9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

The dynamics of writing style in scientific collaborations

Keywords: Stylometrics, writing style, collaborations, networks, coauthorship networks

Extended Abstract

When two scientists write a paper together, does their joint writing style represent some average of their individual writing styles? Or does the joint writing style nearly match that of one author – that of the first author or the oldest author, perhaps? Writing style can be quantified using statistical methods from stylometry. One important application of stylometry is author attribution: Determining the most likely author of an anonymous text [1]. Whereas many stylometric methods have been developed to match texts with likely authors, little is known about the writing style in collaborations. In this work, we compare the writing styles in two-author paper abstracts from seven different scientific fields to the writing styles of the individual authors. For all the scientific fields, we characterize some of the ways in which the joint writing style of authors arises as a mixture of the individual writing styles. In particular, we find that the resulting joint writing style decreases the usage of the most common English words.

We use the Microsoft Academic Graph dataset to study the writing style in scientific collaborations. The data contains detailed information of more than 200 million scientific papers and 250 million authors. The detailed information available for many papers includes the paper's field of study, authors (and their listing order), publication year, publication venue, and abstract. We focus on two-author collaborations in the diverse fields of Ecology, Geography, Geology, History, Psychology, Sociology, and Zoology, where the abstract is in English and given in the data, and for which both authors have at least 5 publications in the data set.

To quantify the writing style in abstracts, we use stylometry methods based on word count and punctuation count [2, 3]. We represent the text style as a vector, $\mathbf{x}$, in a 30-dimensional vector space. The coordinates in the first 21 dimensions are the fraction of abstract words that the 21 most common English words in the dataset each account for (Table 1). The remaining 9 dimension coordinates are given by the fraction of punctuation marks that 9 common punctuation marks each account for in the text. Representing texts by counting words and punctuation (sequences) has previously been used for author and genre attribution [2, 3].

We estimate the personal writing style for each author by computing the mean of the writing styles in the author's abstracts. Although most papers are collaborative works, we find that this average can be used as a representation of personal writing style. We conclude this from an experiment, where we attempt to guess which of two authors (one being an actual author of the abstract, the other a uniformly-randomly chosen scientist publishing in the same field of study) who wrote a specific abstract. We find that for 3 different choices of distance functions, the author whose personal writing style had the smallest distance to the paper's style was indeed the actual abstract author for an overwhelming fraction of the 100 000 papers chosen uniformly at random ($2/5$-norm: 89.6% accuracy; Manhattan distance: 95.5% accuracy; Euclidean distance: 95.2% accuracy). Since Manhattan distance performed best in this attribution experiment, we use this distance measure going forward.

Having established estimates of individual writing styles, we proceed to investigating writing style in two-author abstracts. We examine seven hypotheses of how collaborative writing

style might depend on the writing style of the collaborating individuals and compare how well these hypotheses predict the empirical writing styles in 2-author papers. 4 of the hypotheses assume that the joint writing style is closest to one author – that of the first author, last author, youngest author or oldest author, respectively. The final 3 hypotheses hypothesize that the joint writing style can be estimated as a non-trivial mixture $f(\mathbf{x_1}, \mathbf{x_2})$ of the 30-dimensional representations of writing styles of the individual collaborators, $\mathbf{x_1}$ and $\mathbf{x_2}$. For these 3 hypotheses, the function $f(\mathbf{x_2}, \mathbf{x_2})$ is the element-wise minimum, element-wise maximum and element-wise average of pair of style vectors.

We compute the distance between the observed writing style and the seven predictions of writing style for each paper in the seven fields of interest. Figure 1 Left shows that the element-wise minimum hypothesis matches the empirically observed styles best across all fields. If we rank the performance of each prediction for every paper and compute the mean rank scored by the hypotheses, the element-wise minimum hypothesis also outperforms the others (Figure 1 Right). In Figure 2, we evaluate the performance of the predictions in the 30 individual dimensions. Whereas the element-wise minimum prediction performs best in most dimensions, one interesting exception is the dimension representing the frequency of commas in the abstracts. For this dimension, the prediction of the element-wise minimum hypothesis – that the closest match of comma frequency in the collaborative text will be the comma frequency of the author that most rarely uses commas – is the worst-performing hypothesis of all.

Our finding that the element-wise minimum hypothesis best predicts writing style of two-author abstracts could indicate that collaborating on a text makes the writing more varied. At least, when a smaller fraction of words and punctuation marks are accounted for by the 30 dimensions in our stylometry analysis, other words can be used in their absence. This observation leads to the hypothesis that more authors might increase diversity in language even more. In future work, we intend to test this hypothesis by measuring the word-level entropy of k-author abstracts and see whether the entropy increases with the integer value of k.

Going forward, we also intend to analyze other dimensions of collaborative writing. For example, it would be interesting to analyze text-level quantities such as abstract readability. Another interesting direction would be to analyze changes in writing style over time. How does writing style change over time, and do changes in writing style happen gradually or suddenly?

Stylometry has proven useful in quantifying and recognizing individual writing styles. Here, we used stylometric methods to characterize how collaborative writing style is a mixture of individual writing styles for two-author collaborations. However, much remains to be discovered about how writing style changes in collaborations with two or more participants.

References

[1] Frederick Mosteller and David L. Wallace. Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed *Federalist* Papers. *Journal of the American Statistical Association*, 58(302):275–309, June 1963.

[2] José Nilo G. Binongo. Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. *CHANCE*, 16(2):9–17, March 2003.

[3] Alexandra N. M. Darmon, Marya Bazzi, Sam D. Howison, and Mason A. Porter. Pull out all the stops: Textual analysis via punctuation sequences. *European Journal of Applied Mathematics*, 32(6):1069–1105, December 2021.

the	of	and	a	to	in	is	with	for	by	on	that	an	are	as
be	this	was	from	at	!	”	(	)	,	.	:	;	?	

Table 1: The 30 dimensions used to quantify writing style. The first 21 dimensions are the most common function words in a sample of 1% of the abstracts included in our analysis (corresponding to 2 398 554 abstracts). The final 9 dimensions are common punctuation marks.

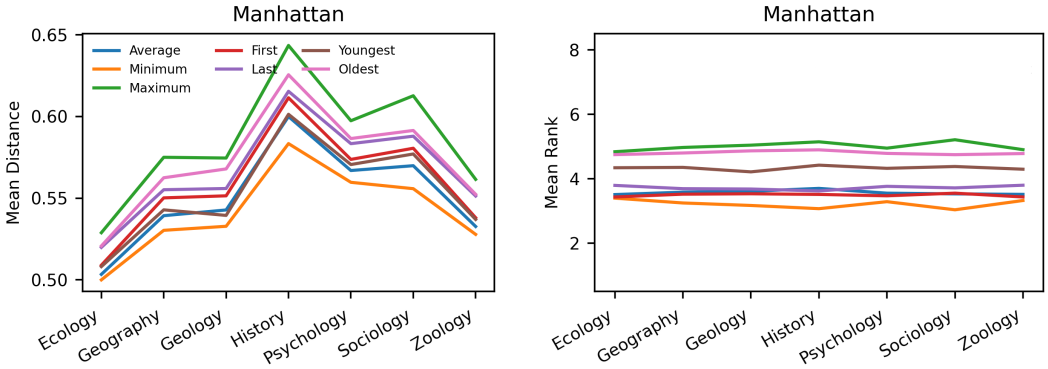

Figure 1: **Left** Mean Manhattan distance between hypothesis-based prediction of writing style and actual writing style for all fields. **Right** Mean rank of hypothesis prediction of writing style when distance between prediction and empirical style is computed with the Manhattan distance. Best prediction has rank 1, worst prediction has rank 7.

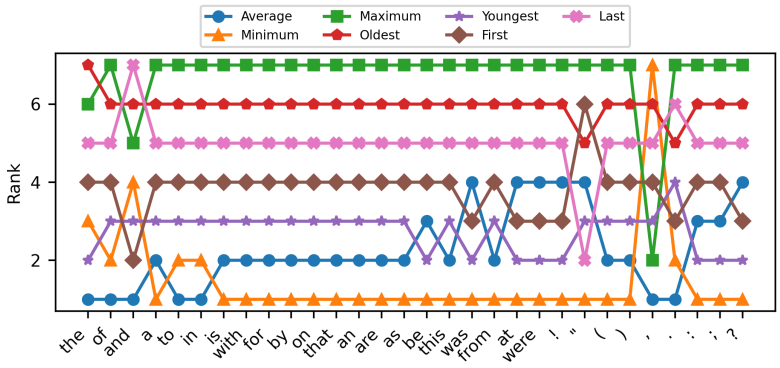

Figure 2: Hypothesis performance in the individual dimensions for papers in the field of sociology. Generally, the element-wise minimum hypothesis performs well in many dimensions (Rank 1 being the hypothesis whose prediction is closest to the empirical value). One interesting exception is the comma dimension, where joint writing generally usage commas closer to the average or maximum comma frequency of the individual authors.