Citing Alike, Writing Alike: Comparing Discourse- and Bibliographic Coupling-Based Science Maps

Bradford Demarest[1], Cassidy R. Sugimoto[2], and Vincent Larivière[3]

[1]*bdemares@indiana.edu*
Indiana University, School of Informatics, Computing, and Engineering, Department of Information and Library Science, 611 N. Woodlawn, 101B, Bloomington, IN 47408

[2]*sugimoto@indiana.edu*
Indiana University, School of Informatics, Computing, and Engineering, Department of Information and Library Science, 919 E. 10th St., 263, Bloomington, IN 47408

[3]*vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7, Canada

Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8, Canada

Abstract

This study proposes a way to map the sciences based on social and epistemic cultural features in writing that can expose heretofore unexposed connections between disciplines. A network based on social and epistemic term frequencies in 1,269,146 journal articles from 14 disciplines is created and compared to a network of the same articles based on bibliographic coupling at the discipline level. The two networks are found to correlate moderately (0.577) with a p-value of 0.0002, and hierarchical clustering conducted on the networks show connections between Health and Clinical Medicine based on bibliographic coupling, and between Health, Psychology, and Social Sciences among others, based on social and epistemic terms in writing.

Introduction

A conflict exists in contemporary science regarding interdisciplinarity. On one hand, interdisciplinary research is widely heralded as critical to solving complex global issues such as climate change (Rylance, 2015). interdisciplinary research has the capacity for great success; Larivière, Haustein, and Börner (2015) find that in a study of 9.2 million papers from 2000 to 2012, a majority of co-cited interdisciplinary research papers result in higher relative citation counts for citing papers, with the highest relative citation counts reserved for interdisciplinary papers that draw from distant disciplines.

On the other hand, interdisciplinary research has been found to have consistently lower success in acquiring funding than disciplinary research (Bromham, Dinnage, & Hua, 2016); this would seem to reflect the perspective of some scholars that interdisciplinary research suffers when evaluated from traditionally disciplinary perspectives (Rylance, 2015). One reason that these evaluations may be hard on interdisciplinary research is due to differing social and epistemological norms in different disciplines, leading evaluators to see interdisciplinary work as an unsatisfactory version of scholarship from the evaluator's discipline rather than a culturally related but distinct product.

Mapping the socio-epistemic cultures of the disciplines is, then, an important first step toward accounting for such disciplinary clannishness and thus eventually opening the door for more productive interdisciplinary innovation. The current study begins this work, using a computational linguistics method (i.e., discourse epistemetrics (DE), per Demarest & Sugimoto, 2015) to extract social and epistemological disciplinary cultural information from scholarly article abstracts. This information is summarized as a pairwise distance metric, which we then use to derive a network of

disciplines. As a basis for comparison, we also create disciplinary networks from references for the same papers. We compare the two networks using Quadratic Assignment Procedure (QAP), graphically based on heatmaps, and with hierarchical clustering.

Literature Review

A wide variety of scholarship in sociology of science bears out the multitudinous ways in which new knowledge is created and verified (Becher & Trowler, 2001; Whitley, 1984). Furthermore, the writing of different scientific communities reflects these different disciplinary identities through social and epistemic language (Argamon, Dodick, & Chase, 2008; Cronin, 2005; Hyland, 2000). While science has been mapped using other measures of similarity including bibliographic coupling (Kessler, 1963; Boyack & Klavans, 2010), co-citation (Small, 1973; Boyack & Klavans, 2010, White & McCain, 1998), and co-authorship analysis (Glänzel & Schubert, 2005), no studies so far have attempted to map science based on social and epistemic written discourse terms. Leveraging the discourse epistemetrics method we previously established as both accurate and interpretable (Demarest & Sugimoto, 2015), the current study undertakes this mapping effort.

Methods

To study disciplinary social and epistemic features in academic writing, this method uses a sample of journal article abstracts from the Web of Science, taken from a single publication year. These abstracts are transformed into frequency vectors of lexical features that previous scholars have found to be indicative of different types of stance. After this transformation, support vector models (SVMs) are then generated for each disciplinary pair, with the accuracy of the model used as a measure of socio-epistemic distance between the disciplines. These accuracy measures are then used to describe the collection of disciplines as a network, which can be compared with a network of disciplines created based on patterns of references. Each of these aspects, including the specifics of sample, features, and model parameterization, are discussed in further detail below.

Sample

The current study utilizes abstracts and references for 1,269,146 English-language scholarly articles from the Web of Science from 2011. Articles from a single year with available abstracts were chosen to avoid any temporal effects on disciplinary socio-epistemic cultures and writing. For article counts by discipline, see Table 1.

Discourse Epistemetrics Features

Each abstract is first converted to a vector of relative frequencies of 568 social and epistemic terms collected from previous scholarship of social and epistemic stance in writing (Biber, 2006; Biber & Finegan, 1989; Hyland, 2005). These terms were found by the scholars to serve one of several functions. Hedging terms mitigate the certainty of an assertion; examples include “perhaps”, “approximately”, or “seem”. Conversely, boosting terms amplify assertions, e.g., “obviously”. Terms that frame an assertion emotionally or judgmentally are affective markers, including terms such as “unfortunately” and “surprisingly”. Aside from these, two other sets of socio epistemic terms exist – those that refer to the author herself (self-references such as “I”, “we”, or “the author”), and those that refer to the reader directly or implicitly (such as “the reader”, and “you”, as well as imperative verbs). For a full list of features, please contact the first author.

Discourse Epistemetrics Model Parameterization

After preparing the data, pairs of disciplines or specializations were then used to train and test SVMs. The LinearSVC from Python’s scikit learn toolkit (Pedregosa et al., 2011) was employed, and thus a linear kernel, such that feature weights could be analyzed. Per Varma and Simon (2006), we used a grid search approach to hyperparameter optimization of C (the total error value). In order to

avoid bias deriving from uneven sample sizes, balanced error values by category size were used. Finally, 10-fold cross validation was employed, with accuracy values averaged across the 10 cycles, to minimize variation due to assignment of samples to the training or test data sets. The resulting average accuracy measures for each disciplinary (or specialization) pair was then used as a distance metric – the higher the accuracy of the optimized model, the more distinct the two disciplines are from one another in terms of the social and epistemic discourse they use.

Table 1. Counts of Web of Science articles by discipline.

Discipline	**Articles**
Arts	1731
Biology	93765
Biomedical Research	153166
Chemistry	129685
Clinical Medicine	340574
Earth and Space	70018
Engineering and Technology	172949
Health	28343
Humanities	13673
Mathematics	42685
Physics	121702
Professional Fields	34590
Psychology	25802
Social Sciences	40463
Total	**1269146**

Disciplinary categories are taken from the U. S. National Science Foundation (NSF) field classification (Hamilton, 2003).

Bibliographic Coupling

To form a reference-based network at the disciplinary level, a matrix of reference counts per discipline was collected, with each row reflecting counts for a given referring discipline, and each column reflecting number of papers for a given discipline referenced by the row-discipline. Cosine distance was then used to calculate distance between each pairwise combination of disciplines. The process was repeated at the specialization level.

Findings

The findings presented here constitute summaries and visualizations of the study's data; for item-level information (such as cosine distance or accuracy for a given disciplinary pair), please contact the first author. Table 2 presents summary statistics for discipline-level networks based on discourse epistemetrics (for which the numbers are accuracy rates) and on bibliographic coupling (for which values reflect cosine distance).

Table 2. Summary statistics for discipline-level networks.

	Maximum Value	Minimum Value	Median
DE (accuracy)	0.988	0.612	0.887
BC (cosine distance)	0.983	0.132	0.898

Notably, pairwise models based on interactive metadiscourse term frequencies achieve accuracy rates of as high as 98.8%, and even the lowest accuracy models improve upon the baseline of 50% accuracy

by 11%. Cosine distance based on bibliographic coupling by discipline reflects a wider range. However, even taking this difference between distributions into account, we ran a 5000-iteration Quadratic Assignment Procedure analysis via UCINET (Borgatti, Everett, & Freeman, 2002) that compared discourse and reference-based distance matrices that yielded a Pearson's Correlation of 0.577 (p= 0.0002), suggesting that moderate correlation does exist between disciplines that cite alike and those that write alike.

Figure 1 presents a heatmap of disciplines based on discourse epistemetrics measures. Of the disciplines shown in Figure 1, the closest disciplines (i.e., those with the lowest DE accuracy scores) are Clinical Medicine and Biology; Social Sciences and Professional Fields; Biology and Biomedical Research; Biomedical Research and Clinical Medicine; and Physics and Engineering. Disciplines with the highest DE accuracy scores (and thus furthest apart) are all paired with Arts: Biomedical Research, Physics, Engineering and Technology, Biology, and Clinical Medicine.

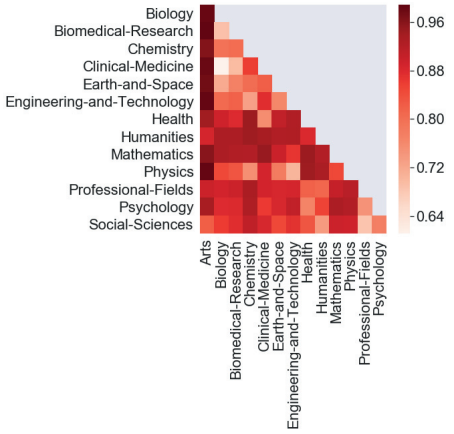

Figure 1. Heatmap of distances between disciplines (DE, accuracy).

Figure 2 presents a disciplinary heatmap showing cosine distance based on discipline-level bibliographic coupling.

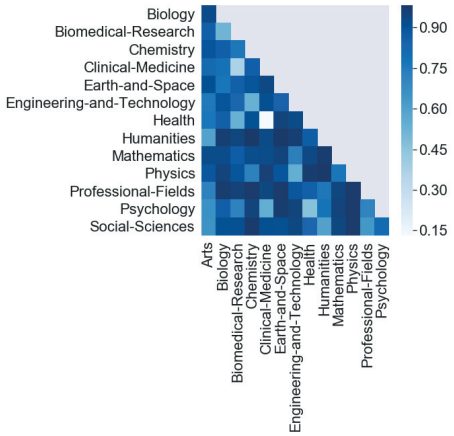

Figure 2. Heatmap of distances between disciplines (Bibliographic Coupling, cosine).

In Figure 2, the closest disciplines (i.e., with the lowest cosine distance) are Health and Clinical Medicine, followed by Biomedical Research and Clinical Medicine; Health and Psychology; Biology and Biomedical Research; and Biomedical Research and Health. Discipline pairs that are

furthest apart by bibliographic coupling measure are Chemistry and Humanities, Physics and Humanities, Earth and Space and Humanities, Earth and Space and Professional Fields, and Chemistry and Professional Fields.

Hierarchical Clustering

Using the accuracy values from the DE modeling in one case and the cosine distances from the bibliographic coupling in the other, we next used the scipy implementation of hierarchical clustering (Jones, Oliphant, & Peterson, 2014) using Ward distance for each of the networks. Figures 3 and 4 below show the resulting dendrograms.

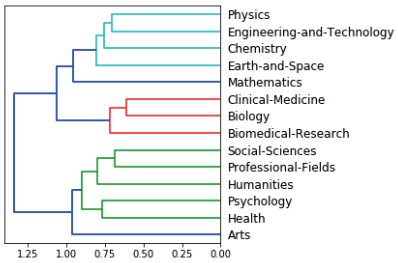

Figure 3. Disciplines clustered using hierarchical clustering (discourse epistemetrics, accuracy)

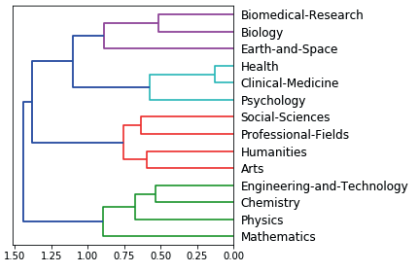

Figure 4. Disciplines clustered using hierarchical clustering (bibliographic coupling, cosine distance)

Figure 3 shows three clusters at the threshold of 1.00 – one for physical sciences, one for the biological sciences, and the last containing human-oriented and applied fields. The last of these clusters notably contains Psychology as well as Health, while the second cluster contains Biology, Clinical Medicine, and Biomedical Research. In contrast, Figure 4 contains four clusters at the same threshold. As before, a physical science cluster and a humanities-social science-professional cluster exist, but Biology, Biomedical Research, and Earth and Space disciplines occupy a separate cluster from Health, Clinical Medicine, and Psychology.

Discussion and Conclusions

This study has established that the discourse epistemetrics method can serve as a useful tool for mapping disciplines in recognizable constellations, and that differences in discourse networks and bibliographic coupling networks expose meaningful differences. Foremost among these is the distinction between Health as a discipline that writes most similarly to fields such as Social Sciences and Psychology, while citing similarly to the biomedical fields; this lays bare the interstitial nature of the Health field in particular, and the pipeline of the biological sciences from research fields (Biology and Biomedical Research) to Clinical Medicine, and then on to Health (with its emphasis on public policy). In consideration of the paradox of interdisciplinary research, it is hoped that this line of research will help to clarify differences when disciplines cite alike but write (and work) differently.

References

Argamon, S., Dodick, J., & Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. Scientometrics, 75(2), 203–238.

Becher, T., & Trowler, P. R. (2001). Academic Tribes and Territories: intellectual enquiry and the cultures of disciplines (2nd edition). Retrieved September 6, 2012, from http://eprints.lancs.ac.uk/3714/

Biber, D. (2006). University language: a corpus-based study of spoken and written registers. Amsterdam ; Philadelphia: J. Benjamins.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. Text, 9(1), 93–124.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for social network analysis.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389–2404. https://doi.org/10.1002/asi.21419

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. Nature, 534(7609), 684–687. https://doi.org/10.1038/nature18315

Cronin, B. (2005). The Hand of Science: Academic Writing and Its Rewards. Scarecrow Press.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. Journal of the Association for Information Science and Technology, 66(7), 1374–1387. https://doi.org/10.1002/asi.23271

Glänzel, W., & Schubert, A. (2005). Analysing Scientific Networks Through Co-Authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems (pp. 257–276). https://doi.org/10.1007/1-4020-2755-9_12

Hamilton, K. (2003). Subfield and Level Classification of Journals (CHI Report No. 2012-R) (No. CHI Report No. 2012-R). Cherry Hill, NJ: CHI Research.

Hyland, K. (2000). Disciplinary discourses: Social interaction in academic writing. Retrieved from http://www.lavoisier.fr/livre/notice.asp?id=OASW3KAR6OKOWH

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. Retrieved from http://books.google.com/books?hl=en&lr=&id=jgfgHpEqPN8C&oi=fnd&pg=PR8&dq=epistemic+metadiscourse+discipline&ots=60Ofr4zJRL&sig=ZYs8Z8BASLtXs3GGExodyDm07h0

Jones, E., Oliphant, T., & Peterson, P. (2014). SciPy: Open source scientific tools for Python.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25.

Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. Plos One, 10(3), e0122565.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, *12*(Oct), 2825–2830.

Rylance, R. (2015). Grant giving: Global funders to focus on interdisciplinarity. Nature News, 525(7569), 313. https://doi.org/10.1038/525313a

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4), 265–269. https://doi.org/10.1002/asi.4630240406

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49(4), 327–355.

Whitley, R. (1984). The intellectual and social organization of the sciences. New York, NY: Clarendon Press.