Reply to Reviews

Submission: EPDS-D-20-00186
Manuscript: Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems **Authors:** Dodds et al.

Assessment from the Associate Editor:

The authors propose a very interesting method. We're sorry this review took a long time, but eventually we were able to secure three highly competent Reviewers. Reviewer 1 presents a few major reviewercomments and recommended a major revision. Upon closer inspection, however, those reviewercomments are mostly interpretational, but addressing them will make the manuscript more readily accessible to a broader readership. The reviewercomments of R1 are in line with the reviewercomments from Reviewer 2, who also suggests the authors make a few minor clarifications. Reviewer 3 raises no criticism. Taken together, all reviews are very positive and only minor clarifications are requested.

We greatly thank the editor for the efforts and their overall positive evaluation, and we thank the reviewers for their helpful and enthusiastic feedback.

We have revised the manuscript to address all concerns, with individual responses below. We have also generally improved the manuscript with some minor edits throughout, as well as by marking up the allotaxonographs in Figures 1 and 2 to better facilitate our explanation of their elements. We have also adjusted some terminology for clarity, and to remove unneeded attributions to individual scientists. For example, we have largely replaced "Zipf distributions" with the more plainspoken "type rankings".

We also apologize for the delay in our response.

Response to Reviewer #1:

Comment 1 by Reviewer 1

Reviewer #1: Thank you for the opportunity to review this paper. It proposes new methods to compare either multiple systems or a single system over time, including graphs and a measure of divergence. The proposed methods are demonstrated on a variety of data examples.

Major Reviewercomments:

Can we use the rank-rank histogram to detect differences in the number of component types between the two systems? For example, can we use the histogram to figure out whether there are more species in 1985 vs. 2015 in Figure 3?

As was presented in the manuscript, the total number of types was not made explicit.

☐ Point to balances

☐ Add some text to paper about this.

☐ In future iterations, we could add the number of tokens (if counts are the measure), number of types, and number of exclusive types to the inset showing balances.

☐ Can see the imbalance from the histogram

☐ The numbers can be reported in text accompanying figures

Comment 2 by Reviewer 1

Also, can we use the graph to gauge the difference in size between the component type of rank r in System 1 and the component type of System 2 of the same rank?

For RTD, we wanted to build a tool that only works with ranked data. While in practice, we will have systems where sizes are typically known, the intent here was to build a fully ranked-based instrument in the spirit of well established rank-based statistical methods (e.g., the Spearman coefficient).

In our second paper on allotaxonometry, we introduce and explore probability-turbulence divergence, where we make size (as measured by probability or rate) explicit.

We note that the allotaxonometer code does include the ability to compare systems where we have sizes that are not counts or probabilities or cannot be interpreted as relative fractions in some way (e.g., heights of buildings).

This manuscript is also in review and can be found here: https://arxiv.org/abs/2008.13078

☐ Zipf distributions comparing cities in 1750 to 2000 in the US

☐ Could compare normalized ones ...

☐ Best thing would be to overlay the two Zipf distributions. We considered ways of adding Zipf distributions to the plots but we felt that the results would be too complicated (and we acknowledge that the allotaxonographs are complicated enough as is)

Comment 3 by Reviewer 1

How do we interpret the percentages in the "Balances" section at the bottom of Figure 1? Are these percentages unrelated to your alpha tuning parameter?

The balances are independent of alpha, and indicate relative percentages of tokens, types, and exclusive types. ☐ We have added some text to make this more clear and ☐ we have also expanded the balance section to included absolute numbers.

Comment 4 by Reviewer 1

The paper states that the standard methods lack transparency and flexibility. It may be helpful to apply some standard methods in one of your case studies to demonstrate the differences from your method.

Yes, this is well spotted. We have separated out work where we compare methods (e.g., JSD vs PTD) and that is still in development, and had left residual guidance in the abstract.

We have removed two sentences from the abstract that misleadingly pointed in this direction.

Comment 5 by Reviewer 1

How sensitive is your method to small perturbations? For example, if you fixed the errors in the Berkshire Hathaway and DowDuPont market caps in the last case study, how much do the divergence and histogram change?

In this first paper, we have explored the effects of truncation only. Generally, using ranks will give a more robust instrument, and is part of our motivation.

We have now noted in the manuscript the following:

"Small errors or perturbations of counts for a Zipf distribution will generally not affect the ranks of the dominant types."

Comment 6 by Reviewer 1

What is the computing time to calculate the divergence and generate the histograms in your four case studies?

Generally, the calculations and graphics require a few seconds of compute time. The computational bottleneck in producing the RTD calculations is confined entirely to the counting of elements to produce ranks. For the Twitter data, for example, enumerating n-grams for each day (20GB of compressed text) can take many CPUs running continuously to keep up with real time, with more than one million previously unseen types encountered on a given day.

Comment 7 by Reviewer 1

Minor Reviewercomments:
The terms "Zipf" and "1-gram" should be defined.

We have made the initial reference to Zipf more clear, connecting only to "Zipf's law". We have now defined Zipf distribution and n-grams in the manuscript. We have replaced "Zipf ranking" with "size ranking" throughout, as we do not need to name this simple kind of ranking after an individual.

Added to the manuscript: 'For ease of reference, we will refer to "component rankings by decreasing size" as "size rankings."'

Definition/clarification of 1-grams now appears in a couple of places, e.g.:

'For Twitter, subsampling n-grams—phrases containing n words—allows for robust estimation of the rates of common n-grams but not rare ones.'

Comment 8 by Reviewer 1

In Paragraph 2 of Part B in the Introduction, it would be helpful to get a concrete example of a component type.

☐ Do

Comment 9 by Reviewer 1

Define the notation $r_(\tau, 1)$ and $r_(\tau, 2)$, which seems to be first used in Part B of Section II. I believe they are the ranks of component in Systems 1 and 2, but this should be clarified.

☐ Do

Comment 10 by Reviewer 1

In Figure 1, the caption has substantial overlap with the main text. Removing some of the overlap would help highlight the new information.

It is our strong preference to make the captions reasonably complete in themselves. Readers may choose to work through the paper in different ways.

Comment 11 by Reviewer 1

In the first two paragraphs on page 6, some of the words that are described (such as "Harambe" and "voted") do not seem to be in Figure 1.

Thanks for pointing this out, we have adjusted the language describing Figure 1 to match words found at the edges of the histogram. 'Gorilla' appears in the figure, so we have retained description of Harambe for context in the main text.

☐ double check that remade figure contains words referenced in the caption

Comment 12 by Reviewer 1

Could you give some intuition of why you chose $\alpha = 0$ for the BCI example? Also, what is the $D^R_{0;rand}$ term mentioned in that example?

We make the choice of $\alpha = 0$ by inspection. In the appropriate supplementary PDF flipbook, the effect of tuning alpha can be observed. Future work, which we discuss at the end of the paper, would seek to find a method to determine an optimal alpha for specific comparisons.

☐ Connect rand properly back to normalization and notation definition, or wherever

Note that we have made the first appearance of our notation for RTD into a standalone equation (the first equation, shifting following equations up).

Response to Reviewer #2:

Comment 1 by Reviewer 2

It was my pleasure to review this article introducing a new framework to compare heavy-tailed ranked lists. Specifically, the work motivates and derives a novel divergence measure for ranked lists, and then illustrates its usefulness across four examples including: Twitter word frequency, baby name popularity, species abundance, and firm sizes. The presentation is rich and clear, elucidating the central framework and offering wonderful insights into the motivating examples and applications. The problem this article addresses is relevant to a broad range of data science applications. Therefore, I strongly recommend to publish the article in EPJ Data Science.

We thank Reviewer 2 for their generous feedback.

Comment 2 by Reviewer 2

Reviewercomments: Even through the presentation is quite clear, it may benefit from an organization that extracts a few non-technical discussions from the rest of the text - i.e. a clearly identified general audience guide to allotaxonomy and the application of rank-turbulence. This could support usage in fields outside of physics / data sciences.

Thank you.

We are authoring a blog post to accompany publication of the manuscript that will advertise the utility of the instrument to other disciplines.

We are also working on julia and python versions of the allotaxonometer code (presently 6000 lines of Matlab) to encourage broader adoption.

We and others have also used allotaxonographs in a number of other papers, and we now cite these in the manuscript's conclusion.

Comment 3 by Reviewer 2

Relatedly, it would be helpful if the authors could elaborate on section IIE and the processes used to select different values of alpha for the different applications. I'm imagining the framework being used by a non-expert who just keeps whatever default value of alpha is encoded instead of exploring the full range - or on the flip side, being overwhelmed with the many potential choices of alpha and not understanding what features of the histograms are being accentuated by the divergence in different limits. Can you distill a few guiding principles or questions to help motivate an appropriate selection?

In the present work, we justify the choice of alpha by looking at the shape of distribution as a guide.

In the appropriate supplementary PDF flipbook, the effect of tuning alpha can be observed.

In future work, which we discuss at the end of the paper, we will seek to find a method to determine an optimal alpha for specific comparisons.

Comment 4 by Reviewer 2

The examples illustrating the effects of partial sampling were helpful, but it would be nice to see how the divergence measure itself changes with subsample size: does it converge or is it sensitive to the influx of different cross-ranks uncovered by larger subsample sizes? Further, how does the divergence measure work in the presence of unbalanced list sizes, say when comparing the word usage of English tweets between countries? I noticed that some imbalance likely appears in the baby names example. Does the skewness of the underlying histogram bias the divergence and resulting understanding of turbulent names?

☐ Performance under sampling: should obey whatever decay rate rank-ordered data of particular scaling law experiences, independent of this methodology.

☐ We've seen undersampled tweet data, looks too sparse near center, too many hapax words appearing in bar chart on right.

Comment 5 by Reviewer 2

The figures function as standalone illustrations of the concepts and enhance the intuition underlying the measure. I especially like Figure 2, which manages to compress an impressive amount of information into a clear visualization. Furthermore, the Supplemental Flipbooks are fantastic and I hope the publication of this article is accompanied by a Twitter campaign sharing these as gifs. I'm already excited to explore the allotaxonomy of my own research questions!

Thank you!

As we noted in the paper, we have an online site for the papers, data, and code that we have produced (and will produce) around allotaxonometry. There are various gifs on display:

https://storylab.w3.uvm.edu/allotaxonometry/

Response to Reviewer #3:

Comment 1 by Reviewer 3

Reviewer #3: This is overall an excellent piece of research. The authors develop an innovative, solid method to compare rankings in systems, and demonstrate its use for word frequencies in social media. The methods are solid, the writing of very quality and the figures as well. It is without hesitation that I recommend this paper for publication as is.

We greatly, greatly appreciate Reviewer 3's extremely positive assessment. Thank you!