This is a preprint version of a paper that has been accepted to appear in *Scientometrics*, 2/8/17.

Investigating the Effect of Global Data on Topic Detection

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 (Astro-set) has been created and is being partitioned by several research groups using different algorithms. For this paper, rather than partitioning the dataset directly, we locate the data in a previously created model of the full Scopus database. This allows comparisons between using local and global data for community detection, which is done in an accompanying paper. We can begin to answer the question of the extent to which the rest of a large database (a global solution) affects the partitioning of a smaller journal-based set of documents (a local solution). We find that the Astro-set, while spread across hundreds of partitions in the Scopus map, is concentrated in only a few regions of the map. From this perspective there seems to be some correspondence between local information and the global cluster solution. However, we also show that the within-Astro-set links are only one-third of the total links that are available to these papers in the full Scopus database. The non-Astro-set links are significant in two ways: 1) in areas where the Astro-set papers are concentrated, related papers from non-astronomy journals are included in clusters with the Astro-set papers, and 2) Astro-set papers that have a very low fraction of within-set links tend to end up in clusters that are not astronomy-based. Overall, this work highlights limitations of the use of journal-based document sets to identify the structure of scientific fields.

Introduction

Partitioning of a dataset into groups of similar objects – alternatively known as clustering, and often (perhaps insufficiently) framed as 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 (or drawing) 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. Although this knowledge is primarily informal, recent comparisons provide evidence that this is the case (Emmons, Kobourov, Gallant, & Börner, 2016; Subelj, Van Eck, & Waltman, 2016). 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 have come together to explore this question. As a first step, each research group has created a partitioning of a dataset of astronomy and astrophysics papers (hereafter called 'Astro-set', and described below) using their own algorithms. Accompanying papers describe the partitioning algorithms and results from each group. Clustering results have been shared within the larger group, and a number of comparisons between the result sets have been created, many of which appear in the accompanying paper by Velden et al. (2017), and some of which appear in the individual papers. 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 (STS) to partition the Astro-set, along with some results. The STS method differs from the other methods in one significant aspect. We have created what we call a 'global' model, meaning that we have partitioned global data (from all of science) rather than partitioning data from a single field or specialty. The other groups have all created what we call a 'local' model because they partition data from a single field or specialty – in this case, the field of astronomy and astrophysics. In this study, we locate the 'local' Astro-set papers within the 'global' model (Klavans & Boyack, 2011), and analyse the partitions where the Astro-set papers are located. 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

The STS 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 documents that are not indexed by Scopus, but that appear in the reference lists of the source documents. We include all cited non-source documents that were cited at least twice by the set of source documents. Non-source documents are not restricted to any particular time window, but are included because 1) they add important content to the model, and 2) they increase the accuracy of the model by increasing the signal that is used in the clustering process. 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 (2014b).

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 (references to other documents and citations received from other documents), and then partitioning the documents into clusters. The citing-cited pairs were provided by SciTech Strategies to Ludo Waltman, who used the CWTS similarity and smart local moving algorithms (Waltman & van Eck, 2012, 2013) to create a four-level hierarchical solution. This algorithm allows desired minimum partition size and resolution parameters to be specified for each level in the hierarchy – the values in Table 1 were chosen to give solutions that resulted in roughly 100k, 10k, 1000, and 100 clusters at the four levels of granularity. Each level in the hierarchical solution is the starting point for the next level. For example, level 2 partitions consist of groups of level 1 partitions. Levels 2-4 contain fewer papers than the level 1 solution because some of the clusters at each level do not link to other clusters, and are thus dropped from the solution as one moves up the hierarchy (e.g., from level 1 to level 2). Details of the partitioning are given in Table 1.

A visual map of the partition solution at level 1 was created using the following process: 1) pairwise similarity between partitions was calculated from the titles and abstracts of the documents in each partition using the BM25 textual similarity measure (Sparck Jones, Walker,

& Robertson, 2000), 2) the 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 and Klavans (2014b). The 91,726 partitions that met the desired minimum size criterion were included in the map. Each partition was assigned a field and color using the journal-to-color scheme from the UCSD map of science (Börner et al., 2012) as shown in Figure 1. The UCSD map was used for coloring only, and did not affect the partitioning in any way.

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

Level	*Partitions Desired*	*Resolution*	*Desired Minimum Size*	*# Partitions*	*# Partitions > Minimum Size*	*# Pubs*	*% Coverage*
1	100000	3e-5	50	114679	91726	48399235	99.72%
2	10000	3e-6	500	13157	10059	47323189	97.51%
3	1000	3e-7	5000	1048	849	46929303	96.70%
4	100	5e-8	50000	122	114	46705047	96.23%

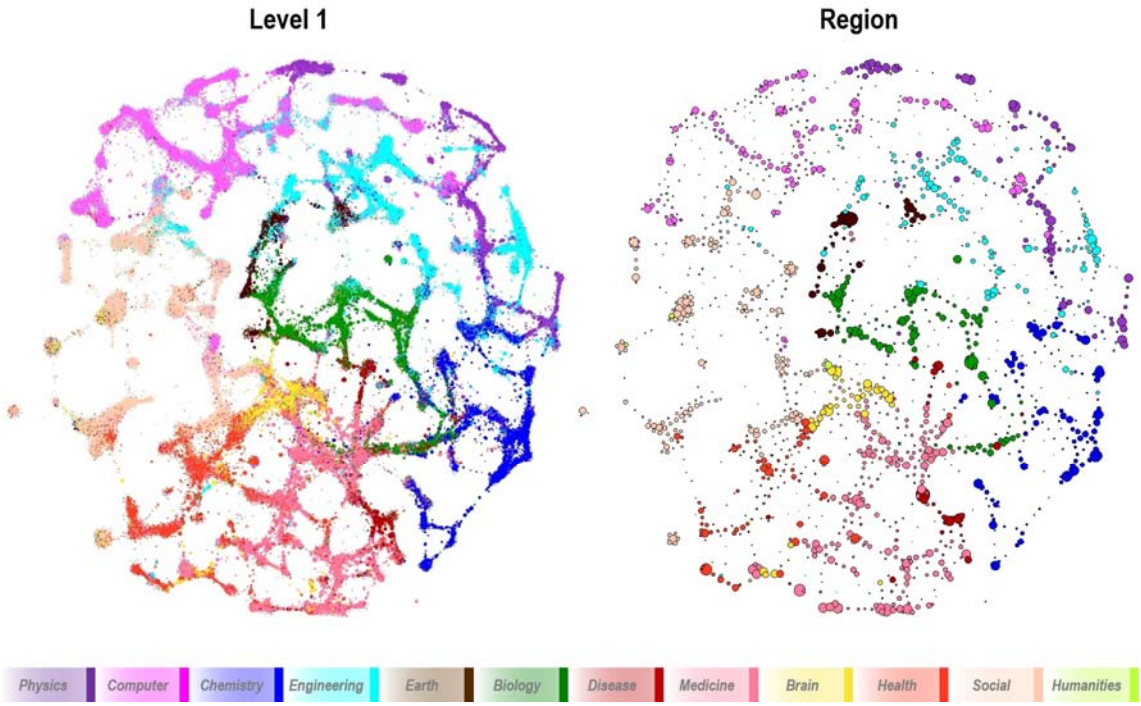

Figure 1. Visual maps of the Scopus database using level 1 partitions, and regions formed from groups of adjacent level 1 partitions.

At STS, we typically work with highly detailed models and maps comprised of ~10^5 partitions, such as those in the level 1 map of Figure 1, because this level of detail is consistent with expert knowledge (Boyack, Klavans, Small, & Ungar, 2014) and is useful for the identification of emerging topics (Small, Boyack, & Klavans, 2014). However, all of the local cluster solutions created by other research groups have been done at a much more aggregated level. Thus, for comparison with other solutions we need to use a more aggregated solution.

In previous work we have found that a combination of clustering using citation followed by a second-level clustering using text is more accurate (i.e., has stronger similarity signals and shorter author distances) and more visually appealing than multi-level clustering based solely on citation information (Boyack & Klavans, 2014a). Thus, rather than using our level 2 or level 3 direct citation solutions for comparisons, we have grouped level 1 partitions into 1,649 regions as shown in Figure 1. This was done algorithmically, using large level 1 partitions as region seeds, and iteratively adding adjacent partitions to regions using cluster-to-cluster textual similarity (moving sequentially from larger to smaller partitions) from the level 1 map. Originally, we started with ~1,000 seeds, but found that some areas of the map did not get added to a region because they were local high density areas lying between lower density areas. Seeds were then added and more regions created until all of the level 1 partitions were assigned to a region.

Astro-set

The Astro-set used by each research group consists of 111,616 document records with accompanying data from the Web of Science. This dataset is comprised of documents published from 2003-2010 in a set of 59 astronomy and astrophysical journals from the Astronomy & Astrophysics subject category, 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 (APJ)	19582
Physical Review D (PRD)	19061
Astronomy & Astrophysics (A&A)	14668
Monthly Notices of the Royal Astronomical Society (MNRAS)	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. Definitive matches (using electronic matching of journal, title, volume, page, year, and first author) 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 which are used often in titles of physics and astronomy papers) 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 Astro-set were located in the level 1 partitions and regions of the STS global map.

Astronomy in the Context of Global Science

Partition size distributions for the full STS global model and Astro-set papers in global model partitions are shown in Figure 2. For both level 1 partitions and regions, the STS distributions are relatively flat for the largest partitions. Distributions for the Astro-set partitions are much more skewed, with a few partitions with large numbers of Astro-set papers, along with a substantial tail of partitions with only a few Astro-set papers. Overlays showing the positions of the Astro-set partitions with at least 50 documents are shown for both the level 1 and region maps in Figure 3. For level 1, this comprises 408 partitions and 90,763 documents (84.1% of the matched documents), while for regions it comprises 51 partitions and 104,711 documents

(97.1% of the matched documents). More of the matched documents are shown in the region map because the aggregated nature of the region map puts more documents in partitions that are above the viewing threshold of 50 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, most of which are in physics, with a few in earth sciences.

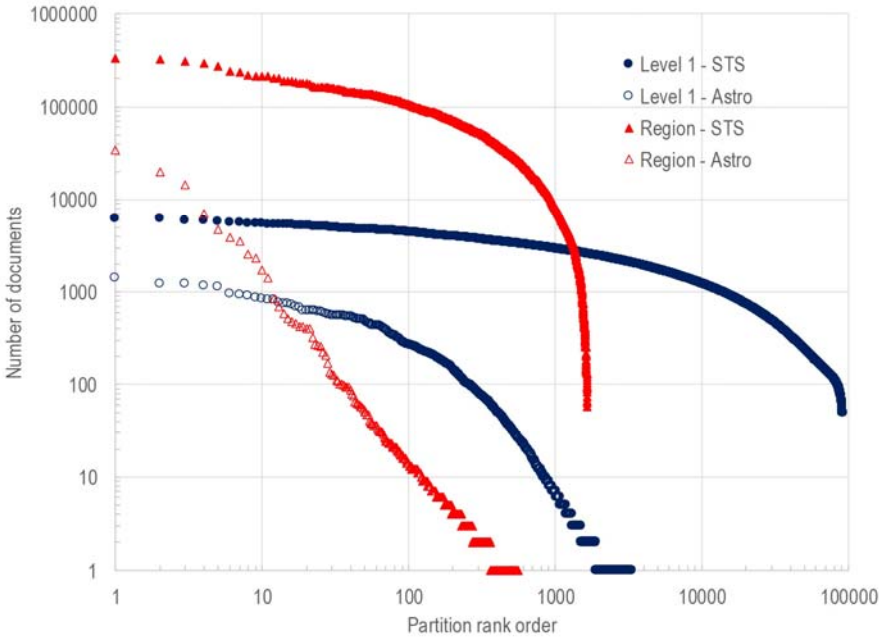

Figure 2. Partition size distributions for the full model and Astro-set. Astro-set partition sizes reflect the numbers of Astro-set papers in the global partitions rather than the full sizes of those partitions. Note the log-log scale.

The top 10 regions comprise 94,384 (87.5%) of the documents in the Astro-set. Of these, the largest three regions comprise over 63%. From this perspective there seems to be some correspondence between local information and the global cluster solution. Table 3 provides a brief description of each of these regions, including the number of documents from the Astro-set and the concentration of Astro-set documents in the region. Astro-set documents comprise roughly a third or more of four regions, while they comprise less than 10% of the documents in two of the regions. It is also notable that when Science-Metrix journal categories (Archambault, Caruso, & Beauchesne, 2011) are used as descriptors, four of the five largest regions have *Astronomy & Astrophysics* as their core. However, the other six regions are more related to *Nuclear & Particle Physics*, and to *Meteorology & Atmospheric Sciences* than to astronomy. These additional categories have obvious relationships to astronomy, but together with the term descriptors provide a nice differentiation between regions in which the Astro-set documents are situated. The Science-Metrix journal classification system is a reasonable choice for labeling in that it was recently shown to be far more representative of the organization of current science than other journal classification systems (Klavans & Boyack, 2016), including those provided by WoS and Scopus.

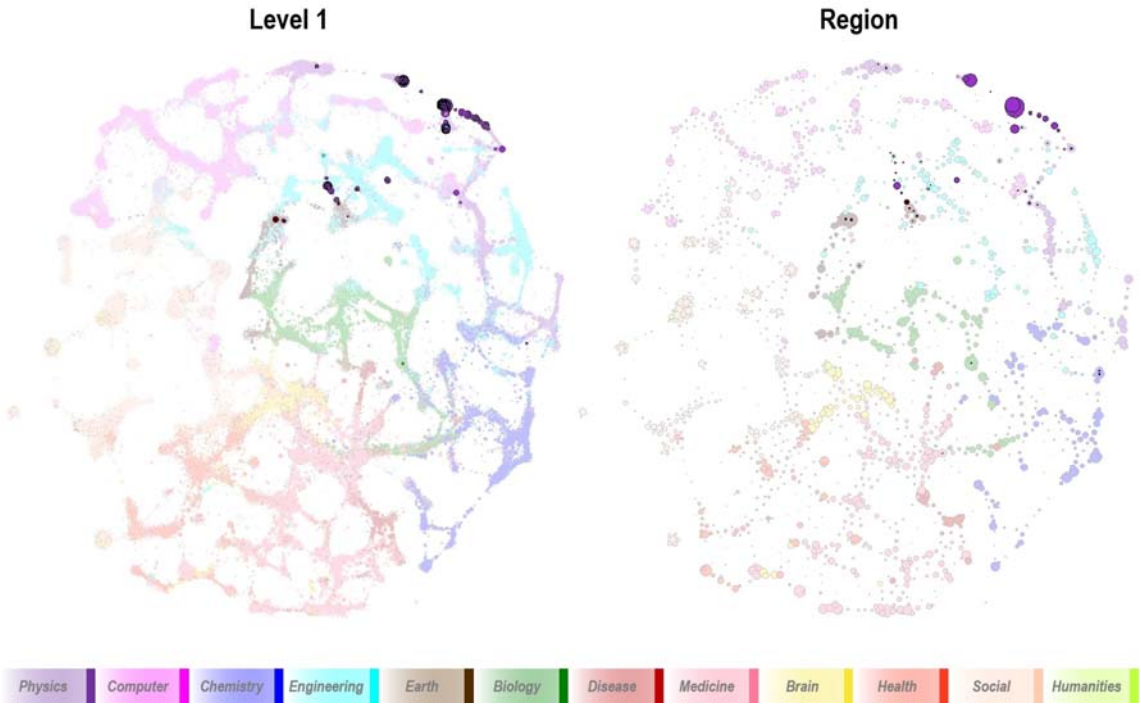

Figure 3. Overlays of the positions of the astronomy set documents on the level 1 and region maps of Figure 1. The Astro-set overlay is dark, the underlying basemap is light. Thus, contrast shows the location of the Astro-set documents.

It is interesting that four of the top 10 regions correlate most highly with the Science-Metrix journal category of *Nuclear & Particle Physics*. This should, however, not be too surprising given the list of journals comprising the astronomy dataset. For example, while three of the top four journals (APJ, A&A, and MNRAS, see Table 2) are clearly astronomy journals, the second ranked journal, Physical Review D (see http://journals.aps.org/prd/about), comprising over 17% of the Astro-set, spans particle physics and gravitational physics, thus bridging astronomy and particle physics. This simple example suggests that the Astro-set inherently contains overlaps with other fields, despite the fact that astronomy is widely considered to be a relatively insular and self-contained discipline (Schubert, 2013; Wallace, Lariviere, & Gingras, 2012). Thus, when global information is included in the partitioning of these data, it should naturally be expected that these additional data will lead to a different partitioning of the papers in the Astro-set than if the global information were excluded.

Table 3 also contains information about the distribution of source documents (Astro-set and non-Astro-set) and non-source documents, along with the percentage of within-set links in each region. There are more Astro-set documents than non-Astro-set documents in three of the top four regions (17, 126, 400), but non-Astro-set documents dominate in the remainder of the regions. The non-source contribution is significant at 22.5% of the documents in the top 10 regions, and suggests that the work that is considered significant by the scientific community (as represented by being cited at least twice) is much broader than what is indexed by the large citation database providers.

Note that while roughly 2/3 of the documents in the top 10 regions are from outside the Astro-set, the level 1 partitions are much more concentrated. Level 1 topics are much more focused on astronomy than is shown by our analysis at the region level. Aggregation to the region level

dilutes the signal from the Astro-set. For example, region 16 (rank 21 in Table 3) is large with 56,765 source documents, but its Astro-set contents are dominated by a single level 1 cluster with 284 Astro-set documents out of its 1087 source documents. This is a case where a level 1 cluster is focused on astronomy, but the larger context of the region in which it resides (because of the text in its titles and abstracts) is that of Geochemistry (see Table 3). Thus, while we are providing a description of regions in this paper to compare with the other solutions, we prefer to work with level 1 clusters for most bibliometrics applications because that level of granularity produces well-defined topics that are easily recognized by experts (Boyack et al., 2014) and that produce validated lists of emerging topics (Small et al., 2014).

These data also suggest that the common strategy of creating journal-based document sets to represent fields is tenuous at best. As interdisciplinary research increases, related documents can be expected to be published in ever increasing numbers of sources. For example, Table 4 lists the leading additional sources with papers (over the same time-frame, 2003-2010) in the top 400 level 1 partitions. These nine sources each have more than 2000 documents that are co-located in level 1 partitions with documents from the astronomy dataset, showing that papers on astronomy-related topics are present in many sources other than those listed in the WoS Astronomy & Astrophysics subject category. In fact, seven of the nine sources in Table 4 are available in the Thomson Reuters databases, but are not assigned to the A&A subject category.

This is a preprint version of a paper that has been accepted to appear in *Scientometrics*, 2/8/17.

Table 3. Characterization of the top 35 regions containing Astro-set documents.

Rank	*Region*	*#Astro-set Docs 2003-10*	*# non-Astro-set Docs 2003-10*	*# Non-source Docs 2003-10*	*% Astro-set Docs (Astro/All)*	*% Links within Astro-set*	*Descriptive terms for Astro-set documents (Velden et al., 2017)*	*Science-Metrix category*
1	17	33874	18546	16487	49.16%	37.48%	star, main sequence, binary, light curve, gamma ray, white dwarf, neutron star, emission, low mass, x ray	Astronomy & Astrophysics
2	126	19988	7058	6461	59.65%	42.89%	galaxies, redshift, star formation, clusters, sample, active galactic, agn, sloan digital, digital sky, halo	Astronomy & Astrophysics
3	48	14588	23936	12132	28.80%	32.19%	scalar field, spacetime, metric, inflation, cosmological constant, dark energy, gravity, general relativity, universe, einstein	Nuclear & Particle Physics
4	400	7076	5080	3577	44.98%	32.44%	solar, coronal, active region, cme, flare, sunspot, mass ejections, magnetic flux, quiet sun, transition region	Astronomy & Astrophysics
5	403	4720	5755	3809	33.04%	24.75%	asteroid, saturn, comet, titan, cassini, jupiter, icarus, albedo, main belt, kuiper belt	Astronomy & Astrophysics
6	191	3941	20220	5805	13.15%	18.64%	decays, meson, qcd, pi pi, j psi, leading order, bar, quark, inclusive, factorization	Nuclear & Particle Physics
7	106	3557	26879	8837	9.06%	16.17%	yang mills, gauge theory, mills theory, noncommutative, lattice, supergravity, string, qcd, finite temperature, branes	Nuclear & Particle Physics
8	425	2593	6793	3088	20.79%	19.79%	standard model, higgs, lhc, minimal supersymmetric, supersymmetric standard, lepton, seesaw, neutrino masses, right handed, leptogenesis	Nuclear & Particle Physics
9	355	2314	11454	3047	13.76%	14.07%	auroral, substorm, solar wind, magnetopause, magnetosphere, plasma sheet, field aligned, cluster spacecraft, ionospheric, ion	Meteorology & Atmos Sciences
10	310	1733	12941	4500	9.04%	9.98%	ionospheric, gps, tec, iri, electron content, mesosphere, degrees n, total electron, ionosonde, summer	Meteorology & Atmos Sciences
11	578	1424	3952	1721	20.07%	31.30%	pamela, matter annihilation, wimp, neutrino, weakly interacting, interacting massive, dark matter, positron, super kamiokande, theta 13	Nuclear & Particle Physics

12	972	844	2513	1032	19.23%	27.01%	ultra high, air shower, cosmic rays, extensive air, uhecr, pierre auger, 19 ev, auger observatory, energy neutrinos, neutrino flux	Nuclear & Particle Physics
13	381	696	8243	3304	5.69%	34.47%	hanle, focal, mirrors, adaptive optics, wavefront, optics, integral field, laser guide, guide star, ifu	Optics
14	**53**	580	42302	17510	0.96%	11.72%	presolar, meteorites, isotopic compositions, chondrules, minerals, inclusions, isotopic, lunar, olivine, solar nebula	Geochemistry & Geophysics
15	644	516	6438	1220	6.31%	14.47%	pentaquark, nucleon, baryon, octet, decuplet, pion, n c, form factors, chiral, strangeness	Nuclear & Particle Physics
16	1381	477	639	310	33.45%	23.09%	meteor, leonid, geminid, radiant, nutation, ablation, video, trails, shower, perseid	Astronomy & Astrophysics
17	238	457	18260	3419	2.07%	28.18%	neutron capture, poor stars, nucleosynthesis, extremely metal, metal poor, capture elements, process elements, cemp, third dredge, isotopes	Nuclear & Particle Physics
18	440	426	7880	4465	3.34%	13.81%	sail, thrust, debris, propulsion, restricted three, earth orbit, body problem, trajectory, maneuvers, geo	Aerospace & Aeronautics
19	502	421	10908	1422	3.30%	21.45%	r matrix, oscillator strengths, breit pauli, dielectronic recombination, transition probabilities, collision strengths, electron impact, 3p, impact excitation, rate coefficients	General Physics
20	**104**	407	28353	7360	1.13%	19.42%	body problem, restricted three, periodic orbits, three body, equilibrium points, photogravitational, lyapunov, collinear, families, planar	General Mathematics
21	**16**	403	56362	22141	0.51%	10.98%	mars, deposits, hesperian, crater, amazonian, volcanic, hirise, geological, gullies, fluvial	Geochemistry & Geophysics
22	503	320	8780	1648	2.98%	16.24%	dynamos, dynamo action, magnetic reynolds, electromotive, reynolds number, prandtl number, geodynamo, magnetic helicity, nonhelical, magnetic prandtl	Fluids & Plasmas
23	474	269	9416	1276	2.45%	35.84%	interferometric gravitational, wave detectors, mirrors, geo 600, gravitational wave, suspension, lcgt, interferometer, fused, thermoelastic	Optics
24	266	262	17638	5837	1.10%	5.15%	ozone, aerosol, aod, toms, stratospheric, envisat, retrieval, gome, sciamachy, modis	Meteorology & Atmos Sciences

This is a preprint version of a paper that has been accepted to appear in *Scientometrics*, 2/8/17.

25	147	260	33134	4535	0.69%	14.88%	rotational transitions, anion, ab initio, c6h, irc 10216, dissociative recombination, vibrational, tmc, molecule, franck	Chemical Physics
26	**36**	224	44491	15469	0.37%	4.02%	rainfall, tropical, precipitation, meteorological, mesoscale, mm5, ocean, sea level, grace, weather	Meteorology & Atmos Sciences
27	**1049**	209	1413	872	8.38%	13.98%	bacillus, subtilis, bacterial, spores, microorganisms, drilling, tinto, biological, rio, mars	Astronomy & Astrophysics
28	330	169	16806	2622	0.86%	15.51%	ethyl, rydberg, lih, formate, vibrational, ch3ch2cn, molecule, cyanide, predissociation, ab initio	Chemical Physics
29	**5**	136	75692	12646	0.15%	9.86%	plants, seedlings, arabidopsis, life support, wheat, grown, gravitropism, bioregenerative, shoots, germination	Plant Biology & Botany
30	628	131	4182	2945	1.81%	12.35%	tourism, economic, human, stiefel, industry, quaternions, kustaanheimo, social, countermeasures, psychological	Aerospace & Aeronautics
31	792	126	4750	797	2.22%	12.67%	dose, hzetrn, dosimetry, liulin, aircrew, shielding, tissue, space station, international space, station iss	Applied Physics
32	1084	112	1409	1044	4.37%	9.75%	elite, terraforming, lander, expeditions, lidov, philae, rendezvous, unsupported, simulant, society	Aerospace & Aeronautics
33	138	109	27563	4190	0.34%	11.19%	nonextensive, tsallis, microcanonical, caloric, mechanics, inequivalence, additivity, bose, coarse grained, self gravitating	Fluids & Plasmas
34	365	102	11736	2457	0.71%	17.34%	sipm, astrosat, irst, fbk, ray astronomy, readout, counters, nct, mega, calorimeter	Nuclear & Particle Physics
35	384	100	11213	3192	0.69%	9.91%	riemannian, osserman, anholonomic, lorentzian, finsler, causal, manifold, nilpotent, chronological, achronal	General Mathematics

Table 4. Additional sources that are highly represented in level 1 partitions.

Source	*Count*
AIP Conference Proceedings [B]	7924
Journal of High Energy Physics [A]	5972
Physics Letters, Section B [A]	4689
Proc. International Astronomical Union	3440
Nuclear Physics B – Proc Supplements [A]	3367
Journal of Geophysical Research A [A]	2966
Proceedings of SPIE [B]	2905
European Space Agency, Special Pub	2754
Physical Review Letters [A]	2672

[A] Available in Web of Science; [B] Available in TR Proceedings Citation Index

While the majority of the 10 largest regions are obviously related to astronomy and astrophysics either directly, or in the related categories of *Nuclear & Particle Physics* or *Atmospheric Sciences*, it is also instructive to explore some of the regions containing smaller, but still significant, numbers of the astronomy set papers. Table 3 contains a number of such examples – regions 53, 104, 16, 36, 1049, and 5 will be analysed further below. If one were to only consider the information in Table 3, it might appear that the Astro-set papers populating these regions have little to do with astronomy. A more detailed analysis suggests that, while these paper sets do have a relationship to astronomy, the primary affiliation of these paper sets is typically to a topic other than astronomy, as will be shown by examples below.

Word clouds created from article titles are shown in Figure 4 for each of the six regions mentioned above. Two word clouds are shown for each region – one based on titles from Astro-set papers, and one based on titles from a random sample of 2000 papers from the full STS region. Comparison of the Astro-set word clouds with STS word clouds for the same regions is very instructive.

For example, the Astro-set word cloud for region 104 suggests that those papers are about orbits and stability (a class of three-body problems). However, a word cloud based on the full region contents shows that region 104 is about mathematical techniques. The Astro-set papers thus appear to be an application space within the larger discipline of mathematical functions, and while these papers are found in astronomy journals, they are more closely linked through their citations to mathematics than astronomy. Most of the other examples in Figure 4 are similar in that the Astro-set word cloud shows the astronomical nature of the Astro-set papers, while the word cloud based on the full set of region papers shows the larger context in which the Astro-set subset occurs. For region 36, the Astro-set papers are about modelling using satellite data, while the region context is climate science. For region 5, the Astro-set papers are about plant growth experiments performed in space environments, while the region context is plant (and particularly, Arabadopsis) genomics research. These thus appear to be cases where a small astronomy-related application is part of another discipline. In each case the non-Astro-set papers have dominated the citation-based signal and have, in essence, pulled these papers into their partition, and away from an astronomy-based partition. In contrast, in the local cluster solutions created by other project groups, these papers are found grouped with other topics in clusters that are astronomy-based because the papers providing the larger context for that work are not part of the Astro-set. In cases such as these, the addition of global information clearly makes a difference in the calculated topic structure.

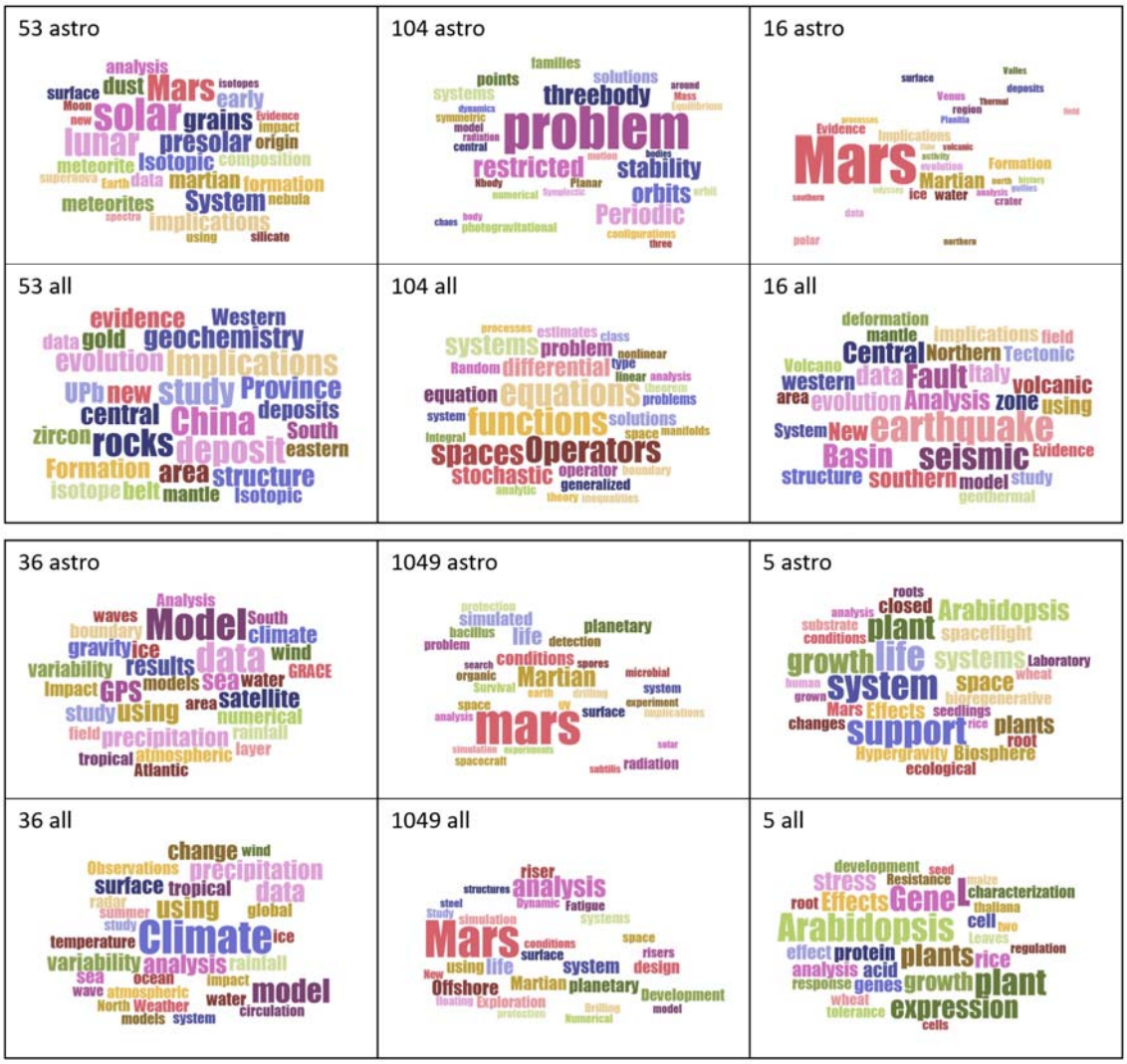

Figure 4. Word clouds created from the titles of papers in six regions with relatively small numbers of Astro-set papers. Word clouds based on Astro-set papers are compared with those representing full regions. (Word cloud generator: www.jasondavies.com/wordcloud)

One exception to this is region 1049. Figure 4 shows that the Astro-set and full word clouds contain similar terms – each word cloud is strongly dominated by the word "Mars". Table 3 shows that the Astro-set papers comprise 8.4% of the region 1049 papers, while for the other five cases in Figure 4, the Astro-set papers comprise 1.1% or less of their region. Interestingly, the fraction of links from the Astro-set papers to other Astro-set papers in region 1049 (14%) is not much different that for the other five regions from Figure 4 (average 11.2%). This suggests that the non-Astro-papers in region 1049 are more similar in content to their Astro-set counterparts than is the case for the other regions. It is also possible that the Astro-set papers for the other five regions are from very small level 1 partitions in the full STS model, and have been diluted by being grouped into regions with papers on other topics.

For the entire set of Scopus matched documents upon which this analysis is based, only 33.1% of the citation links are within the astronomy paper set. The remaining 66.9% of the links are to other papers that are outside the set. When thinking about how papers in the astronomy set might be partitioned when these global links are included, it is natural to assume that those

papers that are highly linked to other papers in the astronomy set will end up clustered together, while those with a large fraction of their links outside the astronomy set will end up in regions dominated by other disciplines. This assumption was checked by calculating the fraction of within-set links as a function of total links for each region. Figure 5 shows that this assumption is confirmed; regions with small fractions of Astro-set papers have a much lower fraction of their links within the astronomy set (~7% on average) than regions with very large numbers of astronomy set papers (>30% within-set links). To some degree this result is trivial, and could be predicted as an application of Bradford's law. What is often forgotten, however, is that all journal-based or keyword-based local datasets are subject to the same principal – namely, dispersion in a broader context – and all local datasets will exhibit this type of dispersion to some degree. All local datasets that attempt to describe a field or specialty will thus contain a boundary set that would contextually fit better somewhere else within the totality of science.

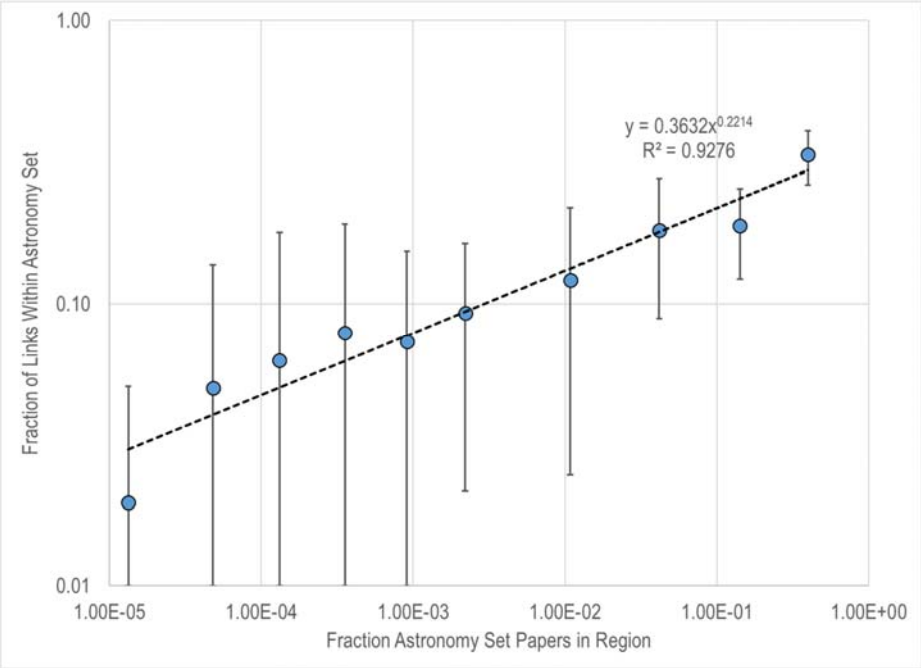

Figure 5. Fraction of within-set links as a function of the fraction of astronomy set papers by region. Error bars represent one standard deviation.

The above analysis of topics that are not primarily related to astronomy only applies to a modest fraction of the papers in the Astro-set. For example, only 20% of the papers are in regions where the fraction of links from those papers to other Astro-set papers is less than 20%, and only 8% of the papers are in regions where the fraction of links to other Astro-set papers is less than 15% (see Table 3). Despite the fact that this is a minority of the Astro-set, it is important to correctly classify these papers (whether 8% or 20%) whose primary contextual environment is something other than astronomy, and the context to do so is lacking in the Astro-set.

Conclusion

This work has located a journal-based set of astronomy and astrophysics papers in a global model of science created from the full Scopus database. As such, it provides a preliminary view into the differences between topics detected from global and local data. Astronomy is typically considered to be a relatively self-contained field. As such, it would be logical to assume that local and global data would give rise to similar topic sets. To some degree we find this to be

true – the three largest regions containing astronomy documents in our global model comprise 63% of the local dataset. However, these three regions also contain a like number of documents from outside the astronomy dataset; these documents may be from astronomy sources that are not in the WoS A&A subject category, or they may be from other linked areas such as physics, space sciences, and geosciences.

In addition, we show that the number of links to papers outside the astronomy dataset is roughly double that of the number of links within the set. Although some of the external signal is to other astronomy documents published in years not covered by the dataset, the external signal is significant and has a strong influence on the partitioning. First, in areas with a high concentration of Astro-set papers, related papers from non-astronomy journals are included in clusters with the astronomy papers. In these cases, the additional papers tend to have similar content to the Astro-set papers. Second, papers from astronomy journals that have a very low fraction of within-set links tend to end up in clusters that are not astronomy-based when external links are considered. In contrast, when external links are not included these same documents will be aggregated (or perhaps mis-aggregated) with dissimilar documents. For example, many Astro-set papers with a high fraction of external links are included in very large clusters in the CWTS partitioning of these data (van Eck & Waltman, 2017). This is a valid direct comparison since the same similarity measure and clustering algorithm were used in both cases. Additional testing is required to determine the extent to which local topics are 'polluted' by global information, or, conversely, the extent to which papers in global topics are misclassified due to weak connections in a local environment. The comparisons in Figure 4 suggest that the differences are large in some cases. Whether these differences are significant or not depends on the purpose of the analysis – some questions do not require the context of global data, while others do.

Using global data clearly has an effect on topic detection, even in a discipline that is considered relatively insular. Use of global data increases the context in two ways: 1) it adds related documents from the same discipline (astronomy, in this case) that were excluded from the initial set formation, and 2) it adds related documents from other disciplines in science. Given the increasingly interdisciplinary way in which science is conducted, we suggest that topic detection from journal-based data sets simply cannot provide the context needed to identify those topics that are small, that are secondary to a discipline, or that are truly interdisciplinary. Global modeling is needed to generate accurate structures in these cases.

Acknowledgments

This paper benefitted greatly from reviews by Theresa Velden, Andrea Scharnhorst, and two anonymous referees.

References

Archambault, E., Caruso, J., & Beauchesne, O. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics*, 66-77.

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.

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

Boyack, K. W., & Klavans, R. (2014a). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology, 65*(4), 670-685.

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

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management, 32*, 147-159.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS ONE, 11*(7), e0159161.

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.

Klavans, R., & Boyack, K. W. (2016). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, forthcoming*.

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.

Schubert, A. (2013). Measuring the similarity between the reference and citation distributions of journals. *Scientometrics, 96*, 305-313.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*, 1450-1467.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. Part 1. *Information Processing & Management, 36*(6), 779-808.

Subelj, L., Van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLoS One, 11*(4), e154404.

Van Eck, N.J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of *Scientometrics*, volume (issue), pages.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of *Scientometrics*, volume (issue), pages.

Wallace, M. L., Lariviere, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PLoS One, 7*(3), e33339.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

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.