Delineating Fields Using Mathematical Jargon

Jevin D. West[1] and Jason Portenoy[1]

Information School, University of Washington, jevinw@uw.edu

Abstract. Tracing ideas through the scientific literature is useful in understanding the origin of ideas and for generating new ones. Machines can be trained to do this at large scale, feeding search engines and recommendation algorithms. Citations and text are the features commonly used for these tasks. In this paper, we focus on a largely ignored facet of scholarly papers—the equations. Mathematical language varies from field to field but original formulae are maintained over generations (e.g., Shannon's Entropy equation). Here we extract a common set of mathematical symbols from more than 250,000 LaTeX source files in the arXiv repository. We compare the symbol distributions across different fields and calculate the jargon distance between fields. We find a greater difference between the experimental and theoretical disciplines than within these fields. This provides a first step at using equations as a bridge between disciplines that may not cite each other or may speak different natural languages but use a similar mathematical language.

Introduction

There has been considerable effort building and designing new recommendation algorithms to help scholars find relevant papers. Most of these methods depend on citations [14], full text [5] or usage data [3]. One feature that has been largely ignored are equations and the mathematical language surrounding these equations. These formal languages can tie together papers and ideas across fields and time periods. Shannon's famous entropy equation (also used in this paper) is an example of this kind of trace [7]. Unlike natural languages, formal languages such as mathematics are exempt from plagiarism rules. The norm is for an author to copy an equation from the original source. This provides a unique opportunity for tracing ideas back to their origins and for tracking them forward in time.

There have been attempts at utilizing the equations as a search facet. For example, Springer's LaTeX Search tool[1] allows authors to search formulae (in LaTeX format) from more than 8 million documents in Springer journals and articles. This is used both as a tool for searching similar formulae and for translating formulae to existing documents. But what if a researcher wants to find not just individual manuscripts with the same equations, but fields of study and groups of papers using similar language? What kind of formalism can be used to map jargon differences across the quantitative sciences?

[1] http://latexsearch.com

In this paper, we measure the communication efficiency or "jargon distance" between fields using mathematical symbols. The jargon metric is derived from a recent paper by Vilhena et al. [11] which measures the distance between disciplines using n-grams extracted from millions of papers in the JSTOR corpus. In our paper, we find that the metric separates fields that are different both in content and mathematical notation[2].

Our ultimate research goal is to find ways to utilize equations and formal notation in scholarly recommendation. The more proximate goal of this paper is to validate that equation jargon can delineate the relationship among fields at the scale of hundreds of thousands of papers in similar ways to citation and text-based clustering. We show in this paper that the jargon metric proposed in this paper does a reasonable job at identifying fields with similar attributes and language. The next steps will involve full-corpus scaling, extracting mathematical grammar, and incorporating the metric into citation-based recommendation algorithms [13,14].

Methods

Data

Research papers were downloaded from arXiv.org[3], an open-access e-print service in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, and statistics. For this study, we used a sample (N = 266,906) of papers published between 2000 and 2009. We downloaded papers using the arXiv's bulk data download, analyzing papers that had field designations in their filenames. We compiled a list of the LaTeX representations of 103 symbols commonly used in mathematical formulae[4]. These symbols included some commonly used Greek letters (e.g. "`\alpha`" [α], "`\omega`" [ω]), arrows (e.g. "`\rightarrow`" [$\rightarrow$]), binary operation/relation symbols (e.g. "`=`", "`>`", "`\geq`" [$\geq$], "`\times`" [$\times$]), and some other symbols (e.g. "`\forall`" [$\forall$]).[5] A shortcoming of this approach is that it does not take into account the ability of authors to redefine commands and refer to symbols with these new command names, but it gives a rough idea of the usage of these symbols over the corpus. Using the LaTeX source for the arXiv papers, we counted the occurrence of each of these symbols. We used field designations provided by the arXiv; however, the fields could be determined by other methods, including citation clustering [14], co-citations [8], and topic modeling [12]. We intend to combine the jargon metric with these different clustering methods in future studies.

[2] These content comparisons were based on inspection of sample papers in the respective fields. More rigorous inspection is needed.

[3] `https://arxiv.org/help/bulk_data`

[4] Modified from `https://www.sharelatex.com/learn/List_of_Greek_letters_and_math_symbols`

[5] Our list of symbols is not exhaustive; for example, it does not include certain structural elements of equations such as sums. We intend to expand the list in future analysis.

Jargon Distance

To measure the communication barrier between fields, we adapt a metric developed by Vilhena et al. [11]. In this study, the authors develop a model of communication to topographically map the JSTOR corpus. Instead of using n-grams from full text like the Vilhena study, we use the language of mathematics—such as symbolic notation and Greek letters—extracted from LaTeX source files.

The jargon distance (E_{ij}) between field i and j is determined by calculating the ratio between two things: (1) the entropy H of a random variable X_i with a probability distribution based on the frequency of mathematical symbols within field i and (2) the cross entropy[6] Q between the distributions in field i and j,

$$E_{ij} = \frac{H(X_i)}{Q(p_i||p_j)} = \frac{-\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_i(x)}{-\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_j(x)} \tag{1}$$

This calculation derives itself from a general communication heuristic whereby a mathematician communicates with another mathematician via a channel [7]. The mathematician writing the formulae in field i has a codebook P_i that maps mathematical concepts to codewords, which mathematician in field j with codebook P_j has to decode. Note that the metric is not symmetrical—field j may more efficiently decode concepts in field i than field j does for field i. The model assumes that the codebooks are optimized based on the frequency of different terms. This power-law assumption holds well for English words [15,16]. It also seems to hold for mathematical terms; we find a Zipfian distribution in our sample (see Figure 1).

This simple metric of communication efficiency has advantages. It is based on a model of communication, which has firm theoretical foundations and additional tools to build upon. Second, it is easy to calculate and can run an entire corpus over a relatively short amount of time[7].

Results

Table 1 shows the top 20 symbols in our sample. The full distribution of symbols follows a power-like law (see Figure 1). This is important given the assumption of the jargon distance metric. We find large numbers of equal signs and inequality symbols but far fewer of others.

After extracting the symbols and distributions of symbols across fields, we calculated jargon distances between fields. We wanted to know if similar fields had smaller jargon distances when compared to more dissimilar fields. To do

[6] The cross entropy is the the entropy of X_i plus the Kullback-Leibler divergence between p_i and p_j [4].

[7] We calculated distances for 200k papers in less than 5 minutes on a micro EC2 instance with Amazon's Web Services (AWS).

Table 1. Top 20 mathematical symbols and letters among the papers sampled.

Symbol	Count	Symbol	Count
$>$	131,163,440	β	2,970,664
$<$	124,248,288	γ	2,876,255
$=$	119,559,867	ρ	2,827,491
$\in$	9,586,274	ϕ	2,820,464
μ	5,655,959	δ	2,677,700
α	5,193,305	θ	2,483,652
π	4,021,629	τ	2,482,211
ν	3,382,075	ω	2,443,861
λ	3,360,273	$\times$	2,320,697
σ	3,207,950	Δ	2,273,644

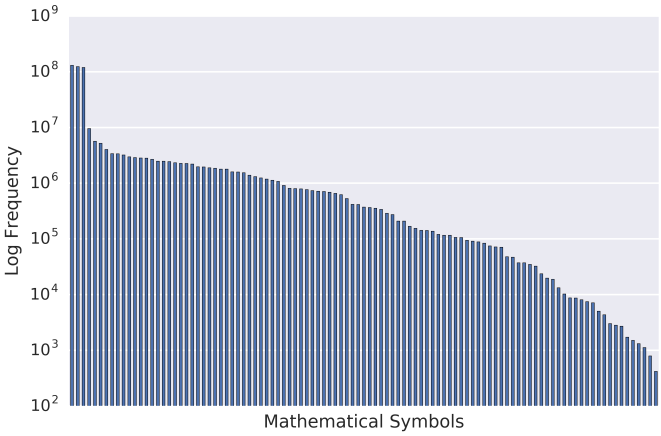

Fig. 1. Distribution of math symbols. The distribution of mathematical symbols in our sample. Most terms occur relatively infrequently compared to the equals (=) and inequality (>, <) symbols which occurred more than 10^8 times in equations and mathematical discussions.

this, we applied standard hierarchical clustering methods to the distance matrix in order to infer which fields were most like each other. We visualized the groupings using dendrograms. The dendrogram in Figure 2 was produced using a hierarchical clustering algorithm implemented in SciPy's linkage function[8]. We used UPGMA [9], which is an agglomerative method that uses average linkage as its criterion. The clustering is done on the adjacency matrix of fields where E_{ij} is the distance. Although the distance metric is not symmetric, for the clustering we symmetrize the matrix by taking the average of the the distances E_{ij} and E_{ji}.

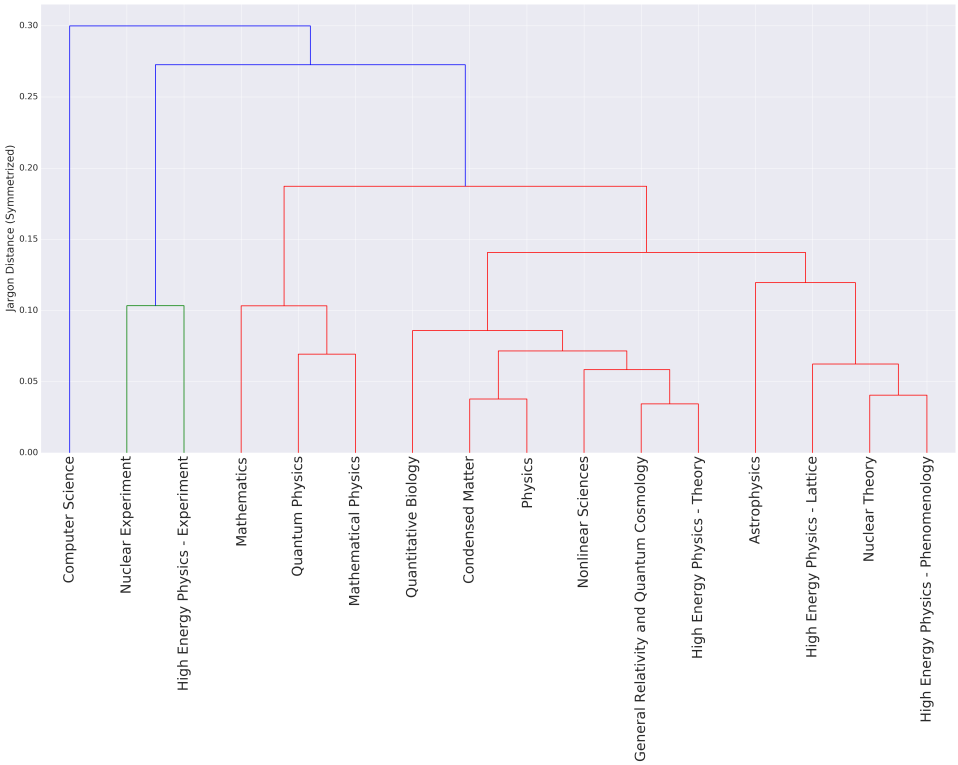

Fig. 2. Hierarchical Clustering. The dendrogram shows the relationships among the different fields represented in the arXiv. The jargon distance is used as the distance metric showing the relationship among disciplines within a sample of 266,906 papers in the arXiv. The distance between fields was determined using the mathematical jargon distance (E_{ij}) [11]. For this clustering, the distance between fields is the mean jargon distance of E_{ij} and E_{ji}.

[8] SciPy version 0.17.0 running on Python 2.7.6

We find a separation among theoretical and experimental sciences. Nuclear Experiment is most similar to High Energy Physics - Experiment (see Figure 2). We see Computer Science as an outlier to the rest of the fields. Mathematics shares a closer branch with Mathematical Physics than with any other discipline other than Quantum Physics. Quantitative Biology sits on its own but within the branch that includes Condensed Matter, Physics, Nonlinear Sciences, among others. We need to further investigate whether these similarities are real.

Figure 3 shows the distances between all fields using a heatmap. The darker colors represent larger jargon distances (lower communication efficiency between the writer and reader). The largest differences are between Computer Science and Experimental Physics, but Computer Science, in general, is quite different than almost all fields except Mathematics. Computer scientists and mathematicians seems to use similar symbols when conveying technical concepts. High Energy Physics (lattice) and High Energy Physics (phenomenology) are among the most similar, as one would expect.

Table 2 shows the top ten symbols for three fields: Computer Science, Nuclear Experiment, and High Energy Physics Experiment. The pairing from computer science to high energy physics experiment had a (mean) jargon distance of 0.171, whereas nuclear experiment to high energy physics experiment had a much smaller distance of 0.033. The top ten symbols help explain the differences (even though ten are not enough to see the full distribution differences).

Table 2. Comparing fields. This table compares two fields that have small jargon distances and two fields that have large jargon distances (see Figure 3). The (mean) jargon distance between the three pairings are the following (CS, Nuc Exp, **0.142**), (CS, HEP Exp, **0.171**), and (Nuc Exp, HEP Exp, **0.033**). The experimental sciences seem to use mathematical languages more similar to each other than to computer science. CS = Computer Science. Nuc Exp = Nuclear Experiment. HEP Exp = High Energy Physics Experiment.

CS		Nuc Exp		HEP Exp	
Symbol	Count	Symbol	Count	Symbol	Count
$=$	1,845,458	$>$	1,624,214	$>$	3,146,439
$>$	1,703,602	$<$	1,592,675	$<$	2,957,547
$<$	1,649,968	$=$	1,188,445	$=$	2,217,220
$\in$	253,658	$!=$	29,422	π	168,633
$\leq$	63,702	π	24,879	μ	100,176
α	63,302	$\in$	24,139	ν	85,847
μ	56,150	γ	18,637	$\in$	71,923
σ	45,041	$\cap$	17,218	γ	68,772
$\cap$	36,338	μ	14,692	η	65,464
λ	34,390	σ	13,867	$\rightarrow$	59,804

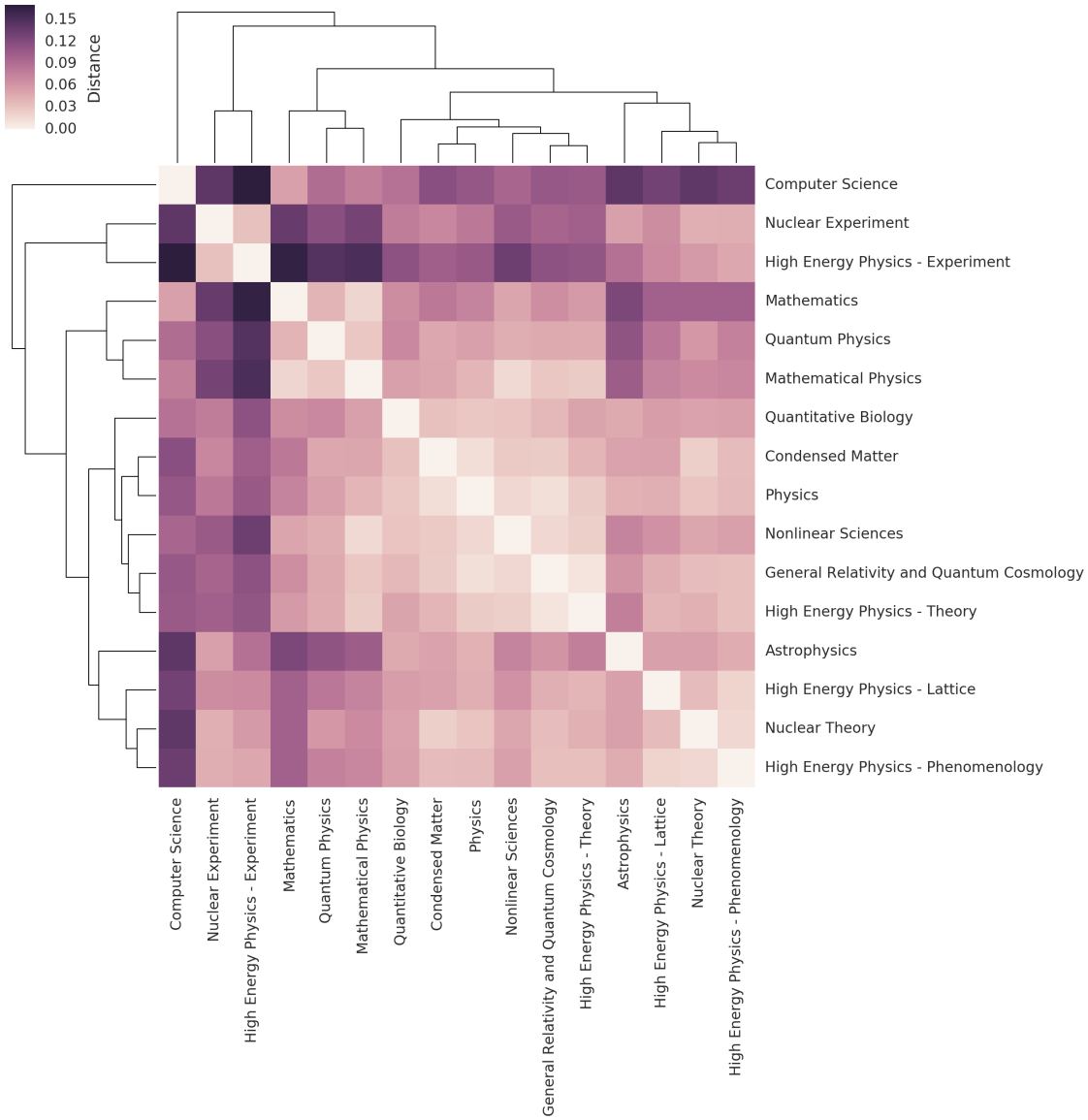

Fig. 3. Distance Matrix. The heatmap provides a visual representation of the distances between fields. We use the symmetrized version of the jargon distance matrix. The darker square represents a greater distance between two fields. For example, the frequency distribution of symbols in Computer Science is quite different than the mathematical notation found in Experimental High Energy Physics.

Discussion

Scientific papers contain many features that are used in search engines and recommendation algorithms: authors, titles, full text, citations, figures, etc. One feature largely ignored within the digital library community (at least relative to the other features) are equations. In this paper, we apply a model of communication efficiency between groups of papers first proposed by Vilhena et al. [11]. We find that field relationships (i.e., how fields are grouped hierarchcially) are recapitulated when using the jargon distance metric (Figure 2). The results are surprising given the simplicity of the data extraction and distance calculation. We only use isolated symbols and the frequency of these symbols to infer the groupings of fields. The resultant groups seem to assemble in logical ways (e.g., the experimental sciences group together, while the more theoretical fields assemble in another area of the tree). However, we see these results as preliminary evidence for using mathematical symbols as a way of clustering papers and topics.

We need to further investigate the true differences in fields. We plan to use citation based clustering as another means of field designation. We also plan to talk to scholars in the various fields to assess the validity of the clusters. In addition, we will extend our analysis to the full arXiv corpus. For corpora with no LaTeXavailable, we plan to use computer vision techniques from the viziometrics.org project for automatically extracting equations from PDFs [6]. We plan also to compare the jargon method to other well-known methods such as cosine similarity [10], LDA [1], and word2vec [2]. The primary difference we see from these methods is the communication theory underlying the jargon metric, but there needs to be analytic work for making this argument. In addition, we plan to expand beyond isolated symbols and analyze mathematical grammar.

Assuming the methods hold, our ultimate goal for this research project is to integrate our methods into existing recommendation engines at the scale of micro-fields. It is at these finer scales where the method could bridge seemingly disparate, emerging fields that are using similar mathematical language.

We also plan to extend this analysis to Science of Science questions, investigating the birth and death of ideas and the sociology surrounding these ideas. We see equations as an effective way of tracing ideas both forwards and backwards in time. The relative stability of equations and mathematical language provides a unique opportunity for tracking the movement of ideas across time and across disciplines.

Acknowledgements

We would like to thank three anonymous reviewers for their helpful feedback.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)

2. Goldberg, Y., Levy, O.: word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
3. Kantor, P.B., Rokach, L., Ricci, F., Shapira, B.: Recommender systems handbook. Springer (2011)
4. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)
5. Lü, L., Medo, M., Yeung, C.H., Zhang, Y.C., Zhang, Z.K., Zhou, T.: Recommender Systems p. 97 (Feb 2012), http://arxiv.org/abs/1202.1112
6. P. Lee, West, J., B. Howe: Viziometrix: A platform for analyzing the visual information in big scholarly data. In: Proceedings of the 25th International Conference on World Wide Web. ACM (2016)
7. Shannon, C.E.: The mathematical theory of communication, vol. 27 (1948)
8. Small, H.: Co-Citation in Scientific Literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science 24(4), 265–269 (1973)
9. Sokal, R.R.: A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38, 1409–1438 (1958)
10. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD workshop on text mining. vol. 400, pp. 525–526. Boston (2000)
11. Vilhena, D., Foster, J., Rosvall, M., West, J., Evans, J., Bergstrom, C.: Finding cultural holes: How structure and culture diverge in networks of scholarly communication. Sociological Science 1(June), 221–238 (2014), http://www.sociologicalscience.com/articles-vol1-15-221/
12. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 448–456. ACM (2011)
13. Wesley-Smith, I., Dandrea, R., West, J.: An experimental platform for scholarly article recommendation. In: Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR2015). pp. 30–39 (2015)
14. West, J., Wesley-Smith, I., Bergstrom, C.: A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Transactions on Big Data (in press) (2016)
15. Zipf, G.K.: The psycho-biology of language. Houghton, Mifflin (1935)
16. Zipf, G.K.: Human behavior and the principle of least effort. Addison-Wesley Press (1949)