August 22, 2024

—Collaboration report—

CUMULATIVE JACCARD DISTANCE FOR SEMANTIC COMPARISON

AARON HUNTLEY

Contents

1. Introduction

Let us consider a set W of n distinct words. A *numerical ranking* with *domain* W is a function from W to the real numbers $\mathbb{R}$.

For a given word g, typically not in W, one can obtain a numerical ranking $r_g \colon W \to \mathbb{R}$ by experimentally collecting similarity judgments from native speakers regarding the relationship between g and each element in W. We refer to g as the *grounding* of the numerical ranking r_g.

A compelling rationale for this approach is that numerical rankings offer a viable method for assessing the semantic proximity of groundings within a fixed domain W. This method facilitates the comparison of how closely related different groundings are based on their numerical representations. However, there are two primary challenges that complicate this endeavor.

First, traditional metrics, such as the Euclidean distance, fail to capture the nuance that words with higher numerical rankings are more indicative of the semantic content of a grounding than those with lower rankings. In semantic analyses, it is crucial to account for the prominence of high-ranking words, as they carry more descriptive weight and thus better represent the underlying meaning.

Second, the variability of domains and scales across different studies presents a significant hurdle for large-scale meta-analyses. To find a common domain,

we can simply take the intersection of all considered domains. However, this approach does not resolve the issue of differing experimental scales, which can lead to inconsistencies that complicate the synthesis and comparison of results from multiple sources. Addressing these scaling disparities is essential to ensure the robustness and reliability of meta-analytic conclusions.

To address the second issue we consider only topological distances between numerical rankings, i.e., those depending only on the order rankings induced on W by these numerical rankings. Explicitly, the *order ranking* defined by a numerical ranking r is the order in W defined by

$$w_1 \leq_r w_2 \text{ in } W \quad \text{if} \quad r(w_1) \leq r(w_2) \text{ in } \mathbb{R}.$$

Topological distances have the added advantage that they are more resilient to noise in the data, since small perturbations of a numerical ranking leave its order ranking unchanged.

To address the first issue, we consider the *cumulative Jaccard distance* (CJD), a principled topological distance for rankings of a domain W that aggregates the proportion of words shared by the two rankings among their top i ranked words for every $i \in \{1, \ldots, n\}$. In precise terms, consider two rankings r and r' of W. For $i \in \{1, \ldots, n\}$, let $A(i)$ and $A'(i)$ be the sets containing the i highest ranked words according to r and r' respectively. Then, the cumulative Jaccard distance is defined by

$$\mathrm{CJD}(r, r') = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{|A(i) \cap A'(i)|}{|A(i) \cup A'(i)|}.$$

This distance function is induced by a similarity measure termed *Cumulative Jaccard similarity* and defined by

$$\mathrm{CJS}(r, r') = \frac{1}{n} \sum_{i=1}^{n} \frac{|A(i) \cap A'(i)|}{|A(i) \cup A'(i)|}.$$

We can think of a similarity measure as inverting the values of its associated distance, so that higher similarity between two rankings indicates a closer relationship.

We can think of the cumulative Jaccard similarity between two rankings r and r' as the area under their *Jaccard curve*, defined for each $i \in \{1, \ldots, n\}$ by

$$j_{r,r'}(i) = \frac{|A(i) \cap A'(i)|}{|A(i) \cup A'(i)|}.$$

This curve can be plotted, providing a visualization tool that facilitates analysis and interpretability.

1.1. **Merge product on numerical rankings.** With this formalism we can explore ways of taking two numerical rankings to produce a third which is sufficiently similar to the first two with respect to the CJS. Since our data started as a numerical ranking our first approach to this idea was to pull-back the monoidal product on $\mathbb{R}$ given by addition (note there are two monoidal products on the real numbers: addition and multiplication). We can without loss of any information re-normalise to the interval [0,1] and so the product of addition becomes taking the average.

So for two numerical rankings $r_g, r_{g'}$ we define the average average ranking.

$$\begin{aligned} r_{g,g'} \colon W &\to \mathbb{R} \\ w &\mapsto \frac{r_g(w) + r_{g'}(w)}{2} \end{aligned}$$

From the average ranking we get the induced order ranking $r_{g,g'}$ which we can now compare to other order rankings. Note that if we were following the theory strictly we would leave the factor of $\frac{1}{2}$ out, however since in our experiments the co-domain of the data is actually an interval (e.g. $[1, 7]$) we add this factor to make the merge product comparable.

2. Experiments

We use the cumulative Jaccard similarity (CJS) to study groundings in an aggregation of experimental open-source data. Specifically, we consider a collection of 8023 words, W, and 20 groundings denoted as g_i, where $i \in \{0, \dots, 19\}$. Our findings indicate that the CJS aligns with the intuitive understanding of the data more effectively than the Pearson correlation.

We now describe the experiments we performed:

(1) **Ranking space:** We conducted a statistical analysis of the CJS on the space of order rankings, by randomly sampling orderings of the 8023 words, we employed kernel density estimation to approximate the underlying Gaussian distribution of cumulative Jaccard similarities.
(2) **Heatmaps:** We computed the CJS between each pair of groundings in our dataset. Figure 2.3 displays the resulting heatmap of similarities. We then compared this heatmap with a Pearson correlation heatmap to evaluate differences and similarities in the patterns detected by both metrics.
(3) **Jaccard curves:** The cumulative Jaccard similarity between two groundings can be interpreted as the area under a curve. We present the curves associated two the extreme values of the cumulative Jaccard similarity.
(4) **Merge products:** We experiment with the idea of merging two of our groundings, Valance and Arousal to create a third higher order grounding, Emotion which we can now compare to our existing groundings.

These experiments collectively provide a comprehensive understanding of how the Jaccard similarity metric can be applied to the analysis of groundings and order rankings in large datasets.

2.1. Data.

The data we used was the same as discussed in [Div+23].

We normalised the raw data by restricting our attention to the available groundings which were rated on more than 8000 words. We further removed groundings which were double-counted due to different studies (e.g., `valence.Mohammad` was removed but `valence.Warriner` was kept). We assigned each grounding a number so we can refer back to each one. There were 20 remaining groundings:

After removing the other groundings, we also removed all the words which were not rated against all of these groundings, of which there are 8023. This way, we have comparable groundings.

Lancaster	Warriner	Other
Auditory 5	Valence 16	Socialness 0
Gustatory 6	Arousal 17	LgSUBTLWF 1
Haptic 7	Dominance 18	OLD 2
Interoceptive 8		PLD 3
Olfactory 9		Concreteness 4
Visual 10		Val_ext 19
Foot_leg 11		
Hand_arm 12		
Head 13		
Mouth 14		
Torso 15		

TABLE 1. Groundings organized by data origin.

2.2. **Expected values.** Here we look closer into the statistical distribution of CJS in the space of order rankings of 8023 words. We do this to try to calibrate significant values for the selection. Our first approach was to take a two randomly ordered lists of 8023 words and compute the Jaccard similarity, we do this 100 times and plot the results in a histogram. We then use Kernel density estimation to estimate the probability density function of the CJS's. Figure 1 presents the histogram and estimated PDF of the distribution of CJS's. We have the mean of this distribution is 0.386 and so the CJS value for no similarity or dissimilarity should be around this value.

There are two problems here are the assumption of Gaussian distribution and the small selection of data. To solve these problems we looked at the same distributions but instead on a smaller word set. Figure 2 shows the histogram of the CJS of every permutation with the identity for permutations of length $n = 8, 9$.

From this we conclude that the distribution of CJS's is not Gaussian and there are some other theoretical constraints on the bounds which we briefly discuss in section A.

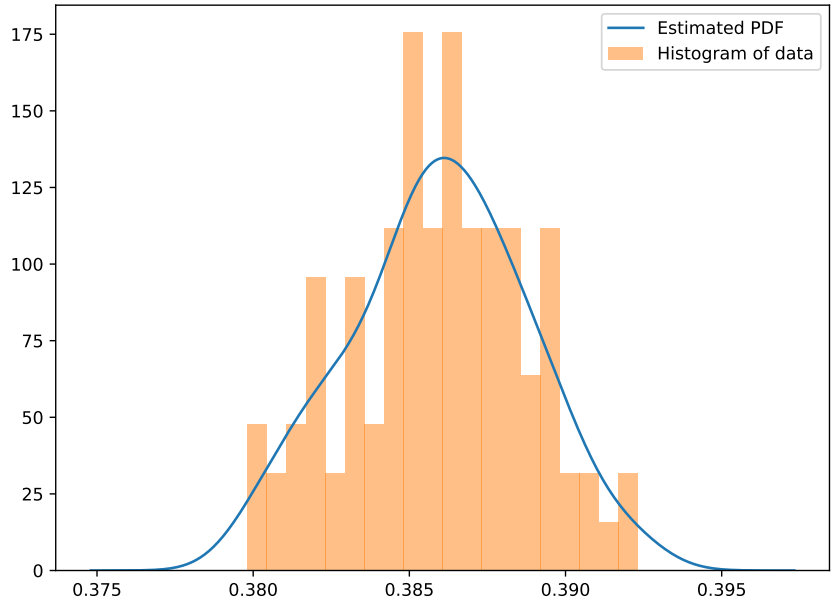

FIGURE 1. Kernel density estimation of 100 random CJS's on 8023 words

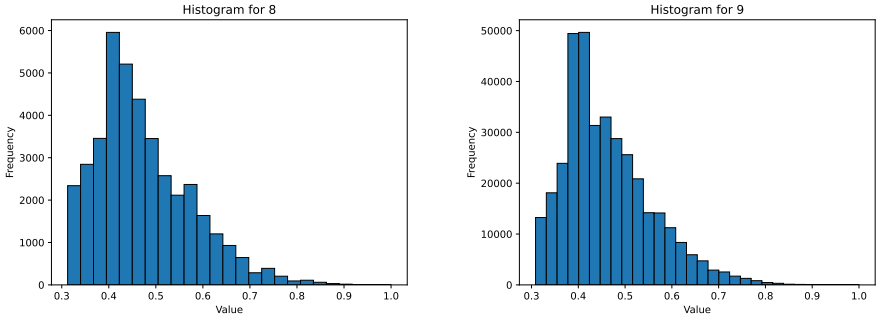

FIGURE 2. Histograms of CJS between every permutation in $\mathbb{S}_n$ and an identity for $n = 8$ (left) and $n = 9$ (right).

2.3. **CJS and Pearson correlation.** First, we created the heat map of the CJS between every pair of groundings. We computed this heat map in python where we wrote a function to which calculates the cumulative Jaccard similarity between every two groundings and collects the results in a matrix. We compare this to the heat map of the Pearson correlation, which is a well studied method to compute the correlation between two random variables.

Note here that the heat map for the CJS is displayed as ranging from values 0 to 1 however there are theoretical constraints meaning the CJS does not actually go below around 0.2. Furthermore, as we discussed in the previous section the

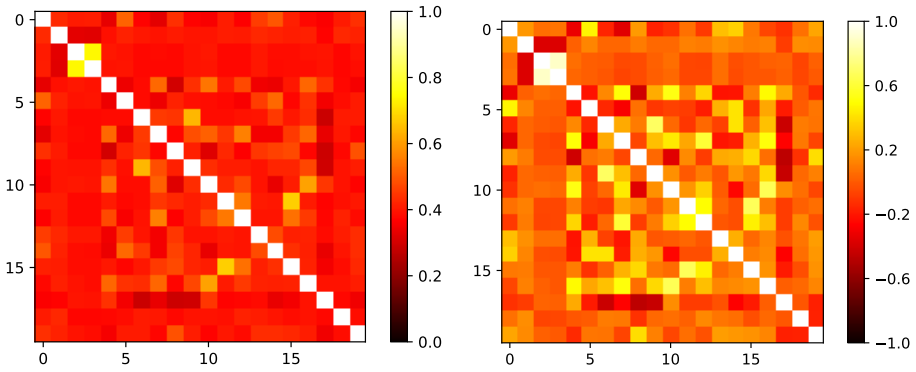

FIGURE 3. Heatmaps for CJS (left) and Pearson correlation (right)

mean value for the CJS is 0.386 not 0.5. By a quick observation we can see some similarities and differences, for example the bright white entry in the top left of both heat maps represents the comparison between `OLD` and `PLD` which is what we would expect. Despite this difference we see these heat maps do have a somewhat similar "shape" which we will now explore more precisely.

2.3.1. *Pearson correlation of matrices.* Here we compute the Pearson correlation between the two matrices. Before we can apply this method we have to linearly normalise the scale of the two matrices, here we changed the scale of the correlation matrix to be from 0 to 1. We then vectorized the matrices and computed the Pearson correlation coefficient. The result was a correlation coefficient of 0.915680136098452.

This shows the method of Jaccard similarity is highly correlated with the method of Pearson Correlation.

2.3.2. *Frobenius norm of matrices.* The Frobenius norm is a way to measure the difference between two matrices by treating them as vectors and computing the Euclidean distance between these vectors. It's particularly useful for comparing matrices, including correlation matrices. Similarly to before we linearly normalised the Pearson correlation matrix and then computed the Frobenius norm getting a value of 2.607239338452729. Since the size of the matrices are 20×20 with entries in the range $[0, 1]$ this is a relatively low Frobenius norm suggesting a high similarity between Pearson correlation and Jaccard similarity.

2.3.3. *Jaccard similarity of matrices.* To compute the Jaccard similarity between these two matrices we again vectorize them creating two vectors of size 400. We can then give each position of the vector a label and order the labels based on the values in the vector. Then we can compute the cumulative Jaccard similarity which gives us a number between 0 and 1. After running this we got the Jaccard similarity was: 0.9005976978218874.

Therefore we can conclude that CJS gives a robust way to measure correlation between numerical rankings.

2.4. **Extreme values.** From the Jaccard analysis we deem pairs of groundings to be significantly similar (dissimilar) if the have a CJS value greater than 0.525 (less than 0.3). In the Pearson Correlation analysis we deem the pairs of groundings to be significantly positively (negatively) correlated if the correlation coefficient is greater than 0.5 (less than -0.5). We include a table of the 10 most similar/dissimilar pairs for both the CJS and Pearson correlation, the significant pairs appear in color:

Cumulative Jaccard Similarity	Pearson Correlation
OLD, PLD (0.736)	OLD, PLD (0.890)
Foot_leg, Torso (0.674)	Gustatory, Olfactory (0.684)
Gustatory, Olfactory (0.638)	Foot_leg, Torso (0.670)
Visual, Valence (0.607)	Visual, Valence (0.663)
Haptic, Hand_arm (0.556)	Haptic, Hand_arm (0.598)
Foot_leg, Hand_arm (0.546)	Concreteness, Visual(0.529)
Concreteness, Visual (0.543)	Haptic, Valence (0.511)
Hand_arm, Torso (0.532)	Concreteness, Haptic (0.507)
Auditory, Mouth (0.522)	Hand_arm, Foot_leg (0.491)
Haptic, Concreteness (0.521)	Valence, Concreteness (0.479)

TABLE 2. Comparison of significantly similar/correlated groundings

Cumulative Jaccard Similarity	Pearson Correlation
Gustatory, Arousal (0.287)	Interoceptive, Arousal (−0.463)
Interoceptive, Arousal (0.289)	Olfactory, Arousal (−0.431)
Olfactory, Arousal (0.293)	Gustatory, Arousal (−0.429)
Concreteness, Interoceptive (0.301)	Concreteness, Interoceptive (−0.414)
Interoceptive, Visual (0.317)	LgSUBTLWF, OLD (−0.370)
LgSUBTLWF, PLD (0.324)	LgSUBTLWF, PLD (−0.368)
LgSUBTLWF, OLD (0.324)	Interoceptive, Visual (−0.355)
Haptic, Socialness (0.325)	Haptic, Socialness (−0.353)
Socialness, Concreteness (0.330)	Socialness, Concreteness (−0.324)
Haptic, Arousal (0.334)	Haptic, Auditory (−0.283)

TABLE 3. Comparison of significantly dissimilar/uncorrelated groundings

2.4.1. *Jaccard curves.* So far we have only looked at the cumulative Jaccard similarity, this is a useful tool as it is a mathematical similarity which naturally weights based on the highest rated words. However, we also have the Jaccard curves which we can use to further analyse specific pairings of groundings and deduce in what sense they are similar. Firstly we will look at the Jaccard curves for the three pairs of groundings with the highest CJS, Figure 4.

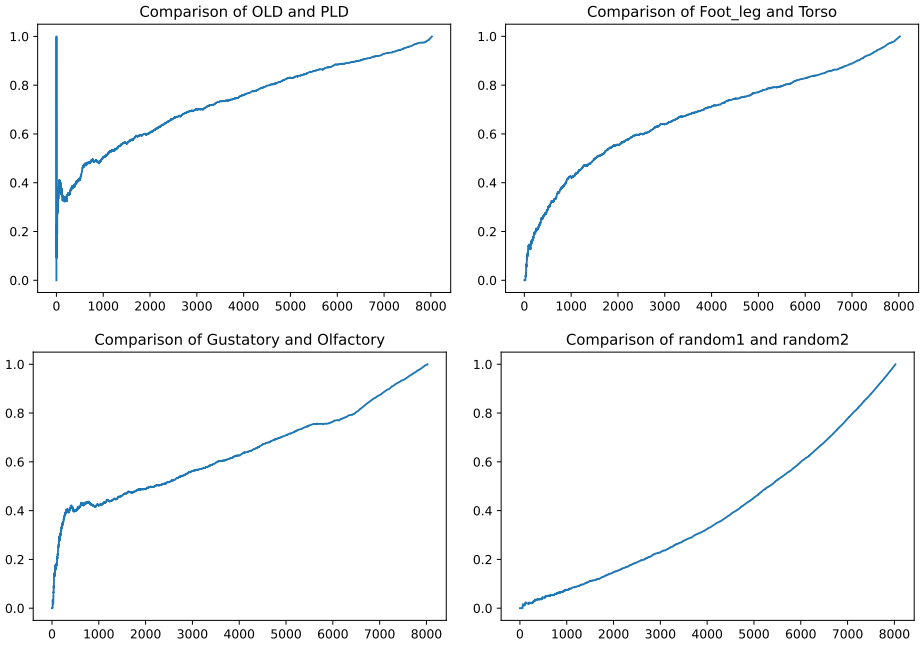

FIGURE 4. Jaccard similarity curves for the three pairs of groundings with the highest CJS and a random control.

We have also included the Jaccard curve for two random orderings on 8023 words, we will discuss this more in the next section but it can be thought of as a control. These visualizations tell us a lot more about the comparisons of the data for example, `OLD` and `PLD` have most of their similarity nearest the highest rated words they are exactly the same. There are other invariants we may be able to discuss from these curves such as the smoothness or point of inflection. For example if we look at the comparison of `Foot_leg` and `Torso` we see they are quite smoothly related but unlike `OLD` and `PLD` they are dissimilar near the start suggesting the highly rated words appear different between these two groundings but overall they have a similar order of words. Other conclusions may be drawn here but we only discuss these few examples.

Here we will also include the Jaccard curves of the three pairs of groundings with the lowest CJS, Figure 5

We can look at the inflection point of the curve to see in what way two groundings may be dissimilar.

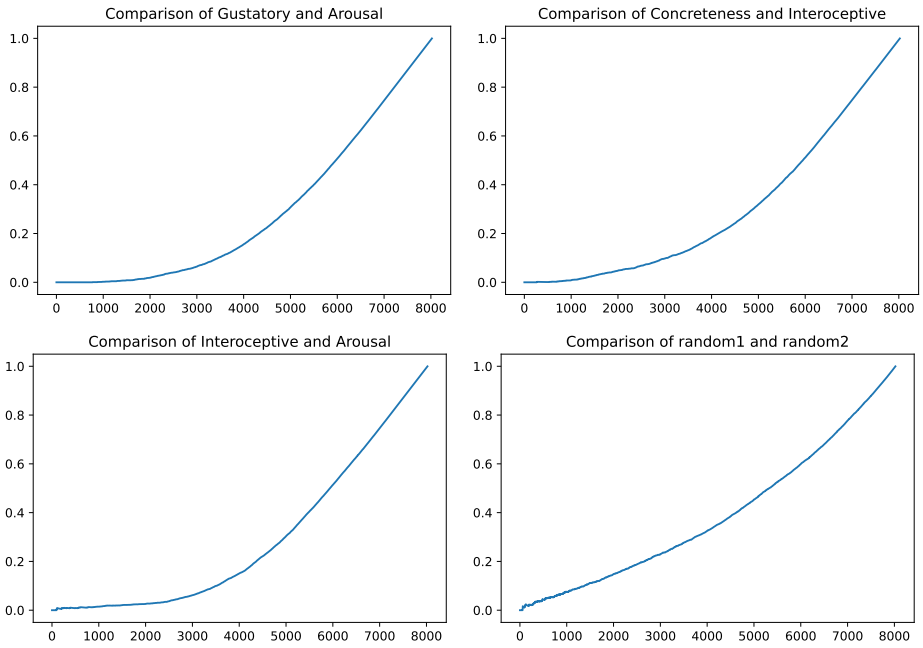

FIGURE 5. Jaccard similarity curves for the 3 groundings with the lowest CJS and a random control.

2.5. **Merge product.** Here we experiment with the new merge product on our data set. We merge arousal and valence to get a new order ranking we call Emotion. We compare emotion with Arousal and Valence.

Figure 6 shows the Jaccard curves of interest we get a CJS score of 0.595 for Valence against Emotion and 0.625 for Arousal against Emotion. As we can see Valence and Arousal are not strongly similar however we get that they are both strongly similar to their merge product Emotion. So we conclude that after merging we get a proportional amount of similarity with the two original groundings. This gives us evidence that the merge product makes sense with respect to CJS.

To explore the idea that Interoception is a higher order grounding made from Emotion, Bodily sensation and cognition we would need the extra data on cognition and direction of what bodily sensation is 'made from' from the other groundings. Here we included the Jaccard curve of Interoceptive and Emotion and found they are not significantly similar so more investigation would need to be done.

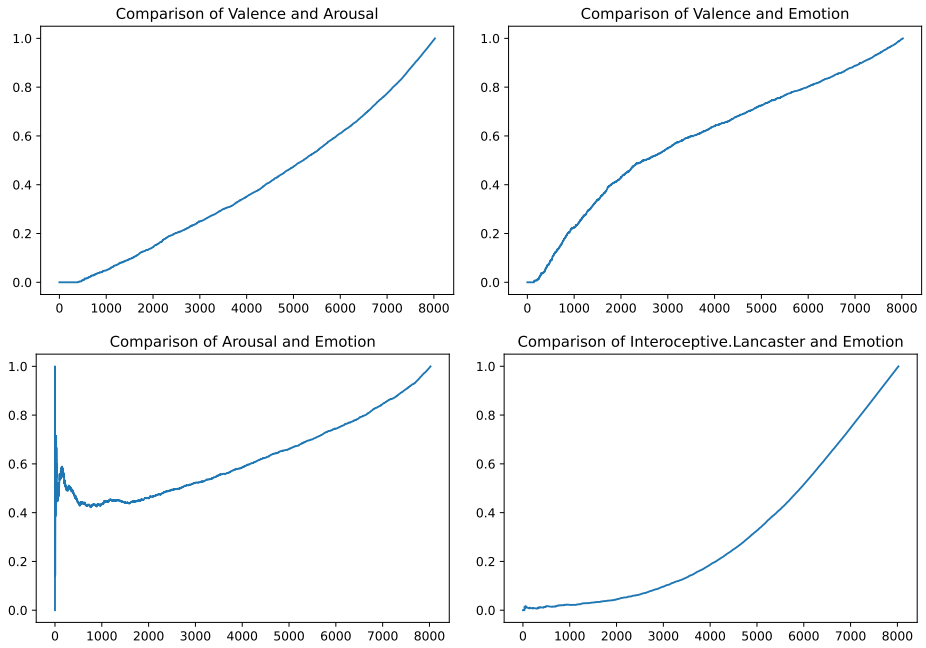

FIGURE 6. Jaccard similarity curves for Emotion against Valence, Arousal and Interoception. We also include the curve of Arousal and Valence

2.6. **Conclusion.** The Cumulative Jaccard Similarity (CJS) coefficient provides a robust, noise-resistant measure to test the similarity (correlation) between two numerical rankings which naturally weights by the higher ranked words. We have noise resistance naturally from forgetting the numerical rankings and passing to the order rankings. This is particularly interesting because, with the metric space formalism for Jaccard similarity, we now have new tools to further analyze the data such as merge products. The CJS coefficient is naturally weighted towards the highest-rated values, making this method more applicable to situations involving numerical rankings that prioritize the highest-rated items. In our experiments, words are ranked based on their perceived similarity to the reference grounding, meaning this analysis is directly applicable to our situation. Furthermore, the Jaccard curves add an extra layer of interpretability to the analysis. We also have an extra visualisation of the data given by the Jaccard curves where we can look more precisely at correlated data sets.

APPENDIX A. MATHEMATICAL FORMALISM

As in the introduction 1 we claim the cumulative Jaccard distance (CJD) defines a metric on the set of here we prove this along with other theoretical constraints on this metric.

Proposition 1. *Let $\mathfrak{X}$ be the collection of all finite sets. Then the function,*

$$\begin{aligned} j\colon \mathfrak{X}\times\mathfrak{X} &\to \mathbb{R} \\ (A,B) &\mapsto 1-\frac{|A\cup B|}{|A\cap B|} \end{aligned}$$

defines a metric on $\mathfrak{X}$.

Proof. □

Theorem 2. *Let W be a set and $\mathfrak{W}$ be the collection of all order rankings on W. Then the function,*

$$\begin{aligned} CJD\colon \mathfrak{W}\times\mathfrak{W} &\to \mathbb{R} \\ (r,r') &\mapsto \frac{1}{n}\sum_{i=0}^{n} j(A(i),B(i)) \end{aligned}$$

where $A(i)$ and $B(i)$ are the subsets of W with the words with ranking greater than or equal to i with respect to r and r', defines a metric on $\mathfrak{X}_n$. Note this aligns with our definition in the introduction by factoring out the 1.

Proof. For order rankings $r \neq r' \in \mathfrak{W}$, clearly we have $cj(r,r') \geq 0$ since it is the sum of non negative real numbers. We have,

$$\begin{aligned} CJD(r,r') &= \frac{1}{n}\sum_{i=1}^{n} j(A_i,A_i) \\ &= \frac{1}{n}0 \\ &= 0. \end{aligned}$$

We have $CJD(r,r') > 0$ since for some i we will have $A(i) \neq B(i)$ and so $j(A(i),B(i)) \neq 0$. Symmetry follows from j being symmetric. For the triangle equality observe for r, r', r'',

$$\begin{aligned} CJD(r,r') &= \frac{1}{n}\sum_{i=0}^{n} j(A(i),B(i)) \\ &\leq \frac{1}{n}\sum_{i=0}^{n} (j(A(i),C(i)) + j(C(i),B(i))) \\ &= \frac{1}{n}\sum_{i=0}^{n} j(A(i),C(i)) + \frac{1}{n}\sum_{i=0}^{n} j(C(i),B(i)) \\ &= CJD(r,r'') + CJD(r'',r') \end{aligned}$$

□

A.1. **As a permutation distance.** In $\mathbb{S}_n$, denote by σ_{ij} the transposition of elements i and j and by e the identity. We simplify notation and write $\mathrm{CJD}(\sigma_{ij})$ instead of $\mathrm{CJD}(\sigma_{ij}, e)$.

Lemma 3. *If $i < j$ then*

$$\mathrm{CJD}(\sigma_{ij}) = \frac{2}{i+1} + \frac{2}{i+2} + \cdots + \frac{2}{j}.$$

In particular, if $j = i+1$ then $\mathrm{CJD}(\sigma_i) = \frac{2}{i+1}$.

Using a Riemann sum estimation of the integral of the function $\frac{2}{x}$ we get the following estimate.

Corollary 4. *If $i < j$ then*

$$\mathrm{CJD}(\sigma_{i,j}) < \log\left(\frac{j-1}{i}\right)^2.$$

References

[Div+23] Veronica Diveica et al. "Mapping Semantic Space: Exploring the Higher-order Structure of Word Meaning". *PsyArXiv* (Dec. 2023) (cit. on p. 3).

Department of Mathematics, Western University, Canada.
Email address: ahuntle@uwo.ca