Part III Project
Dissertation Draft 1

Patrick Lewis
Queens' College

March 2016

Abstract

Abstract goes here

Contents

List of Figures **2**

1 Conclusions **3**

2 Recommendations for Further Work **4**

Bibliography **5**

3 Appendix **7**
3.1 UK Departments scraped 8
3.2 11

List of Figures

3.1 Number of research communities authors are associated with. High values suggest more interdisciplinary work. The bars have been scaled relative to the author's total publication count. 11

1. Analysis with Sample Dataset

Having developed a framework to examine the models, attention was turned to some example analyses that could be carried out. Many possible applications for the framework suggest themselves, but the finite nature of the project necessitated focussing on analyses that could be achieved within the time-frame. [1]. With this in mind, it was decided to focus analysis on the a smaller subset of the training dataset, namely documents from the University of Cambridge Chemistry Department. This dataset is henceforth referred to as *CCD*[2].

1.1 Cambridge Chemistry research clusters

The CCD contained 9467 documents. The cosine matrix was calculated and a network was constructed from the matrix. *communities* within the network (clusters of strongly connected nodes) were identified by applying a high performance modularity algorithm**??**. The result is shown in figure **??**.

[1] Please refer to section 2 for more treatments
[2] Cambridge Chemistry Dataset

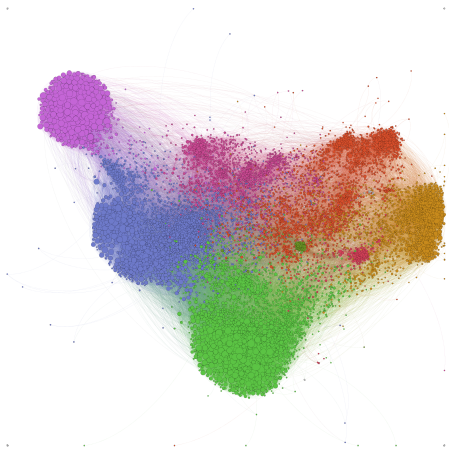

Figure 1.1: A Network visualisation of the CCD. Edges were placed between nodes with weight equalling cosine similarity if S_{Cosine}. Nodes are coloured into their detected communities, and node size is proportional to the number of connections a node has. nodes are arranged by modelling edges as springs.

It is apparent that the CCD contains contains clear communities of documents. This corresponds to different fields of research within the department. Some communities detected were small, but some were large (green, orange, etc...). The algorithm was then applied only to 'green' community to reveal subcommunities within the 'green' documents. A program was then written to recursively detect subcommunities in the CCD. This resulted in the CCD being divided into 300 communities of comparable size. The smallest communities were singleton documents, the largest community was 434 documents, and the mean population was 34.5. The community finding subdivision process is shown in figure **??**

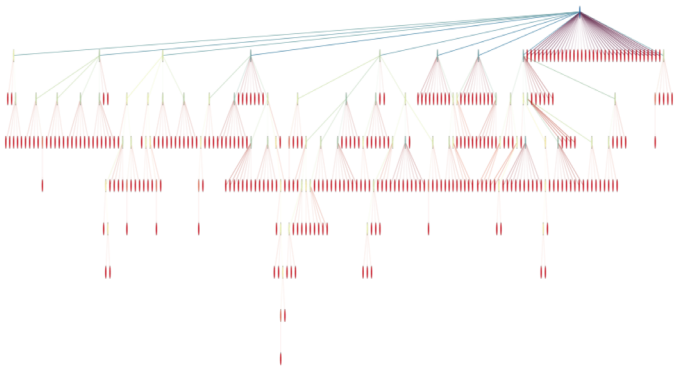

Figure 1.2: Recursion Tree depicting how communities were found. The dataset was partitioned using the modularity algorithm. Partitions with more than 100 documents were then repartitioned recursively. Partitions of less than 100 documents were considered to be communities (shown red in the diagram). The figure shows the maximum depth of partition required was 8, and most communities were found after 3 partitions.

Figure **??** can be interpreted as the *relationship* between different fields of research within the department. It is a shallow tree with highly branched nodes, suggesting that wide research fields, and qualitative overlap between fields. The process described is equivalent to an unsupervised categorisation algorithm. The entire process, from model training to finding communities has been without human labelling or intuition. It was therefore instructive to examine what the algorithm defined as communities. Communities were then analysed vigorously. Community clustering made intuitive sense in the majority of cases. Community 275 is typical:

Table 1.1: Community 275

Community Size	15
Depth down Recursion Tree	2
Contents	Bees, Neonicotinoids, toxicology, pollen.
Article closest to Mean Vector	10.1021/es2035152: Assessment of the Environmental Exposure of Honeybees to Particulate Matter Containing Neonicotinoid Insecticides Coming from Corn Coated Seeds
Community members	(Some omitted for brevity)
10.1007/s00216-012-6338-3	UHPLC-DAD, method for the determination of neonicotinoid insecticides in single bees and, its relevance in honeybee colony loss investigations
10.1021/es2035152	Assessment,of the Environmental Exposure of Honeybees to Particulate Matter Containing,Neonicotinoid Insecticides Coming from Corn Coated Seeds
10.1007/s11356-014-3470-y	Systemic insecticides (neonicotinoids and fipronil): trends,uses,mode of action and metabolites
10.1111/j.1439-0418.2012.01718.x	Aerial powdering of bees inside mobile cages and the extent of neonicotinoid cloud,surrounding corn drillers
10.1098/rsif.2013.0394	Analysing photonic structures in plants
10.1007/s00114-013-1020-y	The influence of pigmentation patterning on bumblebee foraging from flowers of Antirrhinum majus
10.1111/ics.12035	Keratins,and lipids in ethnic hair
10.1021/ja047905n	Photoluminescent,Layered Lanthanide Silicates

Table **??** shows that this particular research community referes mainly to toxicology studies neonicotinoids, bees and flowers[3]. The connections mostly make sense. Note the surprising inclusion of the cosmetics and lanthanide silicate studies. Upon investigation, both studies use very similar analytical techniques used elsewhere in the community, and both examined intercalation. [4]

Note also that the mean vector for the community was closest to a paper in the training set that summarised the community extremely well. This paper could be considered as a *Summary paper.* The uses of this kind of analysis include:

- Analysis of literature field - plotting trees such as figure **??** can give a relational understanding of how facets of a field link up together.
- Research tool: If researching a paper, finding its community immediately provides

[3]Note only some members of the community are shown above. Care was taken to give a representative sample of all 15 articles. The rest refer to Neonicotinoid insecticide studies with honey bees, and honey bee affinity to corn and pollen

[4]Both used made use of powder X-ray diffraction, and the silicates paper used thermogravimetry, the cosmetics study uses FID and several types of liquid chromatography, all methods used in the bee/nicotinoid studies.

the researcher with papers that are related to it. Crucially this is done without simply following citations, so that interesting, perhaps overlooked links between papers can be found.

- Summarising: If a researcher is required to read many papers from a field, they could find the communities involved and begin by reading the 'summary' papers.

1.2 Cambridge Staff Member Similarities

It is not only articles themselves that can be grouped and analysed, but articles can be aggregated together to represent higher collections, such as staff members or research groups, or potentially even departments. To investigate this further, `http://www.ch.cam.ac.uk/publicati` was scraped in order to associate the documents in the CCD with particular staff members groups. A staff member vector $\mathbf{f}$ was defined as $\mathbf{f} = \frac{1}{N}\sum_i^N \mathbf{v}_i$, for an author with N published articles in the CCD, with document vectors $\mathbf{v}_i$ (vector mean).

To investigate author relationships, a cosine matrix was created for each pair of authors A and B, with α and β documents respectively, $\mathbf{C}^{(\mathbf{A}),(\mathbf{B})}$ (see section**??**). The similarity between the author pair was defined as

$$S_{A,B} = \sum_i^\alpha \sum_j^\beta C_{i,j}^{(A),(B)}$$

An Author similarity matrix can then be built up $\mathbf{S}$, with elements $\mathbf{S}_{A,B} = S_{A,B}$. A similar technique could have been used to create clusters of authors. However, the sample size was now much smaller, so a more appropriate technique was hierarchical clustering, specifically UPGMA [5] **?**. This method clusters the authors pairwise in a hierarchical fashion. An effective visualisation of the similarities between staff was to plot a *clustermap* **? ?**.

[5] Unweighted Pair Group Method with Arithmetic Mean

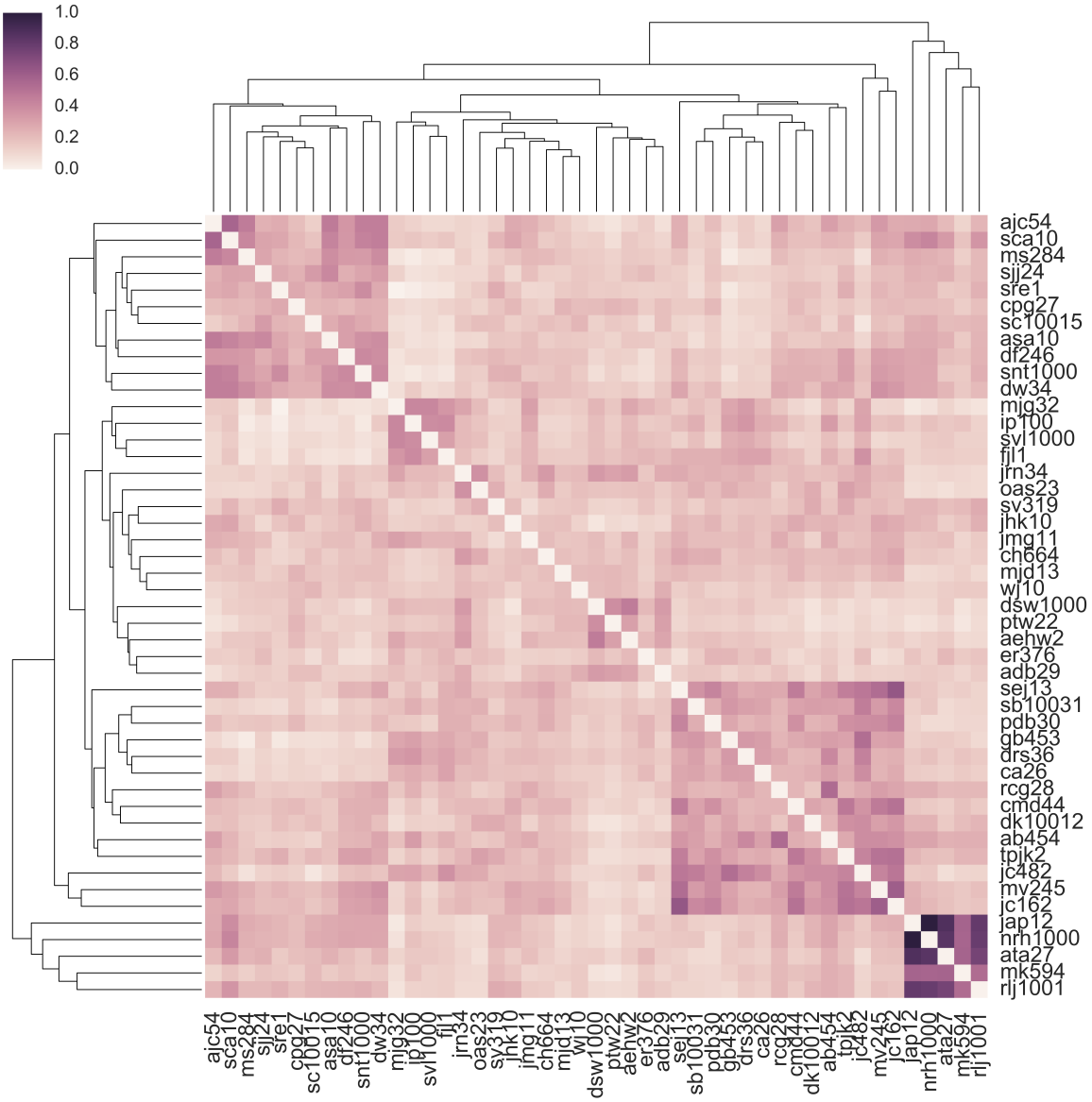

Figure 1.3: This figure shows a heatmap of Author similarity. Dark pixels correspond to the author in the pixel's row having similar research interests to the author in the pixcel's column. The authors are arranged by clusters found in UPGMA of authors. The authors are also diagrammatically connected by a dendrogram. This shows the hierarchical structure of the clustering.

Figure **??** shows the result of generating **S** and performing a UPGMA hierarchical clustering. The authors are labelled by their crsids. The dendrogram tree links authors pairwise, illustrating how the clustering was performed, and how closely related clusters are. An enlarged dendrogram is shown below:

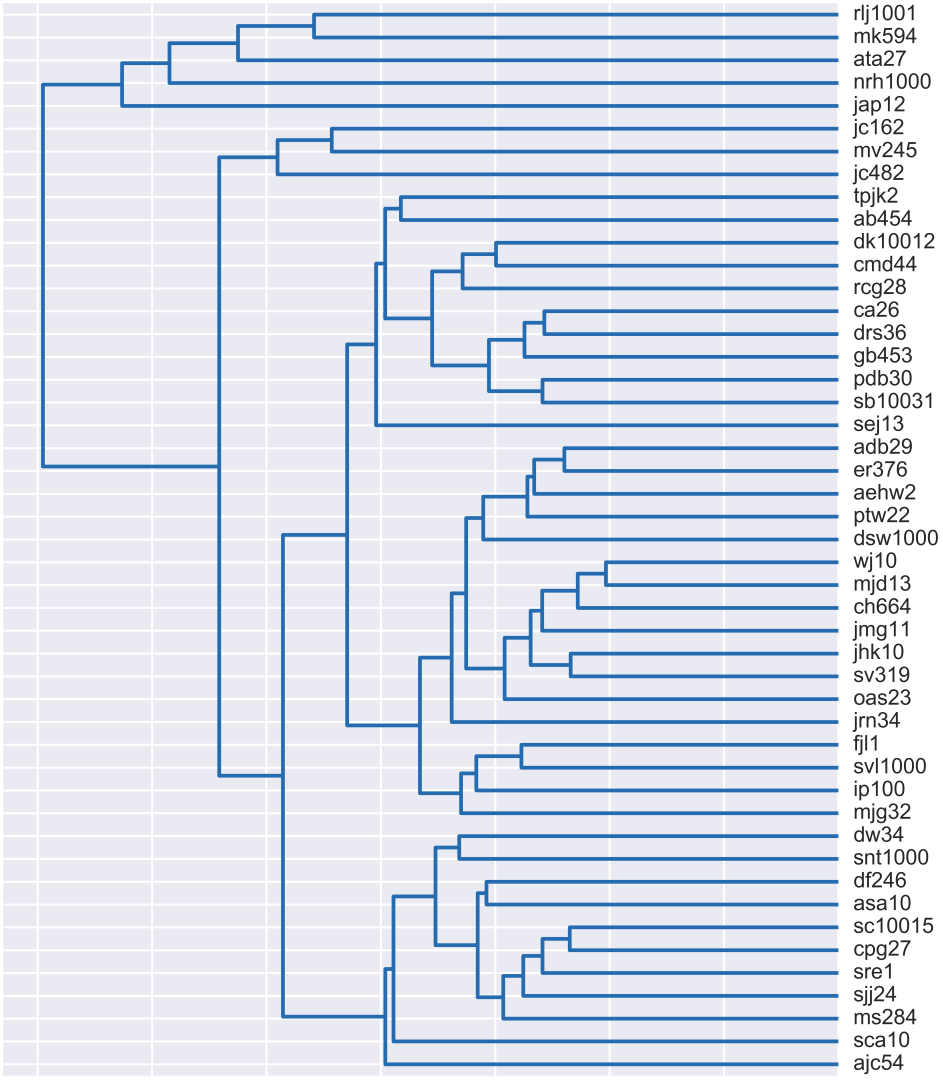

Figure 1.4: The dendrogram of figure **??** plotted for clarity

A striking feature of figure **??** is the cluster in the bottom right corner. The dendro-

gram tree shows the members of this cluster occupy a separate branch of research space than the rest of the department. The staff members involved, Professor Jones, Dr. Harris, Professor Pyle Dr. Archibald and Dr. Kalberer are all members of the Centre for Atmospheric Science. The unsupervised model thus successfully 'predicted' their department, and indicated that their work is quite far removed from most of the work in the Chemistry Department. This is a real success for the model. The dendrogram was then further examined and broken into distinct branches. Each branch was examined and manually labelled. The results are shown in figure **??**. Most clusters make intuitive sense, but there is one core of well connected, more disparate members (wj10 to jrn34). These members could be interpreted as being an interdisciplinary cluster.

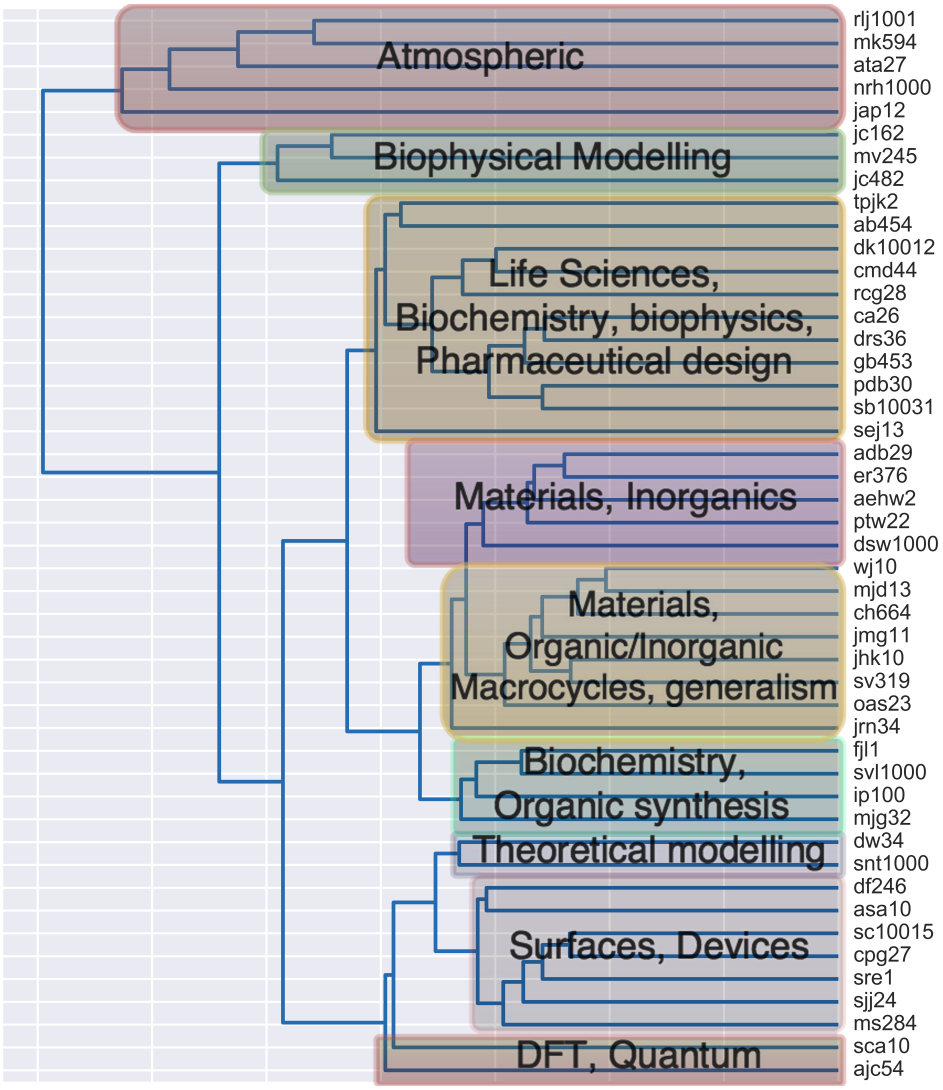

Figure 1.5: Cluster labels overlayed over the distinct branches of the dendrogram.

The analysis's value is self evident. Clusters of similar staff members informs the department about the width of research (number of clusters), and how resources are partitioned (size of clusters). It should also be stressed that authors are associated without any human preconceptions or bias. Thus perhaps the most valuable author associations are the unexpected ones, and authors should be encouraged to examine their cluster and consider their 'neighbours'.

1.3 Combining research clusters and authors

Finally, topic communities found in section **??** were linked to the authors in the department. Firstly, for a topic community $\mathfrak{C}$, with documents $d \in \mathfrak{C}$, and an author $\mathfrak{A}$ with documents $\delta \in \mathfrak{A}$, we can associate the author with the community if $\mathfrak{C} \cap \mathfrak{A} \neq \{\}$, [6]. The function f_{assoc} was defined as

$$f_{assoc}\left(\mathfrak{C}, \mathfrak{A}\right) = \begin{cases} 0 & \mathfrak{C} \cap \mathfrak{A} = \{\} \\ 1 & \mathfrak{C} \cap \mathfrak{A} \neq \{\} \end{cases}$$

It was noted that there was significant variation in the number of communities that researchers were associated with. A plot of $\sum_c^C f_{assoc}\left(\mathfrak{C}_c, \mathfrak{A}\right)$ for each author is shown below.:

[6] or equivalently $\exists \partial : \partial \in \mathfrak{C} \wedge \partial \in \mathfrak{A}$

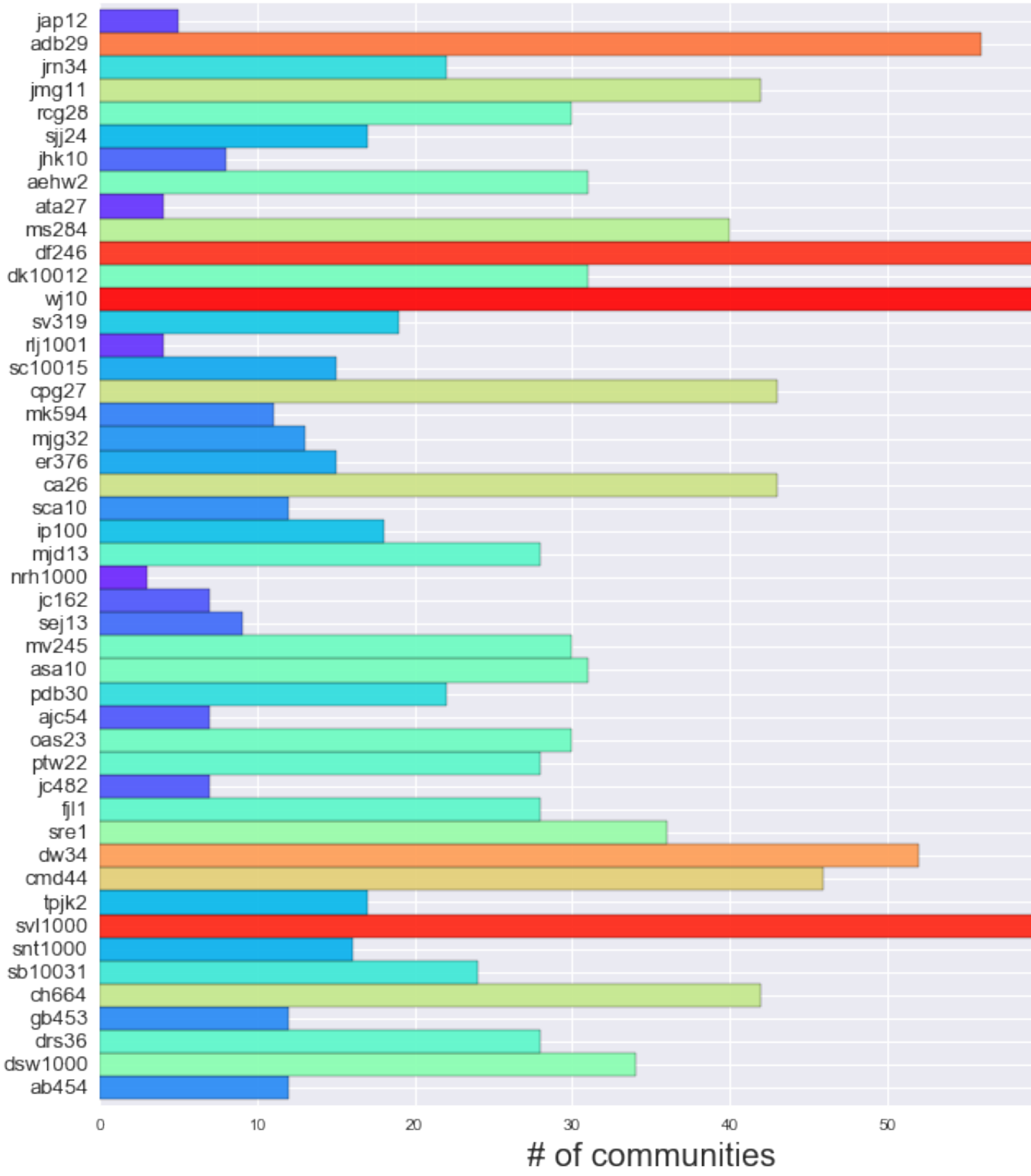

Figure 1.6: Number of research communities authors are associated with. High values indicate an author publishing across many communities, suggest more interdisciplinary work, but also higher publication count per author. (The same plot, scaled for publication count) is included in the appendix

It can be seen that some authors are widely distributed between communities, whereas others are concentrated. It should be appreciated that communities are not uniformly distributed across work done in the department. For example there are many communities in 'Life Sciences' but few in Atmospheric Chemistry, as such, interpretation of high values in Figure **??** corresponding 1:1 to wide research interests is cautious.

An association metric $S_{coincidence}$ between authors $\mathfrak{A}$ and $\mathfrak{B}$ was then defined as

$$S_{coincidence}\left(\mathfrak{A},\mathfrak{B}\right)=\sum_{c}^{C}\left(f_{assoc}\left(\mathfrak{C}_{c},\mathfrak{A}\right)f_{assoc}\left(\mathfrak{C}_{c},\mathfrak{B}\right)\right)$$

Where C is the total number of communities. An association matrix was created, $\mathbf{S}_{\mathfrak{A},\mathfrak{B}}^{Assoc}=S_{coincidence}\left(\mathfrak{A},\mathfrak{B}\right)$, where high values for author pair $\mathfrak{A},\mathfrak{B}$ indicate they appear in lots of research communities together. The matrix was then scaled such that: $\mathbf{S}_{\mathfrak{A},\mathfrak{B}}^{Assoc,scaled}=\mathbf{S}_{\mathfrak{A},\mathfrak{B}}^{Assoc}/\left(\mathbf{S}_{\mathfrak{A},\mathfrak{A}}^{Assoc}+\mathbf{S}_{\mathfrak{B},\mathfrak{B}}^{Assoc}\right)$, and normalised to the range 0,1. This was a measure of how often researchers can be found in the same communities. The matrix is shown below.

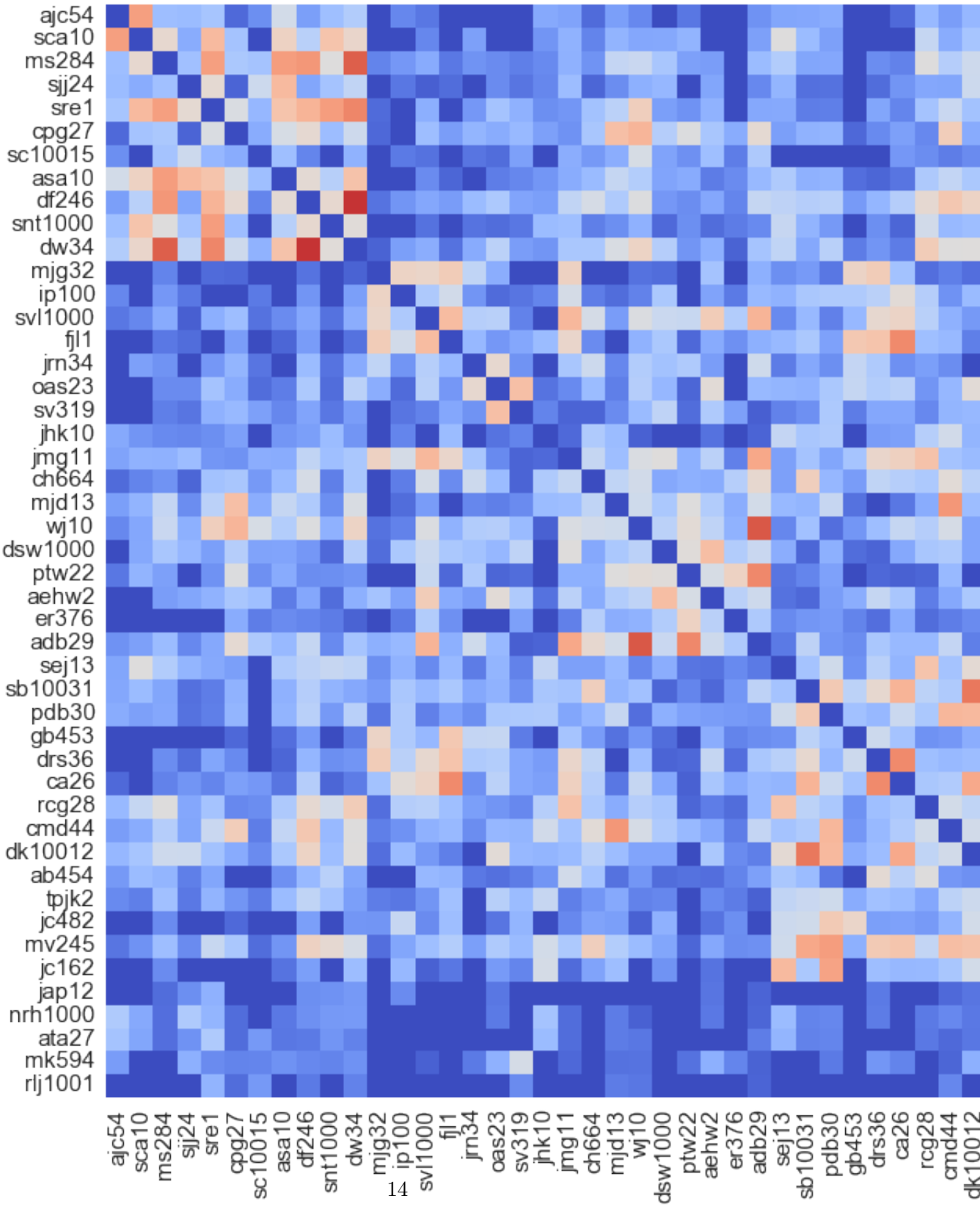

Figure 1.7: Heatmap showing author-author pair values for how often authors publish works in the same communities. High values indicated that Authors are predicted to have similar publication profiles. Note the authors are arranged with the ordering from figure **??**.

Figure **??** indicates where authors have been detected to have similar research community occupations. High values should indicate that authors should ideally collaborate/communicate because they publish in the same research communities. Note also that the square patterns near the diagonal of the plot reproduce the clustering in figure **??**, corroborating the trend, lending weight to the validity of both analyses.[7].

[7]This is because the heatmap has been arranged according to the clustering found in section **??**, but the matrix is derived in a completely different way. Because the clustering is qualitatively visible in figure **??**, the two methods correlate one another

2. Conclusions

To Do

3. Recommendations for Further Work

Bibliography

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013a. URL `http://arxiv.org/abs/1310.4546`.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013b. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=189726`.

A. Ullah and D.E.A. Giles. *Handbook of Empirical Economics and Finance*. Statistics: A Series of Textbooks and Monographs. CRC Press, 2010. ISBN 9781420070361. URL `https://books.google.co.uk/books?id=QAUv9R6bJzwC`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013c. URL `http://arxiv.org/abs/1301.3781`.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. URL `http://arxiv.org/abs/1405.4053`.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495, 9780596516499.

M. F. Porter. *An Algorithm for Suffix Stripping*, volume 14. 1980.

Chris D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, November 1990. ISSN 0163-5840. doi: 10.1145/101306.101310. URL `http://doi.acm.org/10.1145/101306.101310`.

Ingo Feinerer and Kurt Hornik. *wordnet: WordNet Interface*, 2016. URL `https://CRAN.R-project.org/package=wordnet`. R package version 0.1-11.

Mike Wallace. *Jawbone Java WordNet API*, 2007. URL http://mfwallace.googlepages.com/jawbone.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Surasak Seesukphronrarak and Toshikazu Takata. Novel fluorene-based biphenolic monomer: 9, 9-bis(4-hydroxyphenyl)-9-silafluorene. *Chem. Lett.*, 36(9):1138–1139, 2007. doi: 10.1246/cl.2007.1138. URL http://dx.doi.org/10.1246/cl.2007.1138.

Yu. B. Tsaplev. Photochemical transformations of anthraquinone in polymeric alcohols. *Russian Journal of Physical Chemistry A*, 86(12):1909–1914, oct 2012. doi: 10.1134/s0036024412120266. URL http://dx.doi.org/10.1134/s0036024412120266.

Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, nov 1901. doi: 10.1080/14786440109462720. URL http://dx.doi.org/10.1080/14786440109462720.

G.E. van der Maaten, L.J.P.; Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

4. Appendix

4.1 UK Departments scraped

Department	URL
Aberdeen	http://www.abdn.ac.uk/chemistry/
Aston	http://www.aston.ac.uk/eas/about-eas/academic-groups/ceac/
Bangor	http://www.bangor.ac.uk/chemistry/index.php
Bath	http://www.bath.ac.uk/chemistry/
Belfast (Queen's)	http://www.qub.ac.uk/schools/SchoolofChemistryandChemicalEnginee
Birmingham	http://www.birmingham.ac.uk/schools/chemistry/index.aspx
Bradford	http://www.brad.ac.uk/acad/chemistry/
Brighton	http://about.brighton.ac.uk/pharmacy/
Bristol	http://www.bris.ac.uk/Depts/Chemistry/Bristol_Chemistry.html
Cambridge	http://www.ch.cam.ac.uk/
Cardiff	http://www.cardiff.ac.uk/chemistry
Dundee	http://www.lifesci.dundee.ac.uk
Durham	http://www.dur.ac.uk/chemistry/
Edinburgh	http://www.chem.ed.ac.uk/
Essex	http://www.essex.ac.uk/bs/
Glasgow	http://www.chem.gla.ac.uk/
Greenwich	http://www.gre.ac.uk/engsci/study/pharchemenv
Heriot-Watt	http://www.eps.hw.ac.uk/institutes/chemical-sciences.htm
Hertfordshire	http://www.herts.ac.uk/research/hhsri/research-areas-hhsri/pharm
Huddersfield	http://www.hud.ac.uk/sas/chemistry/
Hull	http://www2.hull.ac.uk/science/chemistry.aspx
Keele	http://www.keele.ac.uk/chemistry/
Kent at Canterbury	http://www.kent.ac.uk/bio/
Kingston	http://sec.kingston.ac.uk/research/research-centres/
Lancaster	http://www.lancaster.ac.uk/chemistry/
Leeds	http://www.chem.leeds.ac.uk/
Leicester	http://www.le.ac.uk/chemistry/
Lincoln	https://www.lincoln.ac.uk/home/chemistry/
Liverpool	http://www.liv.ac.uk/chemistry/
Liverpool John Moores	https://www.ljmu.ac.uk/about-us/faculties/faculty-of-science/sch
London Metropolitan	http://www.londonmet.ac.uk/faculties/faculty-of-life-sciences-ar
Loughborough	http://www.lboro.ac.uk/departments/chemistry
Manchester	http://www.manchester.ac.uk/chemistry/
Manchester Metropolitan	http://www.sste.mmu.ac.uk
Newcastle	http://www.ncl.ac.uk/chemistry/
Northumbria	https://www.northumbria.ac.uk/about-us/academic-departments/appl

Department	URL
Nottingham	http://www.nottingham.ac.uk/chemistry/
Nottingham Trent University	http://www.ntu.ac.uk/sat/about/academic_teams/chemistry
Open Univserity	http://www.open.ac.uk/science/chemistry/
Oxford	http://www.chem.ox.ac.uk/
University of the West of Scotland	http://www.uws.ac.uk/schools/school-of-science/departme
Plymouth	https://www.plymouth.ac.uk/schools/school-of-geography-
Reading	http://www.reading.ac.uk/chemistry/
Robert Gordon	http://www.rgu.ac.uk/about/faculties-schools-and-depart
St Andrews	http://ch-www.st-and.ac.uk/
Salford	http://www.salford.ac.uk/environment-life-sciences/resea
Sheffield	http://www.sheffield.ac.uk/chemistry
Sheffield Hallam	http://www.shu.ac.uk/schools/sci/chem/
South Wales	http://www.southwales.ac.uk/chemistry/
Southampton	http://www.soton.ac.uk/chemistry/
Strathclyde	http://www.strath.ac.uk/chemistry/
Sunderland	http://www.sunderland.ac.uk/ug/subjectareas/pharmacychem
Surrey	http://www.surrey.ac.uk/chemistry/
Sussex	http://www.sussex.ac.uk/chemistry/
Teesside	http://www.tees.ac.uk/schools/sst/
UEA	http://www.uea.ac.uk/chemistry
Warwick	http://www2.warwick.ac.uk/fac/sci/chemistry/
York	http://www.york.ac.uk/depts/chem/
Bradford Ploymer IRC	http://www.brad.ac.uk/acad/irc/
Cardiff Pharmacy	http://www.cardiff.ac.uk/pharmacy-pharmaceutical-science
Burbeck Chemistry	http://www.bbk.ac.uk/bcs/
Burbeck Crystallography	http://www.cryst.bbk.ac.uk/
Imperial College London	http://www.imperial.ac.uk/chemistry/
King's College London	http://www.kcl.ac.uk/nms/depts/chemistry/index.aspx
Queen Mary London	http://www.sbcs.qmul.ac.uk/
UCL School of Pharmacy	http://www.ucl.ac.uk/pharmacy
University College London	http://www.ucl.ac.uk/chemistry/
Sheffield Computational Chemistry	http://www.sheffield.ac.uk/is/research/groups/chemoinfor

4.2

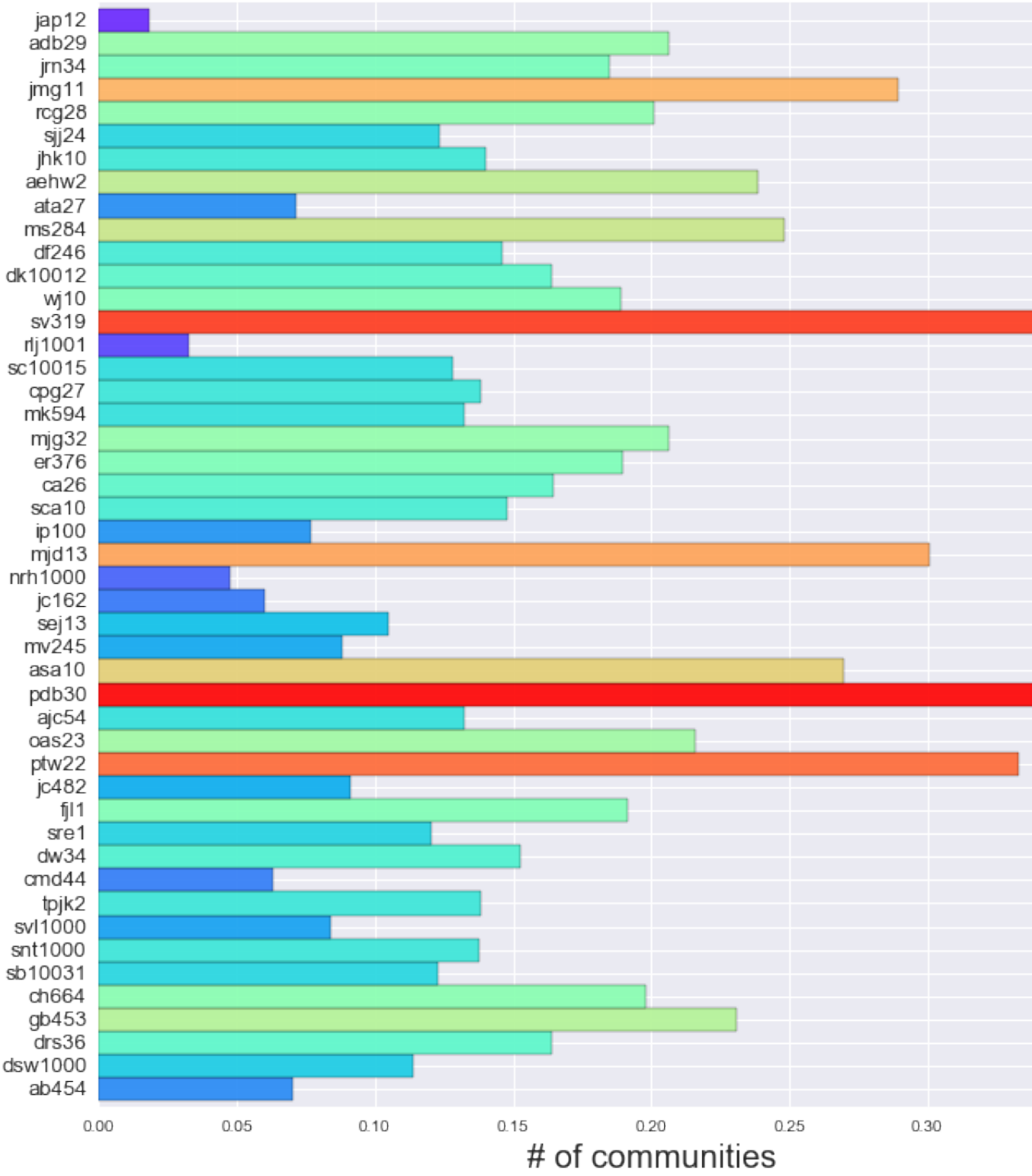

Figure 4.1: Number of research communities authors are associated with. High values suggest more interdisciplinary work. The bars have been scaled relative to the author's total publication count.