.66%) of the documents. Of the 3,728 documents that were not matched, roughly half were in source titles that are not covered by Scopus (such as the IAU Symposium), and thus could only be matched if they were cited non-source materials. The remaining unmatched records were in source titles that are covered by Scopus, but that we could not match. This lack of uniformity between databases is primarily due to differences in the way titles are listed (particularly for non-ASCII characters) and missing records. Despite the unmatched records, we consider a match rate of nearly 96.7% to be very good, and certainly sufficient for reasonable comparison with the partitions from other groups. Once the matching was done, documents from the astronomy dataset were located in global map at three levels (1, 2, and 3 from Table 1).

Astronomy is known to be a relatively insular discipline, with fewer links (percentage basis) to and from other disciplines than for most other disciplines. Thus, we expected the effect of including an additional 48 million documents in a cluster solution to have only a modest effect on the partitioning of the astronomy document set. We did not expect the astronomy documents to be scattered throughout the map. As expected, the astronomy documents are heavily concentrated in the global model. At level 1, 50% of the astronomy documents are in partitions where the astronomy set documents comprise at least 66.5% of the partition contents (limited to the years of study, 2003-2010). In other words, when sorting partitions by concentration of the astronomy document set within the partition, 50% of the total papers are accounted for in partitions with a concentration of over 66.5%. Using an alternative measure, when partitions are sorted by the number of papers from the astronomy document set, the number of non-set papers equals the number of set papers only when 90,000 of the 111,616 papers are accounted for, as shown in Figure 2.

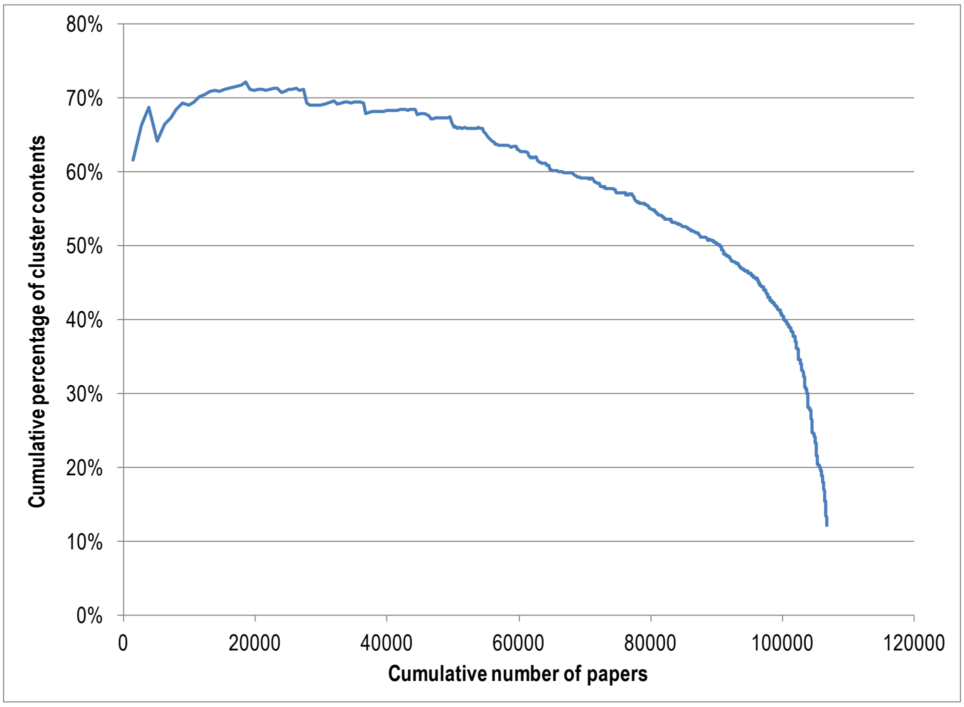

Figure 2. Distribution of the astronomy dataset across partitions in the level 1 solution.

Overlays showing the positions of the partitions with at least 50 documents from the astronomy set are shown for both the level 1 and level 2 maps in Figure 3. For level 1, this comprises 408 partitions and 90,763 documents (84.1% of the matched documents), while for level 2 it comprises 119 partitions and 101,895 documents (94.4% of the matched documents). Both maps make it clear that while the documents are parsed out into hundreds of partitions, each representing distinct topics, these topics are concentrated in only a few areas in the map.

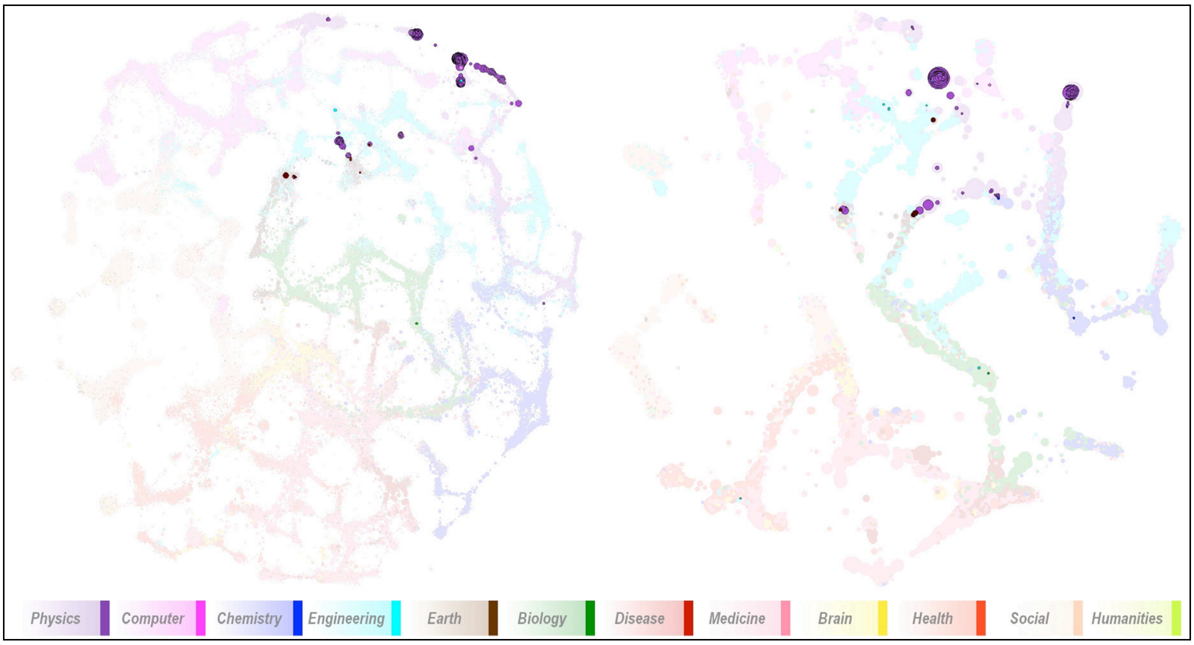

Figure 3. Overlays of the positions of the astronomy set documents on the Scopus level 1 (left) and level 2 (right) maps of Figure 1.

Discussion

Recalling that the astronomy document set was based on a set of journals, the high level of concentration of the overlays shown in Figure 3 suggests that use of journals is a very reasonable strategy for building a dataset in the field of astronomy. Astronomy journals have a very tight profile on a document-based map. By contrast, high profile journals in other fields, such as JACS, Physical Review Letters, and New England Journal of Medicine, have very broad profiles, and overlays for these journals (not shown here) spread across large regions of the map. Thus, while a dataset based on journals is useful to characterize astronomy, journals may be far less useful for characterizing other fields. Correspondingly, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. We would not expect the use of a global map to partition a local document set from another field to work as well. Or, rather, we would expect the journal-based approach to fall short in other fields because it would leave out so much of the relevant contextual literature.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, P10008.

Boyack, K.W., & Klavans, R. (2014). Including non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*, 569-580.

Börner, K., Klavans, R., Patek, M., Zoss, A.M., Biberstine, J.R., Light, R.P., Larivière, V., & Boyack, K.W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Klavans, R., & Boyack, K.W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

Martin, S., Brown, W.M., Klavans, R., & Boyack, K.W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE - The International Society for Optical Engineering, 7868*, 786806.

Newman, M.E.J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Rosvall, M. & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the USA, 105*(4), 1118-1123.

Waltman, L. & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.