Taking uncertainty in word embedding bias estimation seriously: a bayesian approach

Alicja Dobrzeniecka and Rafal Urbaniak

1. Introduction

Natural language processing (NLP) models are used to perform various language-related tasks such as providing email filters, intelligent assistants, search results, translations, text analysis and so on. Such tools require input words to be represented by vectors of real numbers. This can usually be done with word embeddings. A common way to construct an embedding is to use a large natural language corpus to train a neural network to assign vectors to words that optimize for co-occurrence prediction accuracy. The vectors can then be compared in terms of their similarity – the usual measure is cosine similarity – and the results of such comparisons can be used in downstream tasks. Roughly speaking, cosine similarity is an imperfect mathematical proxy for semantic similarity [14].

It has been suggested [1,4,5,8,11,12] that such models can learn implicit biases that reflect harmful stereotypical thinking—for example, the (vector corresponding to the) word *she* might be much closer in the vector space to the word *cooking* than the word *he*. Such phenomena may be undesirable at least in some downstream tasks, such as web search, recommendations, and so on. To investigate such issues, several measures of bias in word embeddings have been formulated and applied. Our goal is to use two prominent examples of such measures to argue that this approach oversimplifies the situation and to develop a Bayesian alternative.

One response to the raising of the issue of bias in natural language models might be to say that there is not much point to reflecting on such biases, as they are unavoidable. This unavoidability might seem in line with the arguments to the effect that learning algorithms are always value-laden [9]: they employ inductive methods that require design-, data-, or risk-related decisions that have to be guided by extra-algorithmic considerations. Such choices necessarily involve value judgments and have to do, for instance, with what simplifications or risks one finds acceptable. Admittedly, algorithmic decision making cannot fulfill the value-free ideal, but this only means that even more attention needs to be paid to the values underlying different techniques and decisions, and to the values being pursued in a particular use of an algorithm.

Another response might be to insist that there is no bias introduced by the use of machine learning methods here, since the algorithm is simply learning to correctly

predict co-occurrences based on what "reality" looks like. However, this objection overlooks the fact that we, humans, are the ones who construct this linguistic reality, is shaped in part by the natural language processing tools we use on a massive scale. Sure, if there is unfairness and our goal is to diagnose it, we should do complete justice to learning it in the model used in a study thereof. One example of this approach is [3]. However, if our goal is to develop downstream tools that perform tasks that we care about without further perpetuating or exacerbating harmful stereotypes, we still have good reasons to try to minimize the negative impact, at least as long as this does not damage the overall performance of the tool. Moreover, it is often not the case that the corpora mirror reality—to give a trivial example, heads are spoken of more often than kidneys, but this does not mean that kidneys occur much less often in reality than heads. To give a more relevant example, the disproportionate association of female words with female occupations in a corpus actually greatly exaggerates the actual lower disproportion in the real distribution of occupations. [6].

In what follows, we focus on two popular measures of bias applicable to many existing word embeddings (*GoogleNews*,[1] *Glove*[2] and *Reddit Corpus*[3]): *Word Embedding Association Test* (WEAT) [8], and *Mean Average Cosine Distance* (MAC) [12]. We first explain how these measures are supposed to work. Then we argue that they are problematic for various reasons, a key one being that by pre-averaging data they manufacture false confidence, which we illustrate in terms of simulations showing that the measures often suggest the existence of bias even if by design it is non-existent in a simulated dataset.

We propose to replace them with a Bayesian data analysis (hierarchical models), which not only provides more modest and realistic assessment of the uncertainty involved, but also allows for inspection at various levels of granularity. Once we introduce the method, we apply it to multiple word embeddings and results of supposed debiasing, putting forward some general observations that are not exactly in line with the usual picture painted in terms of WEAT or MAC (and the problem generalizes to any approach that focuses on chasing a single numeric metric): the word list sizes and sample sizes used in the studies are usually small,[4] posterior density intervals are fairly wide, often the differences between associated, different and neutral human predicates, are not very impressive. Also, a preliminary inspection suggests that the desirability of changes obtained by the usual debiasing methods is debatable. The tools that we have proposed, however, allow for a more fine-grained and multi-level evaluation of bias and debiasing in language models without abandoning modesty about the uncertainties involved.

[1] GoogleNews-vectors-negative300, available at https://github.com/mmihaltz/word2vec-GoogleNews-vectors.

[2] Available at https://nlp.stanford.edu/projects/glove/.

[3] The Reddit-L2 corpus, available at http://cl.haifa.ac.il/projects/L2/.

[4] Depending on a list for [@Caliskan2017semanticsBiases] the range for protected words is between 13 and 100, and for attributes between 16 and 25; for [@Manzini2019blackToCriminal] the range for protected words is between 14 and 18, and for attributes between 11 and 25

Conceptually the idea is that if a particular harmful stereotype is learned in a particular embedding, then certain groups of words will be systematically closer to (or further from) each other. This gives rise to the idea of protected groups—for example, in guiding online search completion or recommendation, female words might require protection in that they should not be systematically closer to stereotypically female job names, and male words require protection in that they should not be systematically closer to toxic masculinity stereotypes, such as "tough", "never complaining" or "macho".[5]

The key role in the measures to be discussed is played by the notion of cosine distance (or, symmetrically, by cosine similarity). These are defined as follows:[6]

$$\mathsf{cosineSimilarity}(A,B) = \frac{A \cdot B}{\|A\|\,\|B\|} \tag{Sim}$$

$$\mathsf{cosineDistance}(A,B) = 1 - \mathsf{cosineSimilarity}(A,B). \tag{Distance}$$

Note that this terminology is slightly misleading, as mathematically cosine distance is not a distance measure, because it does not satisfy the triangle inequality, as generally $\mathsf{cosineDistance}(A,C) \not\leq \mathsf{cosineDistance}(A,B) + \mathsf{cosineDistance}(B,C)$. We will keep using this mainstream terminology.

One of the first measures of bias has been developed in [1]. The general idea is that a certain topic is associated with a vector of real numbers (the topic "direction"), and the bias of a word is investigated by considering the projection of its corresponding vector on this direction. For instance, in [1], the gender direction gd is obtained by taking the differences of the vectors corresponding to ten different gendered pairs (such as $\overrightarrow{she} - \overrightarrow{he}$ or $\overrightarrow{girl} - \overrightarrow{boy}$), and then identifying their principal component.[7] The gender bias of a word w is then understood as w's projection on the gender direction: $\vec{w} \cdot gd$ (which, after normalizing by dividing by $\|w\|\,\|\mathsf{gd}\|$, is the same as cosine similarity). Given a list N of supposedly gender neutral words,[8] and the gender direction gd, the direct gender bias is defined as the average cosine similarity of the words in N from gd (c is a parameter determining how strict we want to be):

$$\mathsf{directBias}_{\mathsf{c}}(\mathsf{N},\mathsf{gd}) = \frac{\sum_{w \in N} |\mathsf{cos}(\vec{w},\mathsf{gd})|^c}{|N|}$$

The use of projections in bias estimation has been criticized for instance in [5], where it is pointed out that while a higher average similarity to the gender direction might be an indicator of bias with respect to a given class of words, it is only one

[5] However, for some research-related purposes, such as the study of stereotypes across history [3], embeddings which do not protect certain classes may also be useful.

[6] Here, "$-$" stands for point-wise difference, "$\cdot$" stands for the dot product operation, and $\|a\| = \sqrt{(a \cdot a)}$.

[7] Roughly, the principal component is the vector obtained by projecting the data points on their linear combination in a way that maximizes the variance of the projections.

[8] We follow the methodology used in the debate in assuming that there is a class of words identified as more or less neutral, such as *ballpark, eat, walk, sleep, table*, whose average similarity to the gender direction (or other protected words) is around 0. We will have more to say about this assumption when we describe our dataset construction.

possible manifestation of it, and reducing the cosine similarity to such a projection may not be sufficient to eliminate bias. For instance, "math" and "delicate" might be equally similar to a pair of opposed explicitly gendered words (*she*, *he*), while being closer to quite different stereotypical attribute words (such as *scientific* or *caring*). Further, it is observed in [5] that most word pairs retain similarity under debiasing meant to minimize projection-based bias.[9]

A measure of bias in word embeddings which does not employ similarities to pre-calculated protected-class-level directions, the Word Embedding Association Test (WEAT), has been proposed in [8]. The idea here is that the bias between two sets of target words, X and Y (we call them protected words), should be quantified in terms of the cosine similarity between the protected words and attribute words coming from two sets of stereotype attribute words, A and B (we will call them attributes). For instance, X might be a set of male names, Y a set of female names, A might contain stereotypically male-related, and B stereotypically female-related career words. The association difference for a particular word t (belonging to either X or Y) is:

$$\mathsf{s}(t,A,B) = \frac{\sum_{a\in A}\cos(t,a)}{|A|} - \frac{\sum_{b\in B}\cos(t,b)}{|B|} \tag{1}$$

then, the association difference between A a B is:

$$\mathsf{s}(X,Y,A,B) = \sum_{x\in X}\mathsf{s}(x,A,B) - \sum_{y\in Y}\mathsf{s}(y,A,B) \tag{2}$$

The authors use it as a test statistic in some tests, and the assumption that X and Y are of the same size is necessary for the statistic to make sense. In the final measure (of effect size), WEAT, means are taken, which makes it insensitive to size differences between X and Y:

$$\mathsf{weat}(A,B) = \frac{\mu(\{\mathsf{s}(x,A,B)\}_{x\in X}) - \mu(\{\mathsf{s}(y,A,B)\}_{y\in Y})}{\sigma(\{\mathsf{s}(w,A,B)\}_{w\in X\cup Y})} \tag{3}$$

WEAT is inspired by the Implicit Association Test (IAT) [17] used in psychology, and it uses almost the same word sets, allowing for a *prima facie* sensible comparison with bias in humans. In [8] the authors argue that significant biases—thus measured—similar to the ones discovered by IAT can be discovered in word embeddings. In [11] the methodology is extended to a multilingual and cross-lingual setting, arguing that using Euclidean distance instead of cosine similarity does not make much difference, while the bias effects vary greatly across embedding models (interestingly, with social media-text trained embeddings being less biased than those based on Wikipedia). A similar methodology is employed in [4]. The authors employ word embeddings trained on corpora from different decades to study the shifts in various biases.[10]

[9] In [1] another method which involves analogies and their evaluations by human users on Mechanical Turk is also used. We do not discuss this method in this paper, see its criticism in [16].

[10] Strictly speaking, these authors use Euclidean distances and their differences, but the way they take averages and averages thereof is analogous, and so what we will have to say about pre-averaging leading to false confidence applies to this methodology as well.

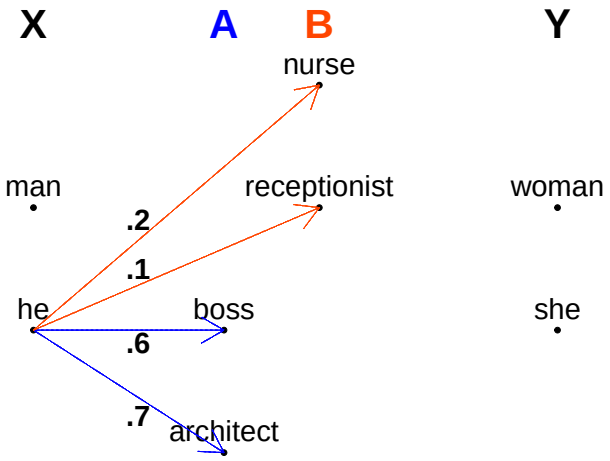


Figure 1: Example WEAT calculations: $s_1 = s(he,A,B) = \frac{.6+.7}{2} - \frac{.2+.1}{2} = .65 - .15 = .5$, $s_2 = s(man,A,B) = .3$, $s_3 = s(woman,A,B) = -.6, s_4 = s(she,A,B) = -.3$, $\mathsf{WEAT}(A,B) = \frac{(s_1+s_2)/2-(s_3+s_4)/2}{sd(\{s_1,s_2,s_3,s_4\})} \approx 1.93$.

- $s_1 = s(he,A,B) = \frac{.6+.7}{2} - \frac{.2+.1}{2} = .65 - .15 = .5$
- $s_2 = s(man,A,B) = .3$, $s_3 = s(woman,A,B) = -.6, s_4 = s(she,A,B) = -.3$

$$\mathsf{WEAT}(A,B) = \frac{(s_1+s_2)/2 - (s_3+s_4)/2}{sd(\{s_1,s_2,s_3,s_4\})} \approx 1.93$$

WEAT, however, has been developed to investigate biases corresponding to a pair of supposedly opposing stereotypes, and so the question arises as to how to generalize the measure to contexts in which biases with respect to more than two stereotypical groups are to be measured. Such a generalization can be found in [12]. The authors introduce Mean Average Cosine distance (MAC) as a measure of bias. Let $T = \{t_1, \ldots, t_k\}$ be a set of protected words, and let each $A_j \in \mathscr{A}$ be a set of attributes stereotypically associated with a protected word where $\mathscr{A}$. For instance, when biases related to religion are to be investigated, they use a dataset of the format illustrated in Table 1. The measure is defined as follows:

$$\mathsf{s}(t,A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \mathsf{cosineDistance}(t,a)$$

$$\mathsf{mac}(T,\mathscr{A}) = \frac{1}{|T|\,|\mathscr{A}|} \sum_{t \in T} \sum_{A_j \in \mathscr{A}} \mathsf{s}(t,A_j)$$

That is, for each protected word $t \in T$, and each attribute set A_j, they first take the mean of distances for this protected word and all attributes in a given attribute class, and then take the mean of thus obtained means for all the protected words and all the protected classes.[11]

[11]The authors' code are in their github repository, https://github.com/TManzini/DebiasMulticlassWordEmbedding.

protected words (T)	attributes	attribute set (A_j)	cosine distance
rabbi	greedy	jewStereotype	1.0306175
church	familial	christianStereotype	0.7087424
synagogue	liberal	jewStereotype	0.7922607
jew	familial	christianStereotype	0.9783245
quran	dirty	muslimStereotype	1.1207093
muslim	uneducated	muslimStereotype	0.5160429
torah	terrorist	muslimStereotype	0.9341137
quran	hairy	jewStereotype	1.1764642
synagogue	violent	muslimStereotype	0.9549743
bible	cheap	jewStereotype	1.2234364
christianity	greedy	jewStereotype	0.9728545
muslim	hairy	jewStereotype	0.8788219
islam	critical	christianStereotype	0.7880706
muslim	conservative	christianStereotype	0.4453191
mosque	greedy	jewStereotype	1.1541524

Table 1: Sample 15 rows of the religion dataset. The whole dataset has 15 unique protected words (T), and 11 unique attributes divided between 3 attribute sets ($A_1 = \mathsf{jewStereotype}, A_2 = \mathsf{christianStereotype}, A_3 = \mathsf{muslimStereotype}$). $\mathscr{A}$ consists of these three sets, $\mathscr{A} = \{A_1, A_2, A_3\}$. The whole dataset has $15 \times 11 = 165$ rows.

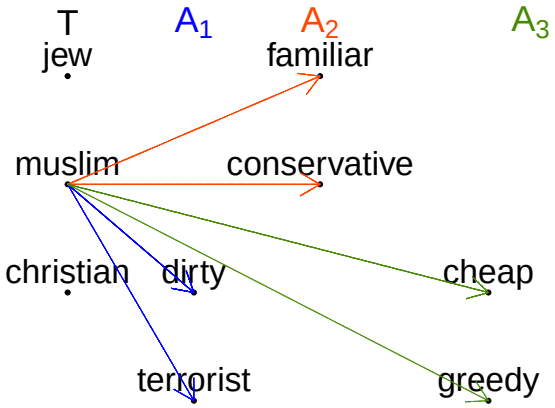


Figure 2: Example of MAC calculations. $s_1 = s(muslim, A_1) = \frac{cos(muslim,dirty)+cos(muslim,terrorist)}{2}$, $s_2 = s(muslim, A_2) = \frac{cos(muslim,familiar)+cos(muslim,conservative)}{2}$, ... Then $\mathsf{MAC}(T,A) = \mathsf{mean}(\{s_i | i \in 1,\ldots,k\})$.

Having introduced the measures, first, we will introduce a selection of general problems with this approach, and then we will move on to more specific but important problems related to the fact that the measures take averages and averages of averages. Once this is done, we move to the development of our Bayesian alternative and the presentation of its deployment.

Here is an example of results of debiasing that can be found in [12] (Table 2). It presents the results of applying the procedure with respect to the Religion dataset. MAC scores are given for the original datasets, and two resulting from two debiasing methods applied to the Reddit embedding. The scores for debiasing come with p-values for the comparison to the original MAC.

Religion Debiasing	MAC	a p-value
Biased	0.859	N/A
Hard Debiased	0.934	3.006e-07
Soft Debiased ($\lambda = 0.2$)	0.894	0.007

Table 2: The associated mean average cosine similarity (MAC) and p-values for debiasing methods for religious bias.

The first conceptual question we should ask is whether the initial MAC values lower than 1 indeed are indicative of the presence of bias? Of course, thinking abstractly, 1 is the ideal distance for unrelated words. But clearly, there is variation in distances, which might lead to non-biased lists also having MAC smaller than 1. How much smaller? What may attract attention is the fact that the value of cosine distance in "Biased" category is already quite high (i.e. close to 1) even before debiasing. High cosine distance indicates low cosine similarity between values. One could think that average cosine similarity equal to approximately 0.141 is not large enough to consider it as bias. The authors, though, still aim to mitigate such "bias" to make the distance even larger. The question is, on what basis is this small similarity still considered as a proof of the presence of bias, and whether these small changes are meaningful.

The problem is that the original paper did not employ any control group of neutral attributes for comparison to obtain a more realistic gauge on how to understand MAC values. Later on, in our approach, we introduce such control word lists. One of them is a list of words we intuitively considered neutral (see the Appendix for word lists).

Moreover, it might be the case that words that have to do with human activities in general, even if unbiased, are systematically closer to the protected words than merely neutral words. This, again, casts doubt on whether comparing MAC to the abstractly ideal value of 1 is a methodologically sound idea. For this reason we also use a second list with intuitively non-stereotypical human attributes (again, see the Appendix for a word list).

Another important observation is that MAC calculations do not distinguish whether a given attribute is associated with a given protected word, simply averaging across all such groups. Let us use the case of religion-related stereotypes to illustrate. The full lists from [12] can be found in Appendix A.3.1. In the original paper, words from all three religions were compared against all of the stereotypes. No distinction between cases in which the stereotype is associated with a given religion, as opposed to the situation in which it is associated with another one, is made. For example, the protected word *jew* is supposed to be stereotypically connected with the attribute *greedy*, while from the protected word *quran* the attribute *greedy* comes from a different stereotype, and yet the distances between these pairs contribute equally to the final MAC score.

This is problematic, as not all of the stereotypical words have to be considered as harmful for all of the religions. To avoid the masking effect, one should pay attention to how protected words and attributes are paired by stereotypes.

In Figure 3 we look at the empirical distributions, while paying attention to such divisions. The horizontal lines represent the values of 1-MAC (that is, we now talk in terms of cosine similarity rather than cosine distance; this is a trivial linear transformation) that the authors considered indicative of bias for stereotypes corresponding to given word lists. For instance, in religion, MAC was .859 and considered a sign of bias, so we plot $0 \pm (1 - .859) \approx .14$ lines around similarity = 0 (that is, distance = 1). Notice that most distributions are quite wide, and the proportions of even neutral or human neutral words with similarities higher than the value of 1-MAC deserving debiasing according to the authors is quite high.

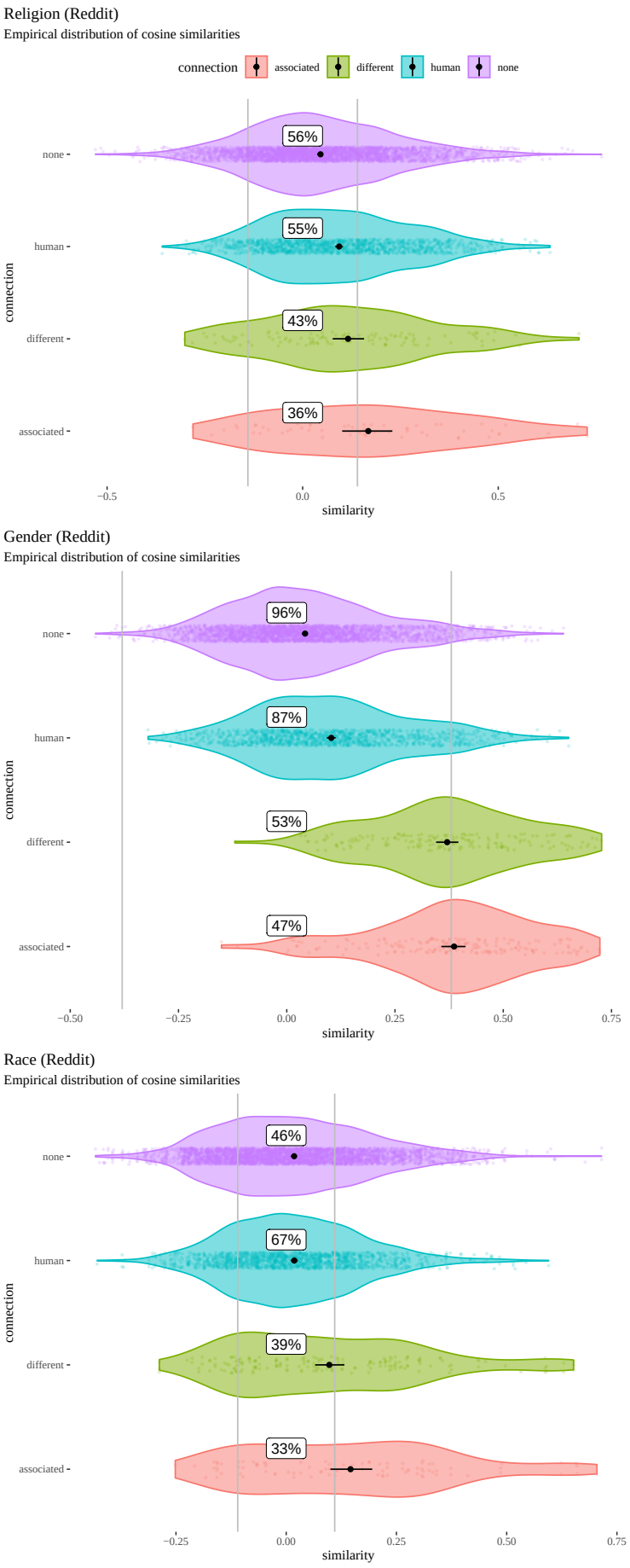


Figure 3: Empirical distributions of cosine similarities (1-distances) for three word lists used in the original paper.

Another issue to consider is the selection of attributes for bias measurement. The word lists used in the literature are often fairly small (5-50). The papers in the field do employ statistical tests to measure the uncertainty involved and do make claims of statistical significance. Yet, we will later on argue that these method are not proper for the goal at hand. By using Bayesian methods we will show that a more appropriate use of statistical methods leads to estimates of uncertainty which suggest that larger word lists would be advisable.

To avoid such problems, later on, we employ control groups and in line with Bayesian methodology, use posterior distributions and highest posterior density intervals instead of chasing single-point metrics based on pre-averaged data. But before we do so, let us spend some time explaining why pre-averaging is a sub-optimal strategy.

4. Pre-averaging, bootstrapping and testing with the null model

The approaches we have been describing use means of mean average cosine similarities to measure similarity between protected words and attributes coming from harmful stereotypes. However, a closer inspection look at the individual values that are taken for the calculations, it turns out that there are quite a few outliers and surprisingly dissimilar words. This problem becomes transparent when we examine the visualizations of the individual cosine distances, following the idea that one of the first steps in understanding data is to look at it.

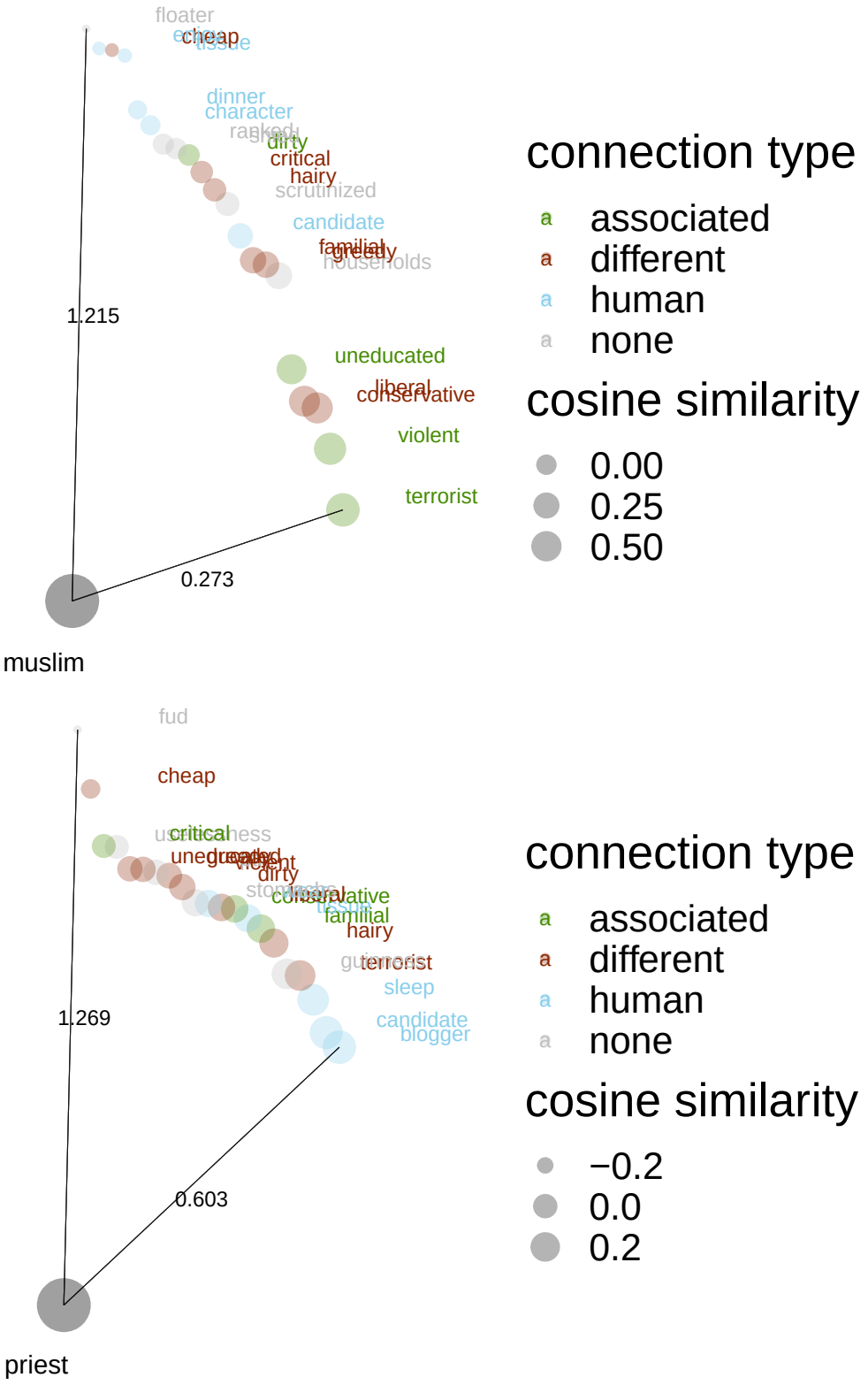


Figure 4: Actual distances for two protected words. For the protected word "muslim", the most similar attributes tend to be the ones associated with it stereotypically, but then words associated with other stereotypes come closer than neutral or human predicates. For the word "priest", the situation is even less as expected: the nearest attributes are human attributes, and all there seems to be no clear pattern to the distances to other attributes.

With such a method the uncertainty involved is not really considered which makes it even more difficult to give reasonable interpretations of the results. We propose the use of Bayesian method to obtain some understanding of the influence the uncertainty has on the interpretation of final results.
$s(X,Y,A,B)$ is the statistic used in the significance test, and the p-value is obtained by bootstrapping: it is the frequency of $s(X_i,Y_i,A,B) > s(X,Y,A,B)$ for all equally sized partitions X_i, Y_i of $X \cup Y$. The effect size is computed by normalizing the difference in means as follows:

$$bias(A,B) = \frac{\mu(\{s(x,A,B)\}_{x\in X}) - \mu(\{s(y,A,B)\}_{y\in Y})}{\sigma(\{s(w,A,B)\}_{w\in X\cup Y})} \tag{4}$$

The t-tests they employ are run on average cosines used to calculate MAC.

The effect sizes reported by [8] for lists of words for which the embeddings are supposedly biased range from 2.06 to 1.81, and the reported p-values are in the range of $10^{-7} - 10^{-2}$ with one exception for Math vs arts, where it is .018. The question is, are those results meaningful?

One way to answer this question is to think in terms of null models. If the words actually are samples from two populations with equal mean, how often would we see effect sizes in this range? How often would we reach the p-values that the authors reported?

We start our exploration with considering a list of 16 protected words, each class containing 8, with 16 attributes, 8 in each attribute class. In each particular sample we will investigate the consequence of the model being in fact null: we draw the cosine distances, every time, from the $\mathsf{Normal}(0,.08)$ distribution. 0.08 is approximately the empirical standard deviation observed in fairly large samples of neutral words, and 16 is the sample size used in the WEAT7 word list, which is not much different from the other sample sizes in word lists used by [8]

So let's draw one sample of this type. We calculate the s-values, the test statistics and the effect sizes. Then, following the original methodology, we obtain bootstrapped distributions of the effect sizes and the test statistics by calculating the same for each possible equal split of the initial random data set. The observed test statistic is 0.386222 and 1.2671948. The bootstrapped distributions of the test statistics and effect sizes are illustrated in Figure 5. Quite notably both (two-sided) p values are uneven and rather low.

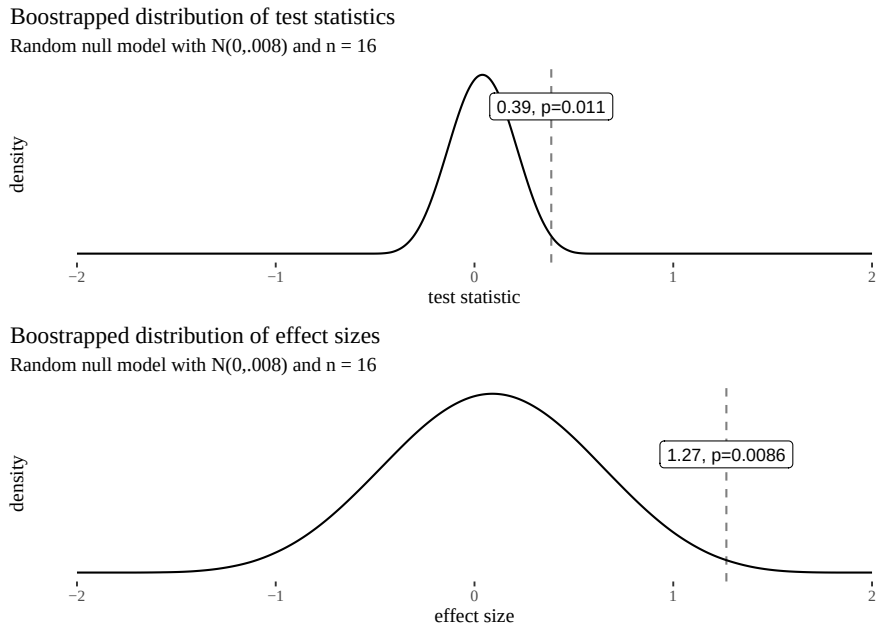


Figure 5: Bootstrapped distributions of test statistics and effect sizes in a random sample given the null hypothesis.

At this point, we might think—right, so while drawing on the assumption of the null hypothesis we just stumbled into a data set that randomly happened to display strong bias. We decide to double-check this by visual inspection expecting exactly this: a strong, clearly visible bias (Figure 6).

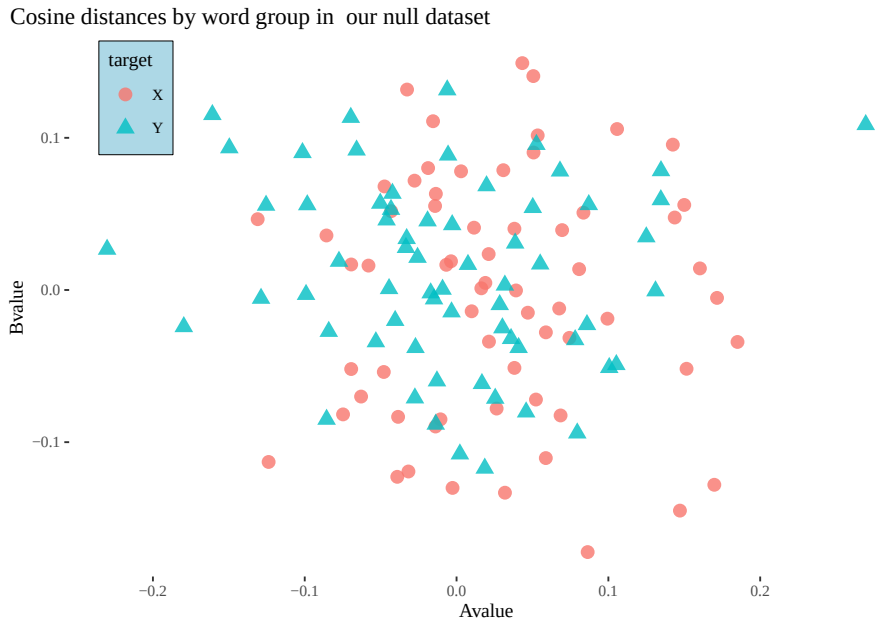


Figure 6: Cosine distances to two attribute sets by protected word groups. Observe nothing unusual except for a few outliers.

In fact, while there might be some outliers here and there, saying that a clear bias on

which one group is systematically closer to As than another is definitely a stretch. What happened?

In the calculations of WEAT means are taken twice. The s-values themselves are means, and then means of s-values are compared between groups. The test statistic used itself is of this sort as well—while it compares sums, these are sums of sets of the same size, so they are really linear transformations of means of means. Statistical troubles start when we run statistical tests on sets of means, for at least two reasons One, by pre-averaging data we throw away information about sample sizes. For the former point, think about proportions: 10 out of 20 and 2 out of 4 give the same mean, but you would obtain more information by making the former observation rather than by making the latter. And especially in this context, in which the word lists are not huge, sample sizes should matter. Two, when we pre-average, we remove variation, and so pre-averaging tend to manufacture false confidence. Group means display less variation than the raw data points, and the standard deviation of a set of means of sets of means is bound to be lower than the original standard deviation in the row data. Now, if you calculate your effect size by dividing by the pre-averaged standard deviation, your quite likely to get something that looks like a strong effect size, but the results of your calculations might not track anything interesting.

So let us think again about the question that we are ultimately interested in. Are the X terms systematically closer to (further from) the A attributes (B attributes) than the Y words? But now let's use the raw data points to try to answer these questions. As the first stab, let us run two quick t-tests to gauge what the raw data illustrated in Figure 6 tell us.

First, distances to A attributes for X words and Y words. Whoa, the result is significant! The p-value is 0.02. So sure, the sample is in some sense unusual. But the 95% confidence interval for the difference in means is $[.0052.061]$, clearly nothing that a reader would expect given that the calculated effect size seemed quite large. How about the B attributes? Here the p-value is .22 and the 95% confidence interval is $[-0.03, .009]$, even less of a reason to think a bias is present.

The difficulties are exacerbated by the fact that statistical tests are based on bootstrapping from a relatively small data sets, which is quite likely to underestimate the population variance. To make our point clear, let us avoid bootstrapping and work with the null generative model with $\mathsf{Norm}(0, .08)$ for both word groups. We keep the sizes the same: we have eight protected words in each group, sixteen in total, and for each we randomly draw 8 distances from hypothetical A attributes, and 8 distances from hypothetical B attributes. Calculate the test statistic and effect size the way [8] did. Do this 10000 times and look at what the distributions of these values are on the assumption of the null model with realistic empirically motivated raw data point standard deviation.

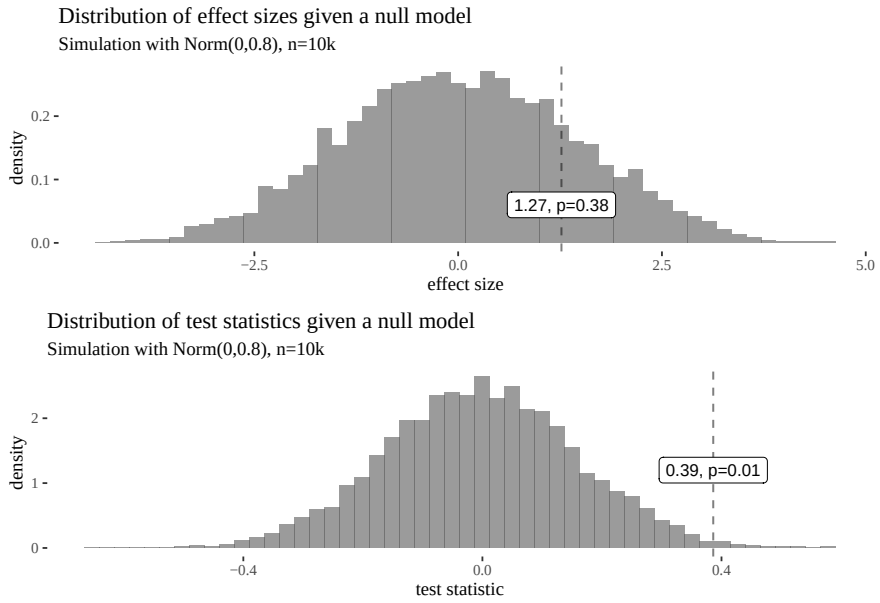


Figure 7: Distributions of test statistics and effect sizes based on 10k simulations on the assumption of a null model in which all distances come from normal distribution with $\mu = 0, \sigma = .08$.

The first observation is that the supposedly large effect size we obtained is not that unusual even assuming a null model. Around 38% of samples result in effect size at least as extreme. This illustrates the point that the effect size used does not track anything interesting. Second, the distribution of test statistics is much more narrow, which means that if we use it to calculate p-values, it is not too difficult to obtain a supposedly significant test statistic which nevertheless does not correspond to anything interesting happening in the data set.

Where do we stand now? We have seen that seemingly high effect sizes might arise even if the underlying processes actually have the same mean. The uncertainty resulting from including the raw data point variance into considerations is more extensive than the one suggested by the low p-values obtained from taking means or means of means as data points. In this section, we discussed the performance of the WEAT measure, but since the [12] one is a generalization thereof, including the method of running statistical tests on pre-averaged data, our remarks, *mutatis nutandis*, apply. Moreover, a general problem is that studies such as [8] or [12] do not use control groups, and this makes the interpretation of the analysis additionally difficult.

What is the alternative? As we already emphasized: focusing on what the real underlying question is, and trying to answer it using a statistical analysis of the raw data. Moreover, since the data sets are not too large and since multiple evaluations are to be made, we will pursue this method from the Bayesian perspective.

5.1. Introduction and model definition

Bayesian data analysis takes prior probability distributions, a mathematical model structure and the data, and returns the posterior probability distributions over the parameters of interest, thus capturing our uncertainty about their actual values. One important difference between such a result and the result of a classical statistical analysis is that classical confidence intervals (CIs) have a rather complicated and somewhat confusing interpretation, which has little to do with the posterior probability distribution.[12] [7]

In fact, Bayesian highest posterior density intervals (HPDIs, the narrowest intervals containing a certain ratio of the area under the curve) and CIs end up being numerically the same only if the prior probabilities are uniform. This illustrates that (1) classical analysis is unable to incorporate non-trivial priors, and (2) is therefore more susceptible to over-fitting, unless regularization (equivalent to a more straightforward Bayesian approach) is used. In contrast with CIs, the posterior distributions are easily interpretable and have direct relevance for the question at hand. Moreover, Bayesian data analysis is better at handling hierarchical models and small datasets, which is exactly what we will we dealing with.

In a standard Bayesian analysis, the first step is to understand the data and select potential predictors and the outcome variable. Once the data is understood, the next step is to formulate a mathematical description of the generative model of the relationships between the predictors and the outcome variable. Prior distributions must then be chosen for the parameters used in the model. Next, Bayesian inference must be applied to find posterior probabilities over the possible parameter values. Finally, after finding the posterior probabilities, it is important to check how well the posterior predictions reflect the data with a posterior predictive check. In our analysis the outcome variable is the cosine distances between the protected words and attribute words. The predictor is a factor determining whether a given attribute word is a neutral word, a human predicate, is stereotypically associated with the protected word, or comes from a different stereotype connected with another protected word. The idea is, if bias is present in the embedding, distances to associated attribute words should be systematically lower than to other attribute words.

Furthermore, conceptually there are two levels of analysis in our approach (see Figure 8). On the one hand, we are interested in the general question of whether related

[12] Here are a few usual problems. CIs are often mistakenly interpreted as providing the probability that a resulting confidence interval contains the true value of a parameter. CIs bring confusion also with regard to precision, it is a common mistake to interpret narrow intervals as the ones corresponding to a more precise knowledge. Another fallacy is to associate CIs with likelihood and stating that values within a given interval are more probable than the ones outside it. The theory of confidence intervals does not support the above interpretations. CIs should be plainly interpreted as a result of certain procedure (there are many ways to obtain CIs from a given set of data) that will in the long run contain the true value if the procedure is performed a fixed amount of times. For a nice survey and explanation of these misinterpretations, see [15]. For a psychological study of the occurrence of such misinterpretations, see REF. In this study, 120 researchers and 442 students were asked to assess the truth value of six false statements involving different interpretations of a CI. Both researchers and students endorsed, on average, more than three of these statements.

attributes are systematically closer across the dataset. On the other hand, we are interested in a more fine-grained picture of the role of the predictor for particular protected words. Learning in hierarchical Bayesian models involves using Bayesian inference to update the parameters of the model. This update is based on the observed data, and estimates are made at different levels of the data hierarchy. We use hierarchical Bayesian models in which we simultaneously estimate parameters at the protected word level and at the global level, assuming that all lower-level parameters are drawn from global distributions. Such models can be thought of as incorporating adaptive regularization, which avoids overfitting and leads to improved estimates for unbalanced datasets (and the datasets we need to use are unbalanced).

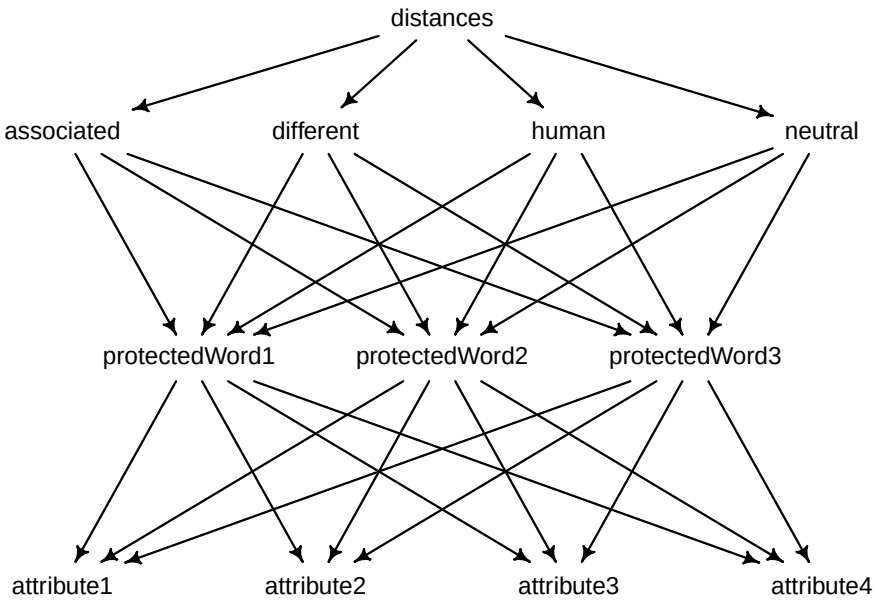


Figure 8: At a general level, we will be estimating the coefficients for distances as grouped by whether they are between protected words and attributes coming from their respective associated/different/human/neutral attribute groups. At a more fine-grained level, for each protected word we will be estimating the proximity of that word to attributes that are associated with its respective stereotype, come from a different stereotype, or come from the human/neutral attribute lists.

To be more specific, the underlying mathematical model is as follows. First, we assume that distances are normally distributed:

$$\mathsf{distance}_{\mathsf{i}} \sim \mathsf{dnorm}(\mu_i, \sigma_i)$$

Second, for each particular protected word pw there are four parameters to be estimated. Its mean distance to associated stereotypes $a[pw]$, its mean distance to attributes coming from different stereotypes, $d[pw]$, its mean distance to human attributes, $h[pw]$, and its mean distance to neutral attributes, $n[pw]$:

$$\mu_i = d_{\mathsf{pw[i]}} \times \mathsf{different}_i + a_{\mathsf{pw[i]}} \times \mathsf{associated}_i + h_{\mathsf{pw[i]}} \times \mathsf{human}_i + n_{\mathsf{pw[i]}} \times \mathsf{neutral}_i$$

where $\mathsf{different}, \mathsf{associated}, \mathsf{human}$ and $\mathsf{neutral}$ are binary variables.

This completes our description of the underlying mechanism. Now the priors and the hierarchy. We assume all the a parameters come from one distribution, that is normal around a higher-level parameter $\bar{a}$ and so on for the other three groups of parameters. That is, $a_{\text{pw[i]}}$ is the average distance of a given particular protected word to attributes stereotypically associated with it, while $\bar{a}$ is the overall average distance of protected words to attributes associated with them [10,13].

$$
\begin{aligned}
d_{\text{pw[i]}} &\sim \text{Norm}(\bar{d}, \overline{\sigma_d}) \\
a_{\text{pw[i]}} &\sim \text{Norm}(\bar{a}, \overline{\sigma_a}) \\
h_{\text{pw[i]}} &\sim \text{Norm}(\bar{h}, \overline{\sigma_h}) \\
n_{\text{pw[i]}} &\sim \text{Norm}(\bar{n}, \overline{\sigma_n})
\end{aligned}
$$

According to our priors, the group means $\bar{a}$, $\bar{d}$, $\bar{h}$ and $\bar{n}$ all come from one normal distribution with mean equal to 1 and standard deviation equal to .3. The standard deviations $\bar{\sigma}_a$, $\bar{\sigma}_d$, $\bar{\sigma}_h$ and $\bar{sigma}_n$ to be estimated, according to our prior, come from one distribution, exponential with rate parameter equal to 2. Our priors are slightly skeptical. They do reflect our knowledge and intuition on the probable distribution of the cosine distances in the data. We know that the cosine distances lie in the range $0-2$, and we expect two randomly chosen vectors from the embedding to have rather small similarity, so we expect the distances to be centered around 1. However, we use a rather wide standard deviation (.3) to easily account for cases where there is actually much higher similarity between two vectors (especially in cases where the embedding is supposed to be biased). Our priors for the standard deviations are also fairly weak.[13]

$$
\begin{aligned}
\bar{d}, \bar{a}, \bar{h}, \bar{n} &\sim \text{Norm}(1, .3) \\
\overline{\sigma_d}, \overline{\sigma_a}, \overline{\sigma_h}, \overline{\sigma_n} &\sim \text{Exp}(2)
\end{aligned}
$$

5.2. Posterior predictive check

A posterior predictive check is a technique used to evaluate the fit of a Bayesian model by comparing its predictions with observed data. The underlying principle is to generate simulated data from the posterior distribution of the model parameters and compare them with the observed data. If the model is a good fit to the data, the simulated data should resemble the observed data. A posterior predictive check is crucial because it provides a way to assess the validity of a Bayesian model. This is important because a model that fits the observed data well may still fail to generalize to new data. In the figure below we can see that the model fits the the data well.

[13] To build a model one may use original *Stan* which is a state-of-the-art platform for statistical modeling and high-performance statistical computation (https://mc-stan.org/). The simpler option is to use *Rethinking* package from the *Statistical rethinking* course. With one of the functions *ulam* one may build RStan models from the formulas in an easy manner

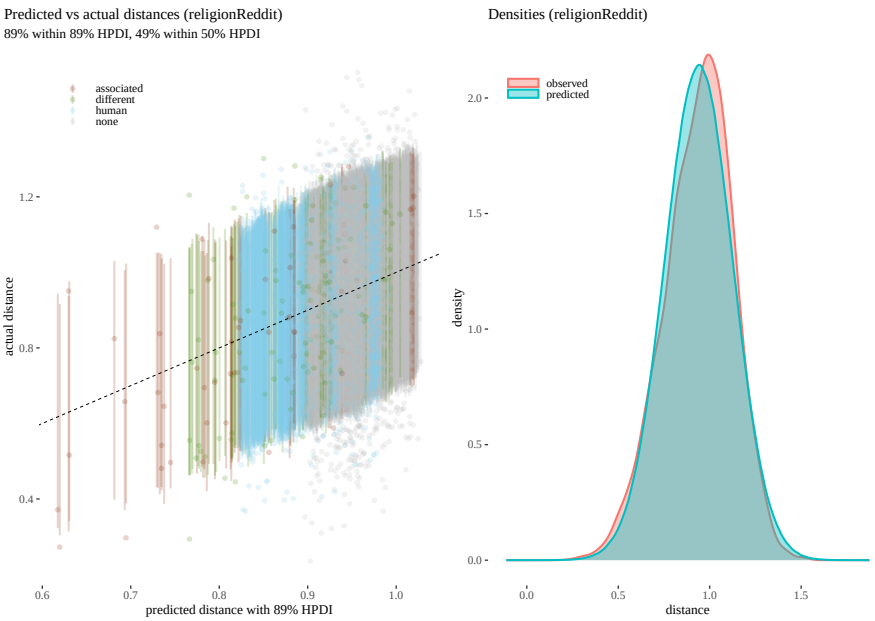


Figure 9: Example of a posterior predictive check. (Left) Actual cosine distances are plotted against mean predictions with 89% highest posterior density intervals. Notice that 90% of actual values fall within the 89% HPDI and 55% of actual values fall into 50% HPDI, which indicates appropriate performance of the model. The left-right alignment of differrent colors coressponds to the fact that cosine differences between elements of different categories differ, to some extent systematically (this will be studied in the results section). (Right) Densities of predicted and observed distances.

6. Results and discussion

6.1. Observations

In brief, despite one-number metrics suggesting otherwise, our Bayesian analysis reveals that insofar as the short word lists usually used in related research projects are involved, there usually are no strong reasons to claim the presence of systematic bias. Moreover, comparison between the groups (including control word lists) leads to the conclusion that the effect sizes (that is, the absolute differences between cosine distances between groups) tend to be rather small, with few exceptions. Moreover, the choice of protected words is crucial — as there is a lot of variance when it comes to the protected word-level analysis.

In a bit more detail, the visualizations in Appendix A.1 show that the situation is more complicated than merely looking at one-number summaries might suggest. Note that the axes are sometimes in different scales to increase visibility.

To start with, let us look at the association-type level coefficients (illustrated in the top parts of the plots). Depending on the corpus used and word class, there is a large variety as to the posterior densities. Quite aware of this being a crude approximation, let's compare the HPDIs and whether they overlap for different attribute groups.

- In Weat 7 (Reddit) there is no reason to think there are systematic differences between cosine distances (recall that words from Weat 7 were mostly not available in other embeddings).
- In Weat 1 (Google, Glove and Reddit) associated words are somewhat closer, but the cosine distance differences from neutral words are very low, and surprisingly it is human attributes, not neutral predicates that are systematically the furthest.

- In Religion (Google, Glove, Reddit) and Race (Google, Glove), the associated attributes are not systematically closer than attributes belonging to different stereotypes, and the difference from neutral and human predicates is rather low, if noticeable. The situation is interestingly different in Race (Reddit) where both human and neutral predicates are systematically further than associated and different attributes - but even then, there is no clear difference between associated and different attributes.
- For Gender (Google, Glove), despite the superficially binary nature, associated and opposite attributes tend to be more or less in the same distances, much closer than neutral words (but not closer than human predicates in Glove). Reddit is an extreme example: both associated and opposite attributes are much closer than neutral and human (around .6 vs. .9), but even then, there seems to be no reason to think than cosine distances to associated predicates are much different from distances to opposite predicates.

Moreover, when we look at particular protected words, the situation is even less straightforward. We will just go over a few notable examples, leaving the visual inspection of particular results for other protected words to the reader. One general phenomenon is that–as we already pointed out–the word lists are quite short, which contributes to large uncertainty involved in some cases.

- For some protected words the different attributes are somewhat closer than the associated attributes.
- For some protected words, associated and different attributes are closer than neutral attributes, but so are human attributes.
- In some cases, associated attributes are closer, but so are neutral and human predicates, which illustrates that just looking at average cosine similarity as compared to the theoretically expected value of 1, instead of running comparison to neutral and human attributes is misleading.
- The only group of protected words where differences are noticeable at the protected word level is Gender-related words– as in Gender (Google) and in Gender (Reddit) — note however that in the latter, for some words the opposite attributes seem to be a little bit closer than the associated ones.

6.2. Rethinking debiasing

These visualizations can be also handy when it comes to the investigation of the effect that debiasing has on the embedding space. Below one may see chosen pair depicting the difference in means with 89% highest posterior density intervals before and after applying debiasing.

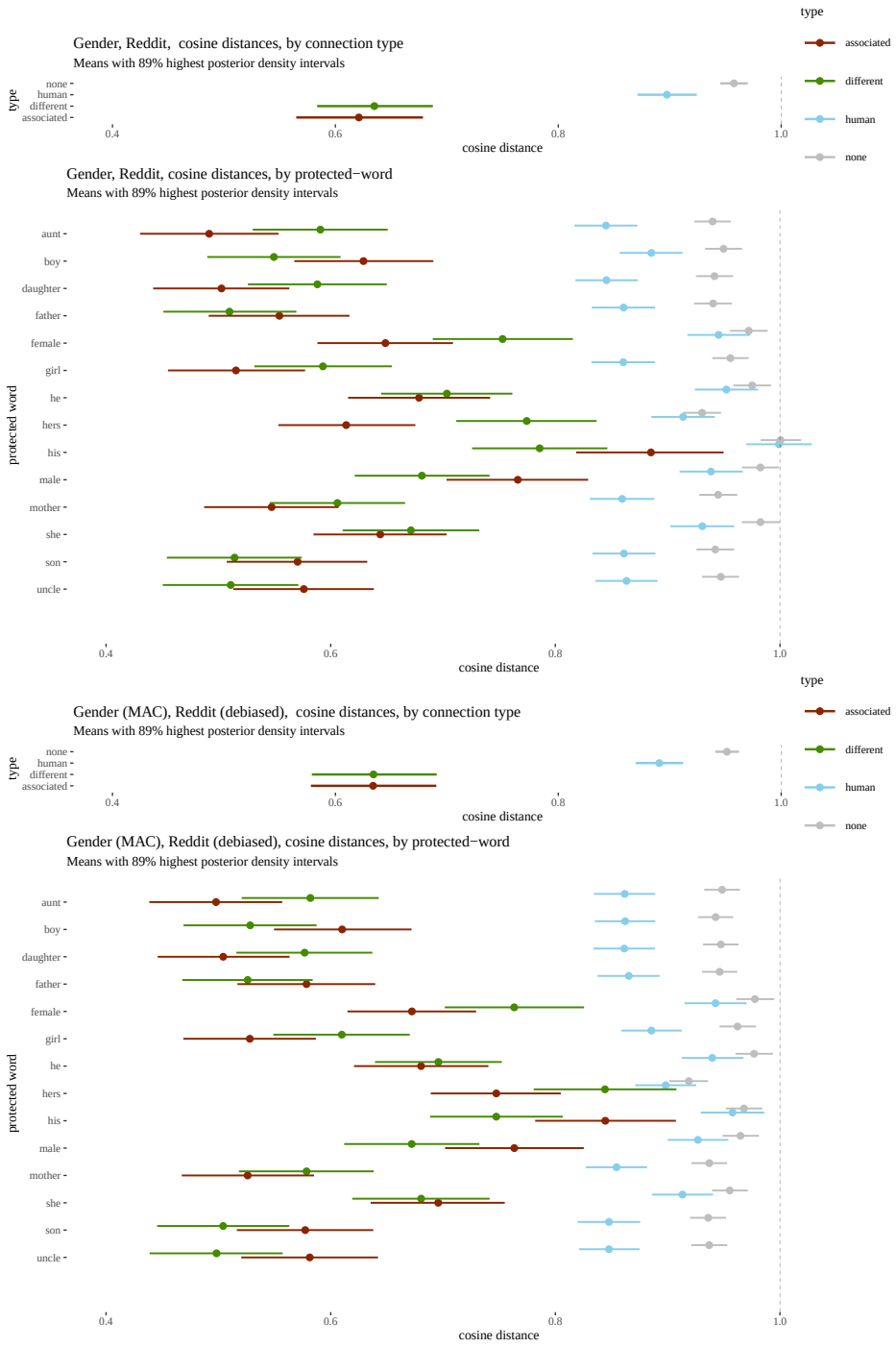


Figure 10: Means with 89% highest posterior density intervals for gender before and after debiasing.

In Figure 11 we look at the empirical distributions for the debiased embeddings. Comparing the results to the original embedding, one may notice that for the Religion group the neutral and human distribution has changed slightly. Before within the "correct" cosine similarity boundaries there were 56% of neutral and 55% of human word lists. After the debiasing the values changed to 59% (for neutral) and 59% (for human). The different and associated word lists were more influenced. The general shape of both distributions is less stretched. Before debiasing 43% of the different word lists and 35% of the associated word lists were within the accepted boundaries. After the embedding manipulation the percentage has increased for both lists to 63%. Visualization for Gender group illustrates almost no

change for the neutral and human word lists before and after debiasing. The values for different and associated word lists are also barely impacted by the embedding modification. In the Race group, the percentage within the boundaries for neutral and associated word lists have increased. The opposite happened for human and different word lists, where the percentage of "correct" cosine similarity dropped from 67% to 55% (human) and from 39% to 36% (different).

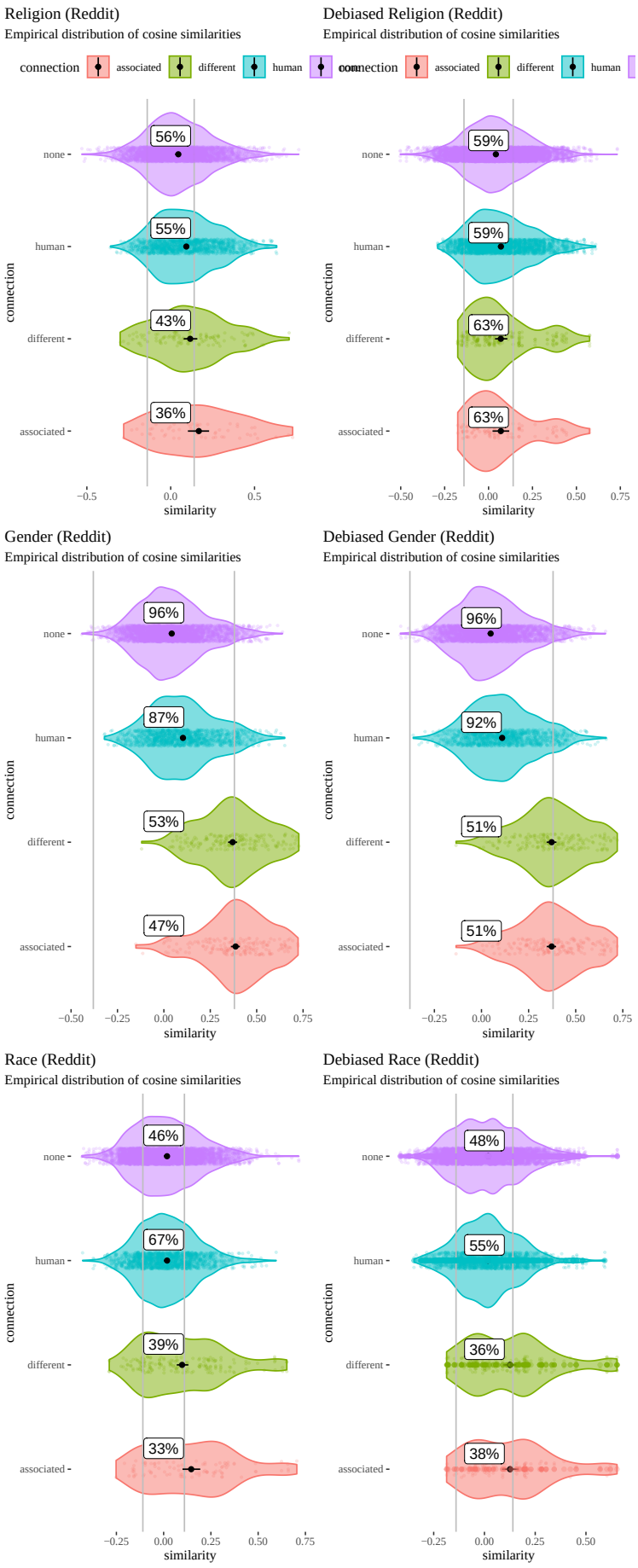


Figure 11: Empirical distributions of debiased cosine similarities (1-distances) for three word lists used in the original paper.

7. Related work and conclusions

It has been suggested [2] to use Bernstein bounds to state the uncertainty of the sample bias in a form of confidence interval. The interpretation of this interval is as follows, 95% confidence interval contains such a range of values, that one can be 95% certain that it contains the population mean. As we will prove later in the text, confidence intervals are quite problematic for several reasons, among others the confusing interpretability. In [2] the current problem of bias estimation was correctly identified. Indeed, the bias estimate should not be expressed as a single number without taking into account that the estimate is made on a sample of data and therefore has intrinsic uncertainty. However, we argue that Bernstein bounds do not provide the best solution to this problem. Applying their method to a popular WinoBias dataset leads to the conclusion that more than 11903 samples are needed to claim a 95% confidence interval for a bias estimate. This amount exceeds existing datasets for bias measurement. We propose a more realistic method that can be applied to commonly used datasets to estimate the uncertainty of the bias measurement.

More references to similar papers

A Bayesian data analysis with hierarchical models of cosine distances between protected words, control group words, and stereotypical attributes provides more modest and realistic assessment of the uncertainty involved. It reveals that much complexity is hidden when one instead chases single bias metrics present in the literature.

After introducing the method, we apply it to multiple word embeddings and results of supposed debiasing, putting forward some general observations that are not exactly in line with the usual picture painted in terms of WEAT or MAC (and the problem generalizes to any approach that focuses on chasing a single numeric metric): the word list sizes and sample sizes used in the studies are usually small,[14] posterior density intervals are fairly wide, often the differences between associated, different and neutral human predicates, are not very impressive. Also, a preliminary inspection suggests that the desirability of changes obtained by the usual debiasing methods is debatable. The tools that we propose, however, allow for a more fine-grained and multi-level evaluation of bias and debiasing in language models without losing modesty about the uncertainties involved.

[14] Depending on a list for [8] the range for protected words is between 13 and 100, and for attributes between 16 and 25; for [12] the range for protected words is between 14 and 18, and for attributes between 11 and 25.

A. Appendix

A.1. Visualizations

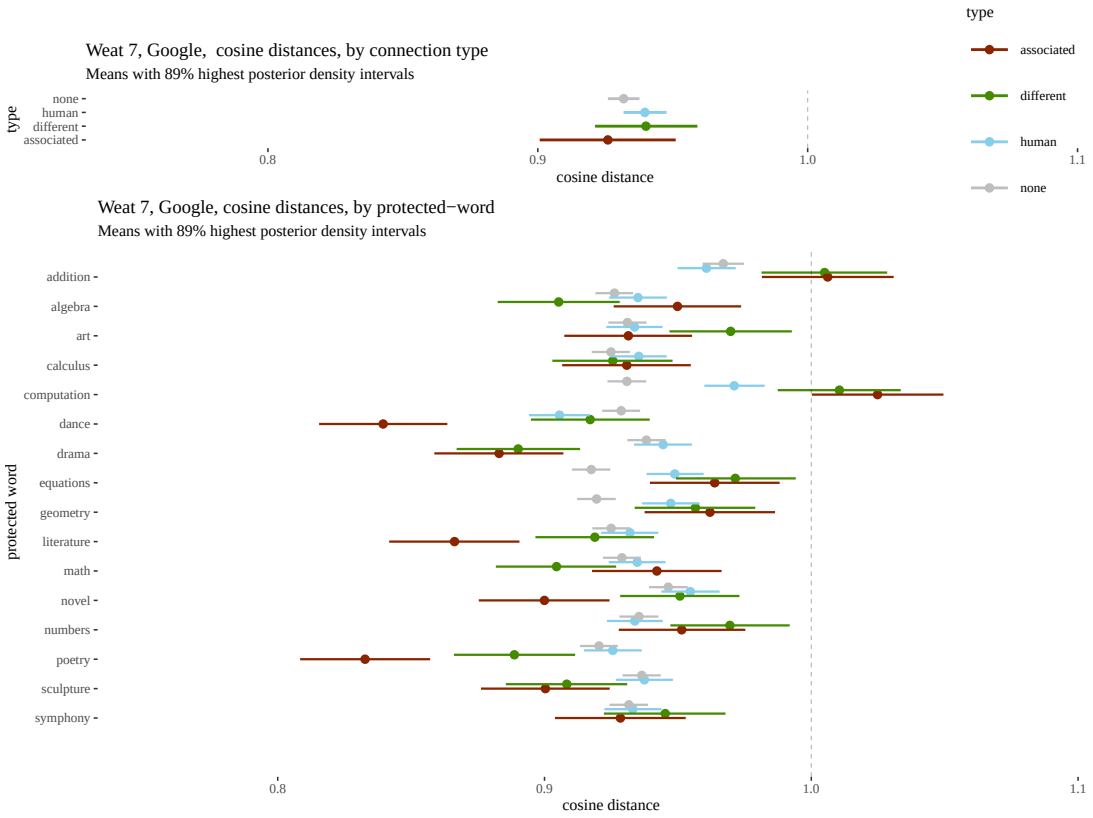


Figure 12: dsds

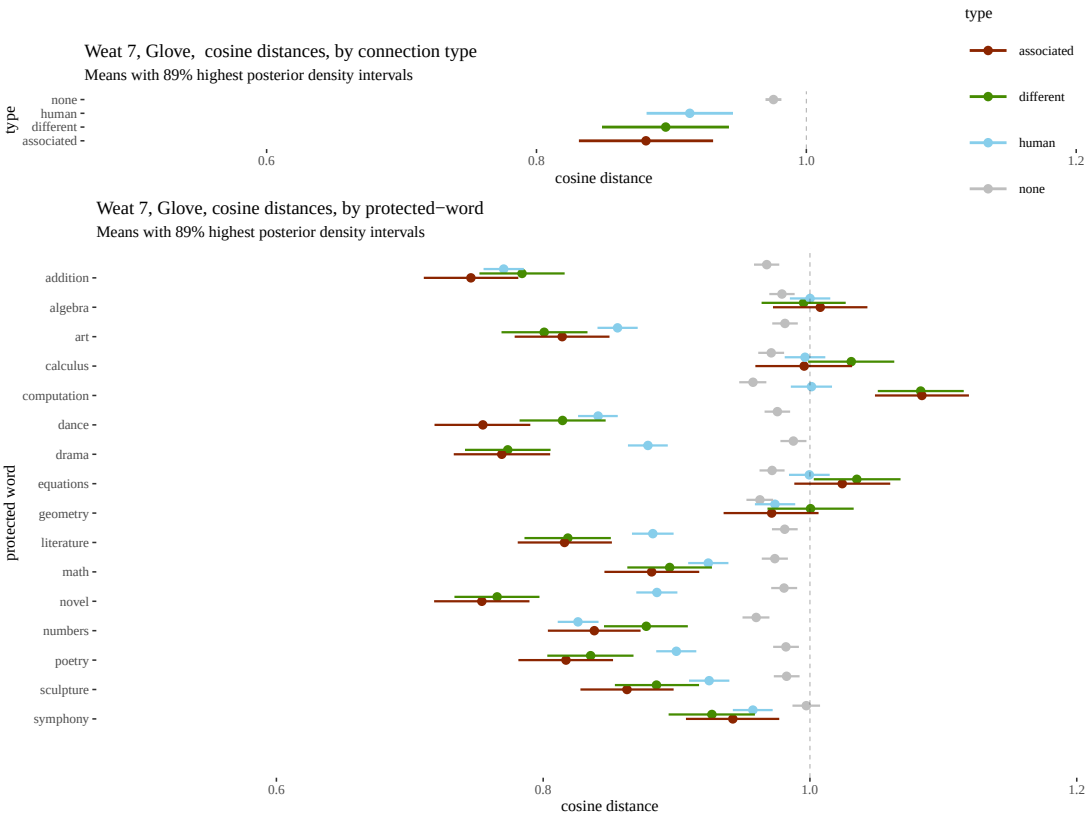


Figure 13: dsds

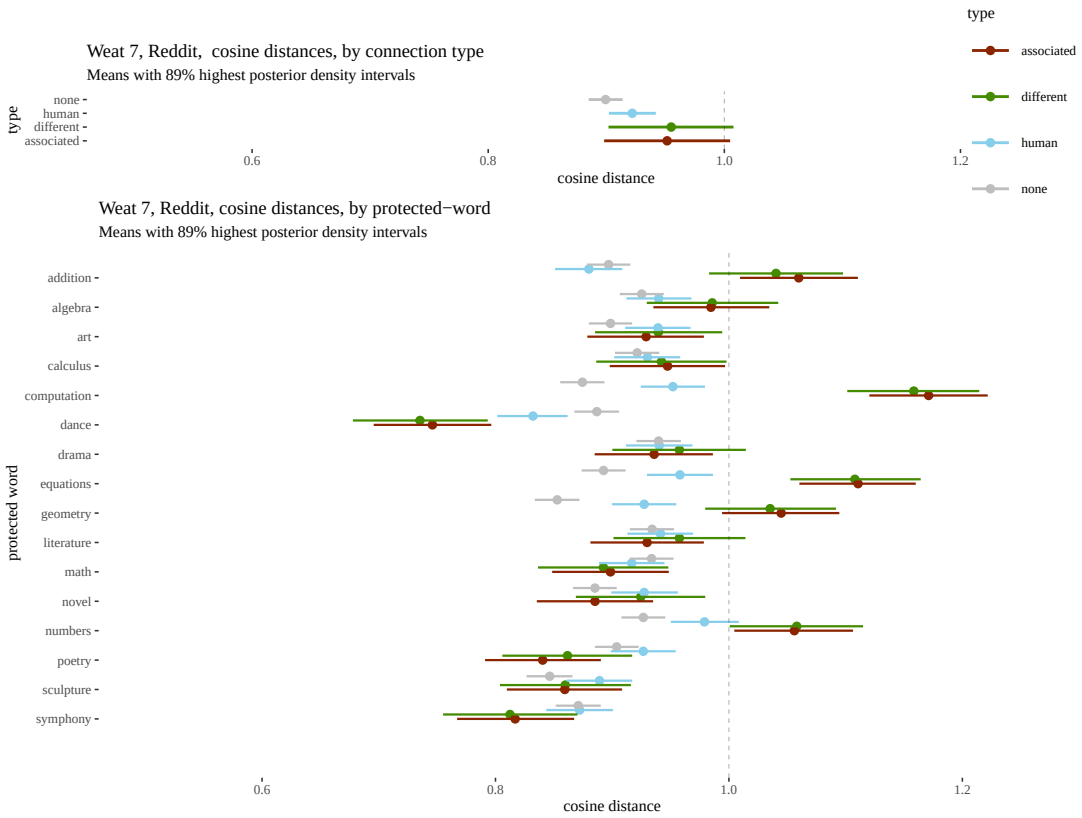


Figure 14: dsds

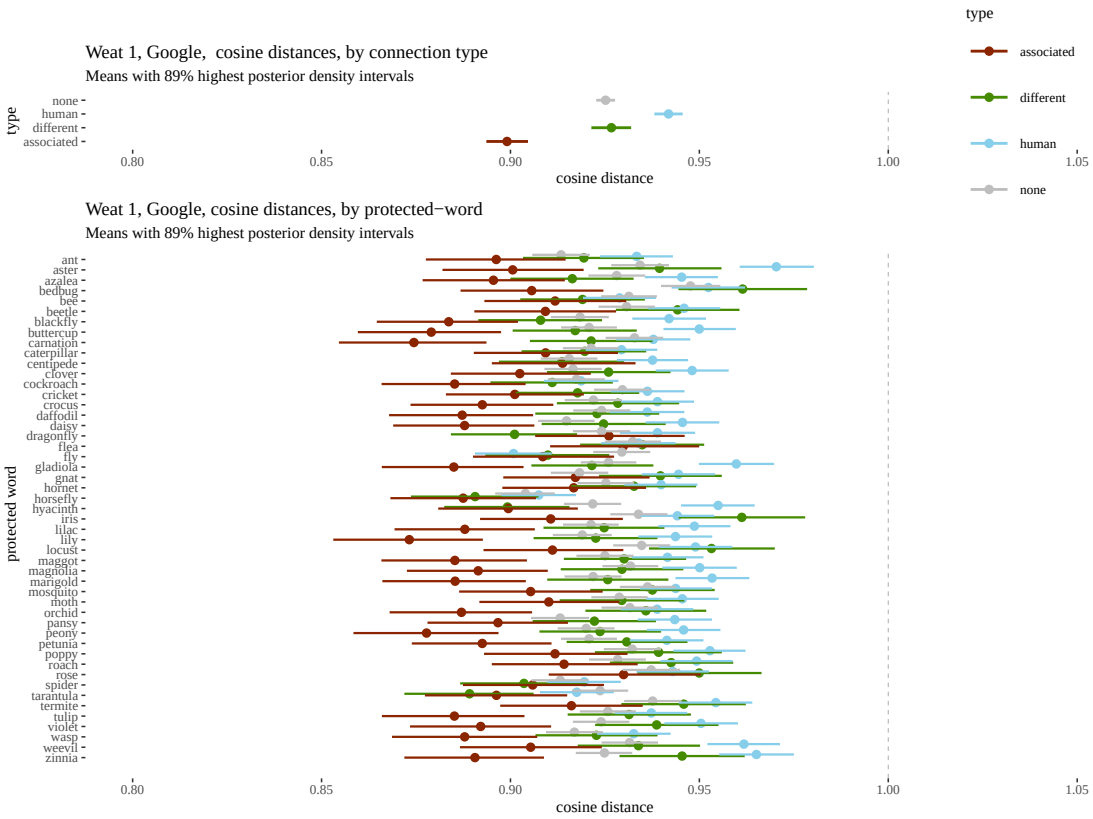


Figure 15: dsds

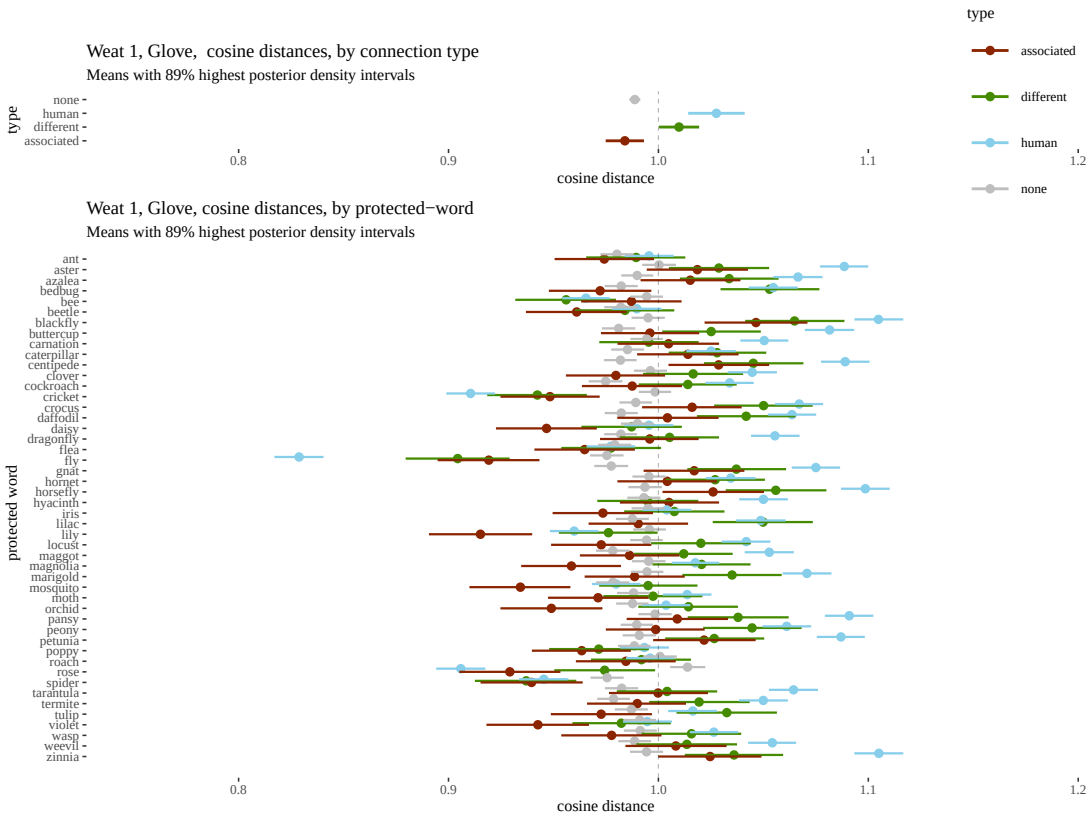


Figure 16: dsds

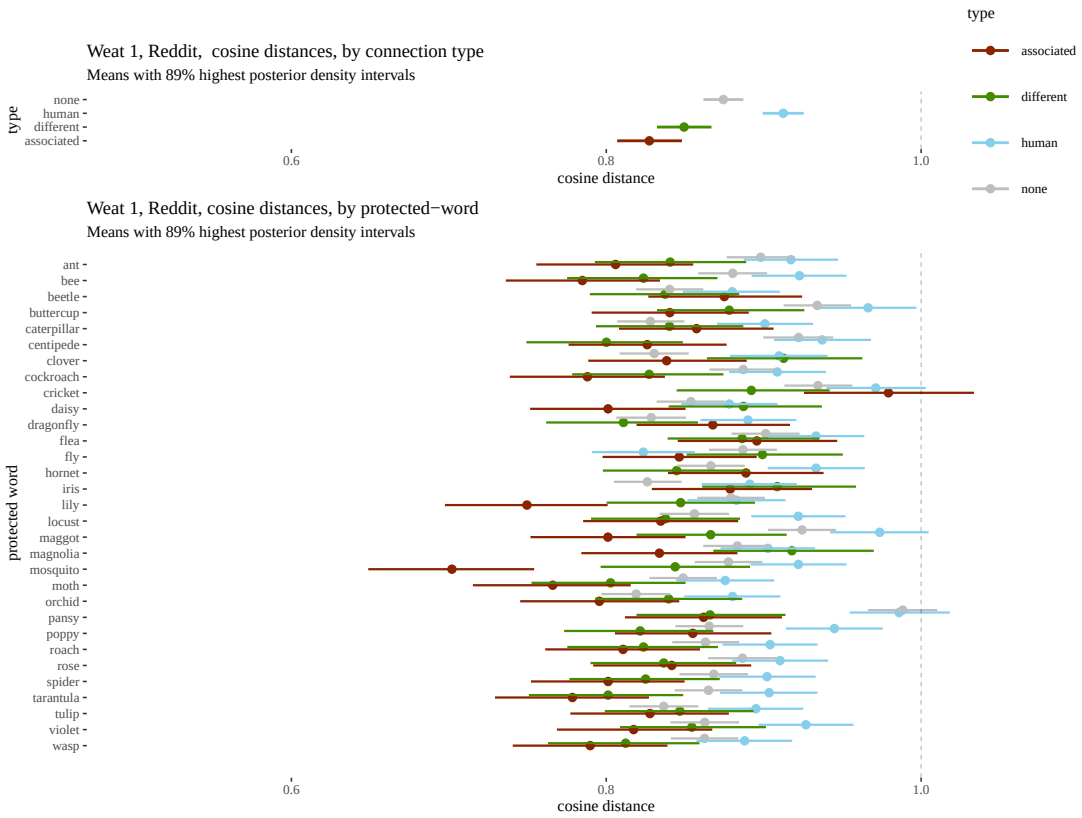


Figure 17: dsds

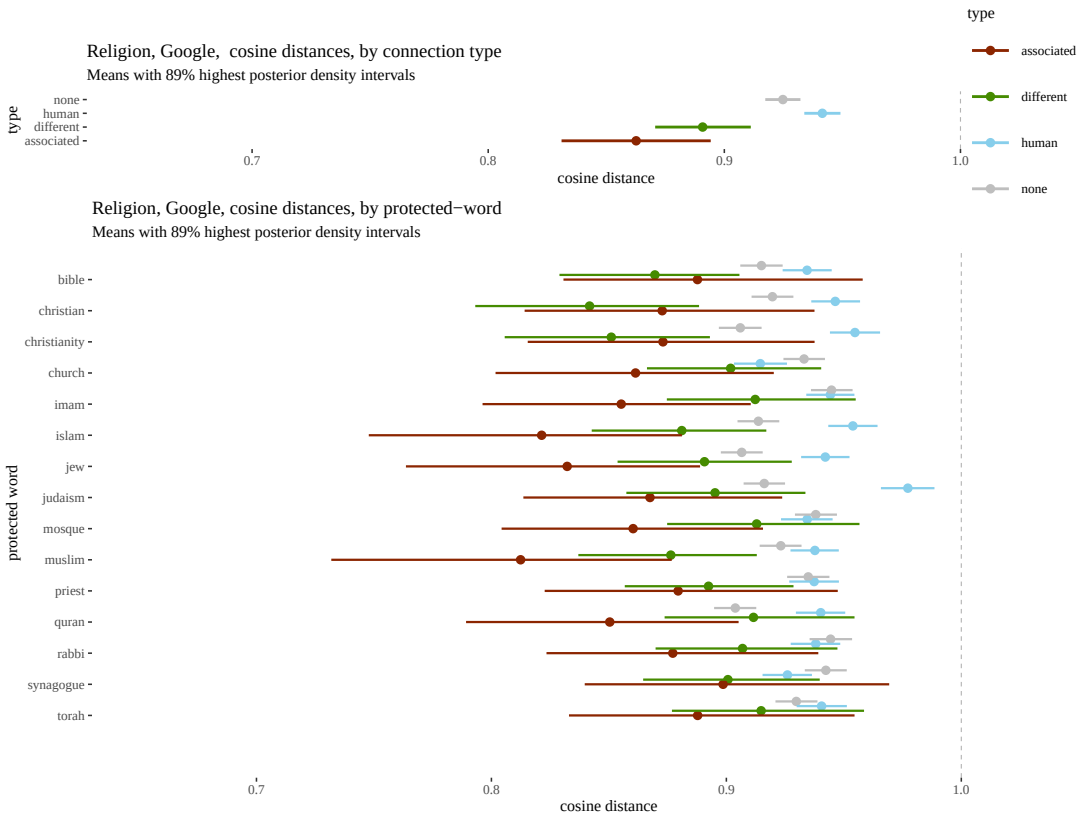


Figure 18: dsds

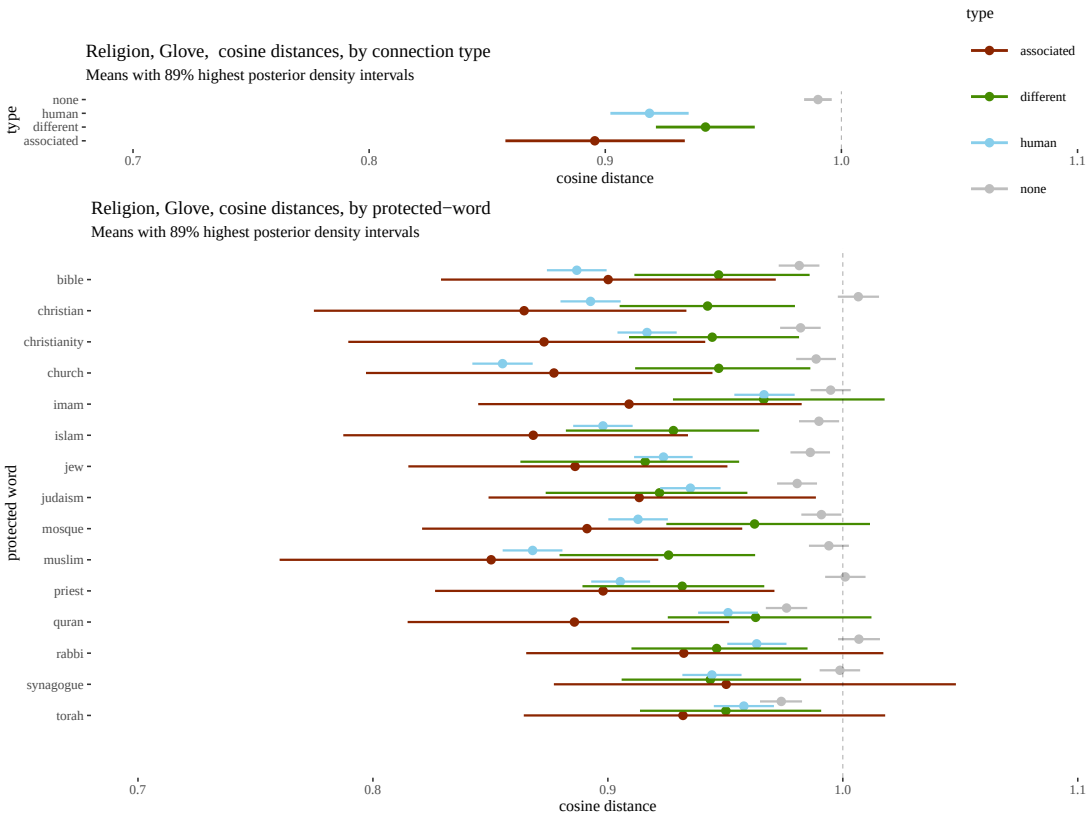


Figure 19: dsds

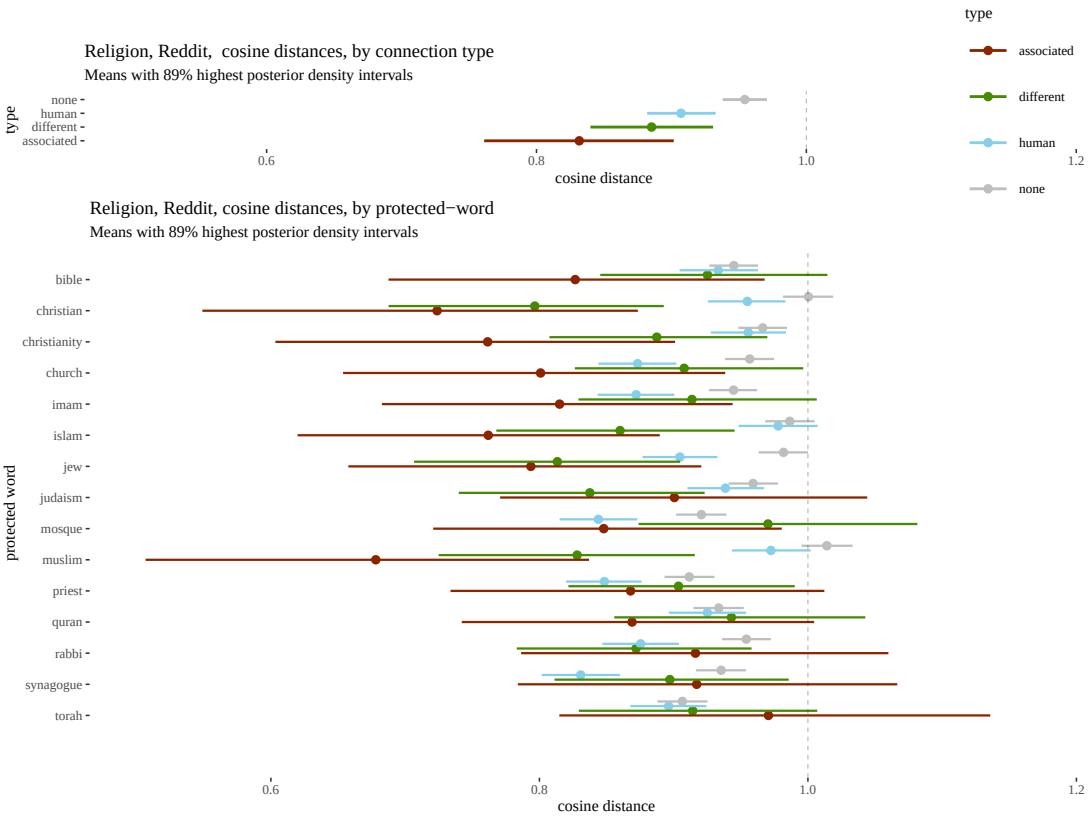


Figure 20: dsds

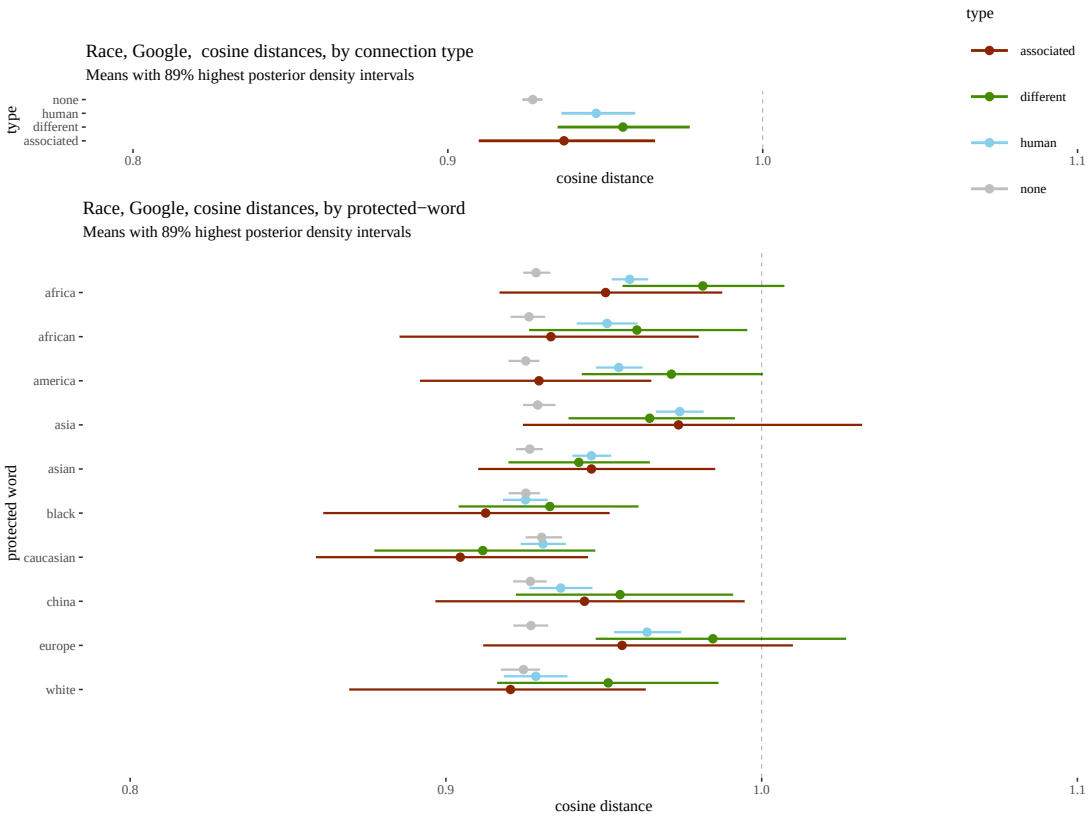


Figure 21: dsds

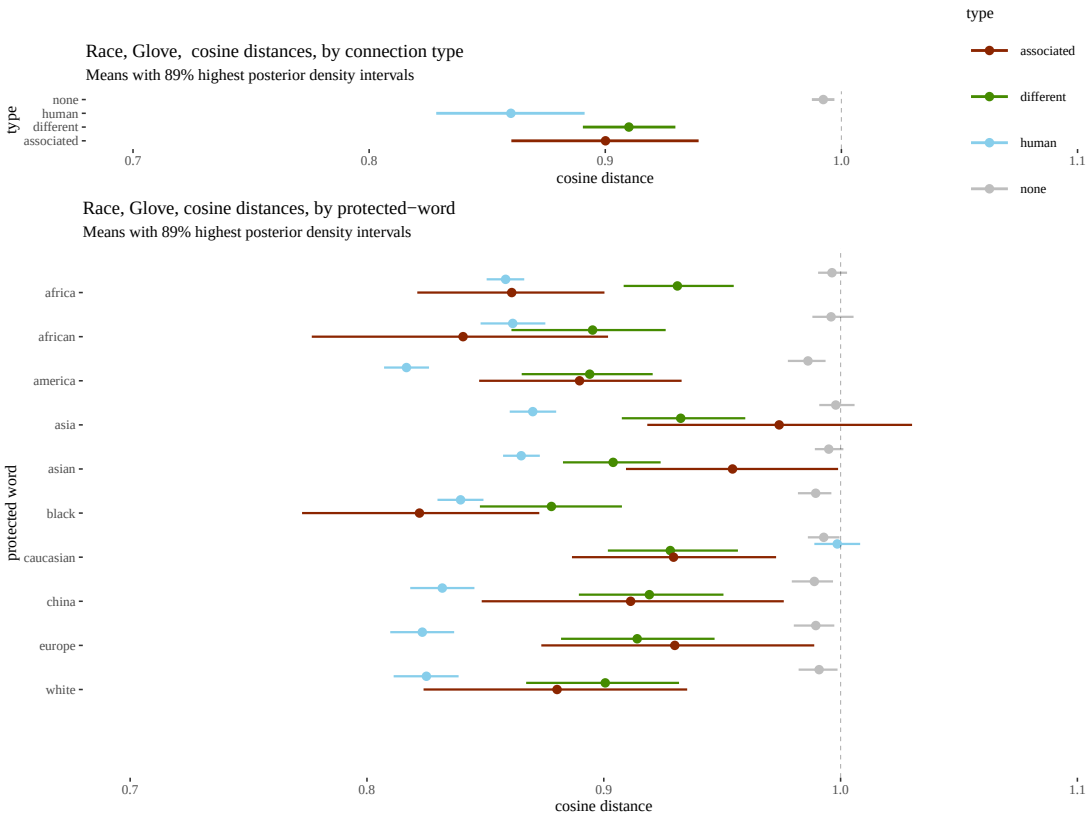


Figure 22: dsds

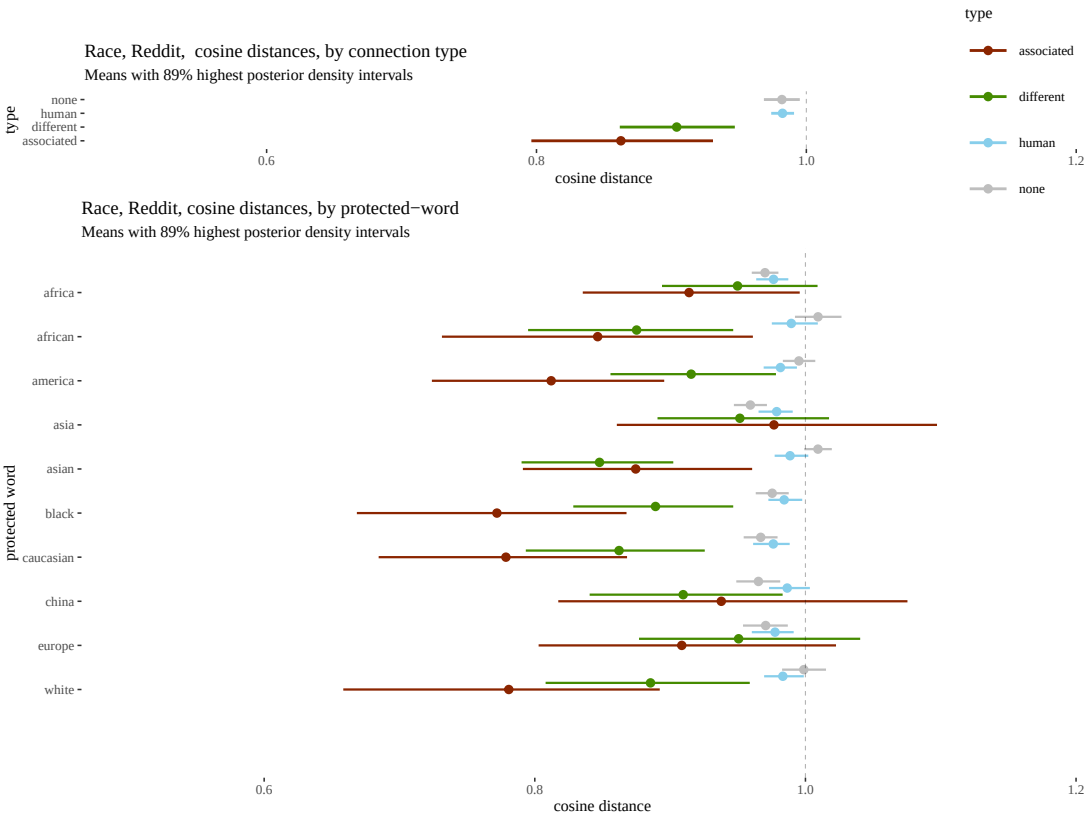


Figure 23: dsds

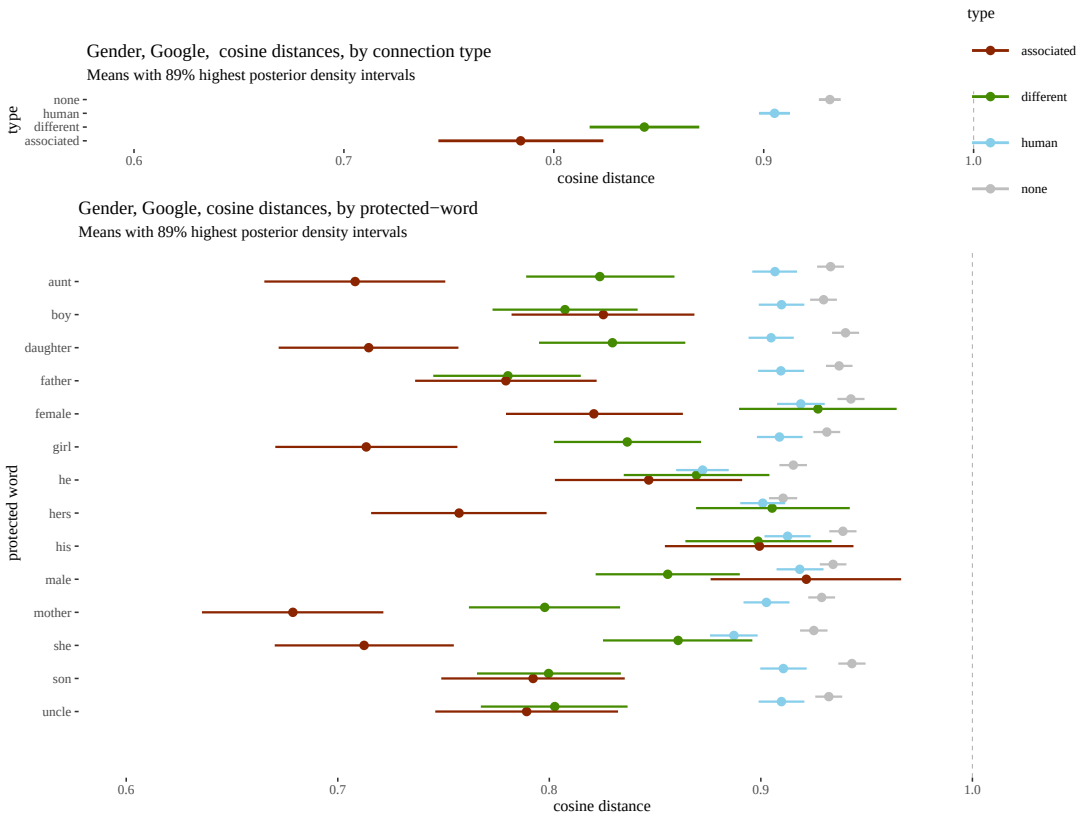


Figure 24: dsds

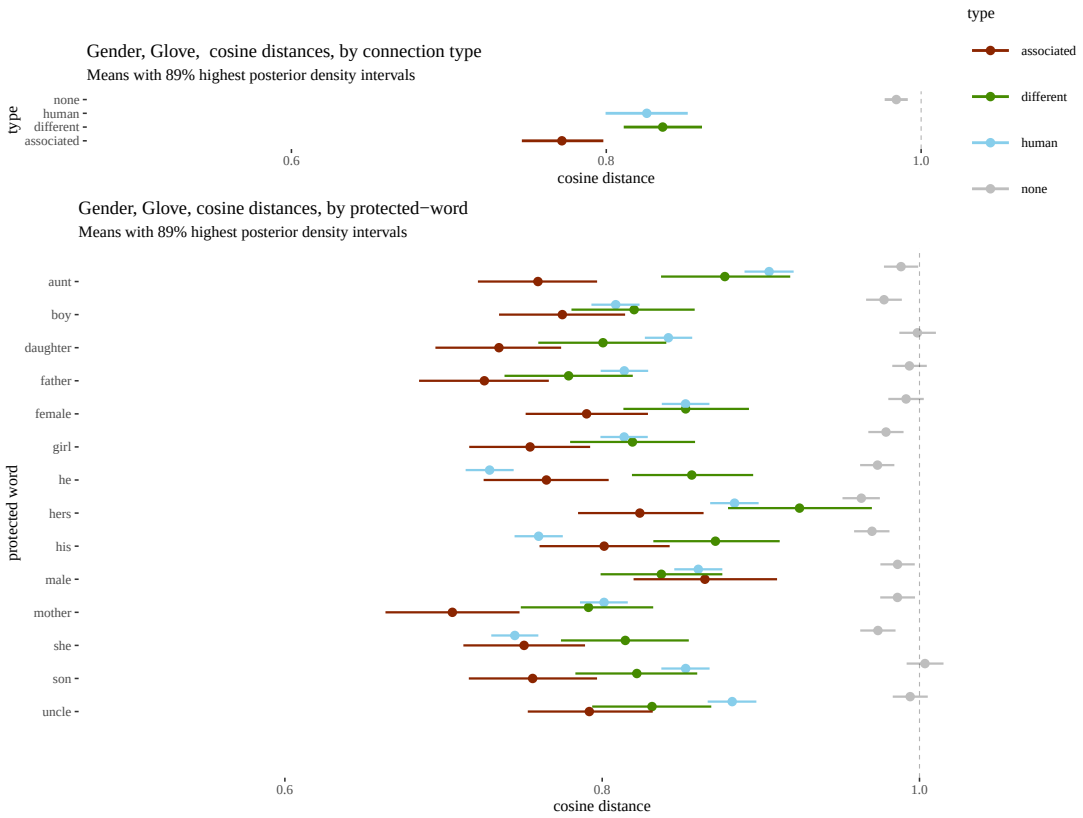


Figure 25: dsds

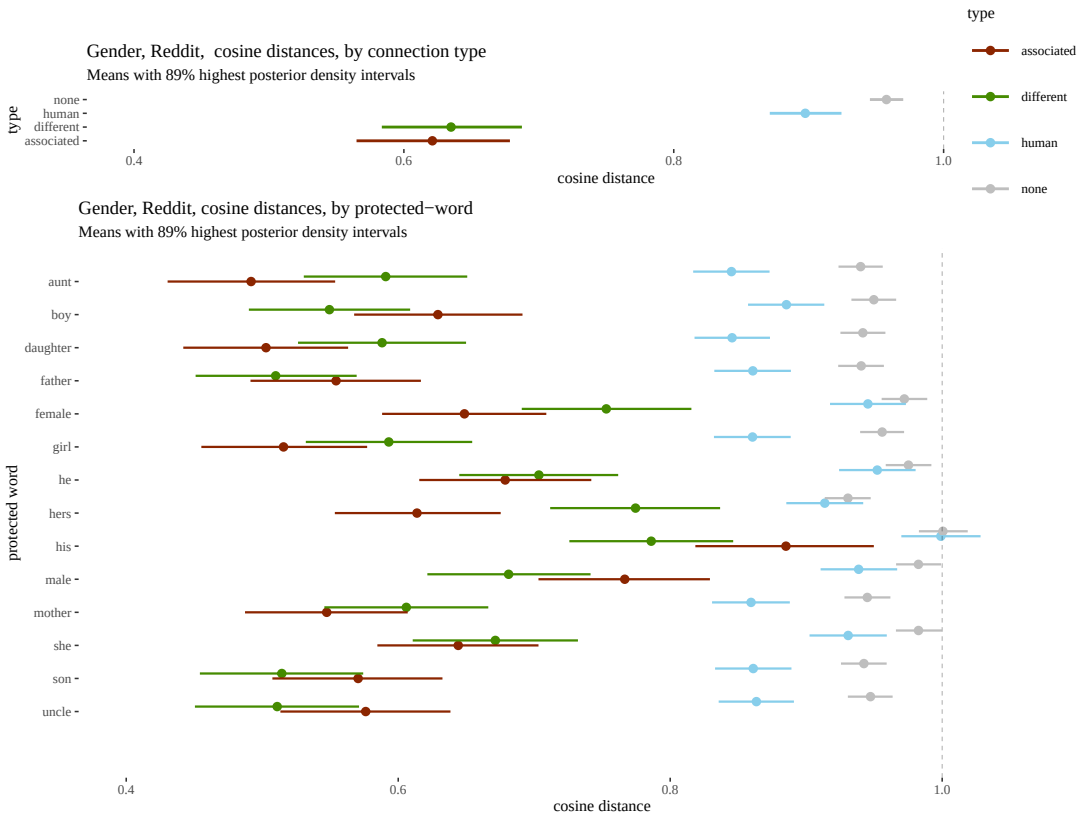


Figure 26: dsds

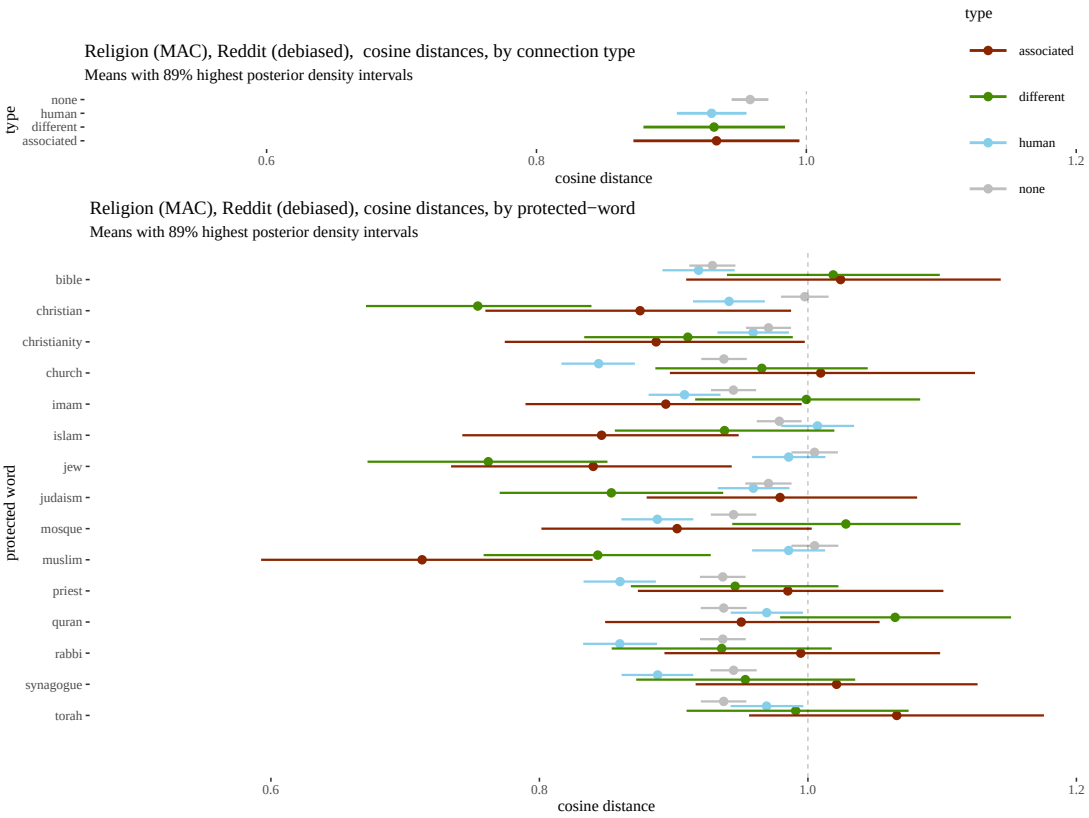


Figure 27: dsds

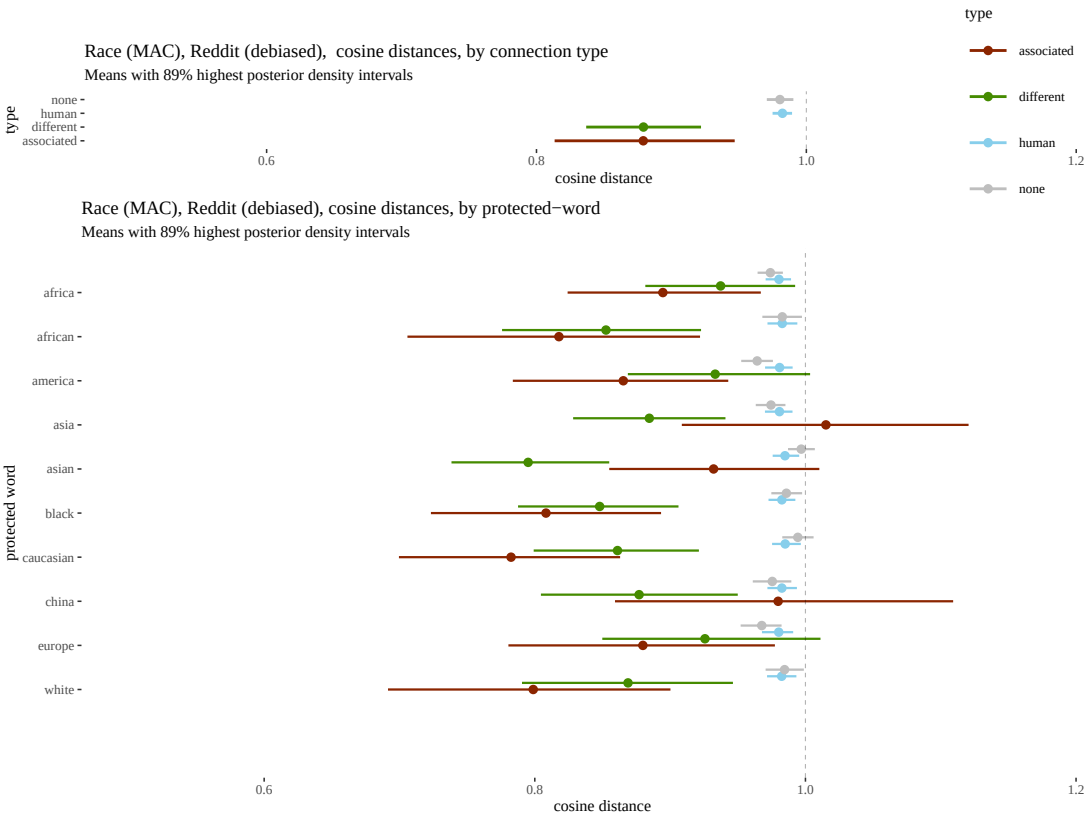


Figure 28: dsds

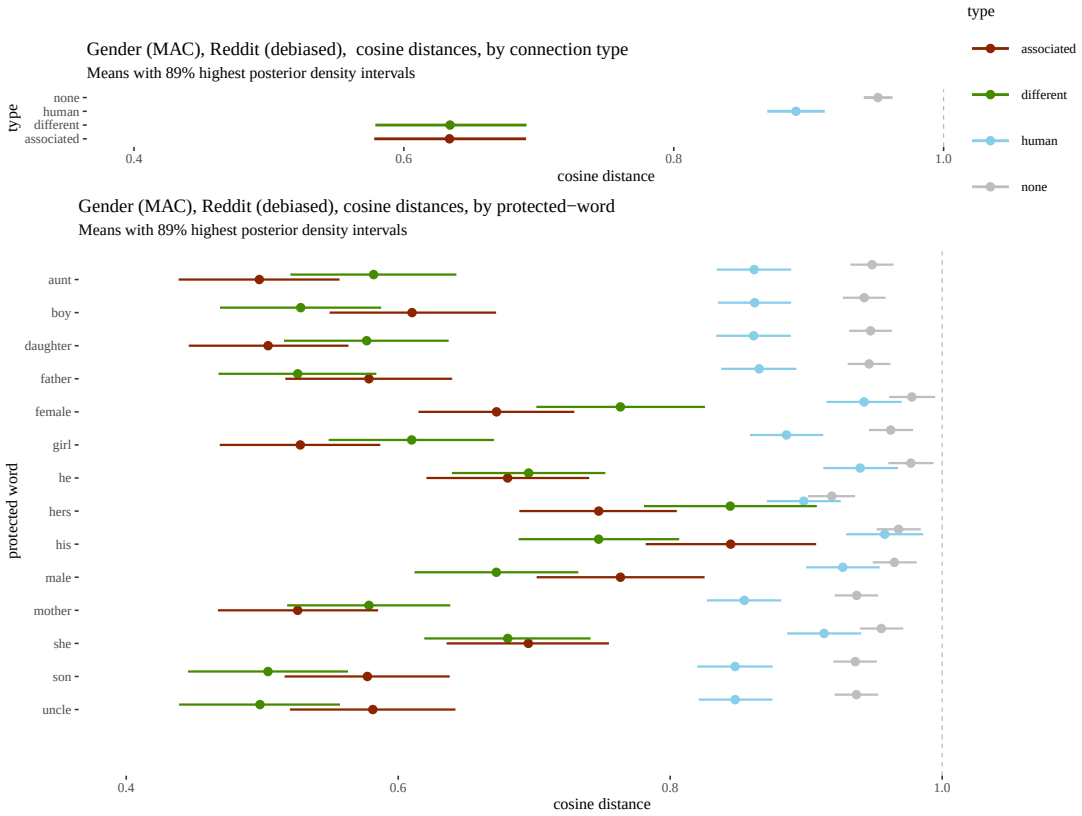
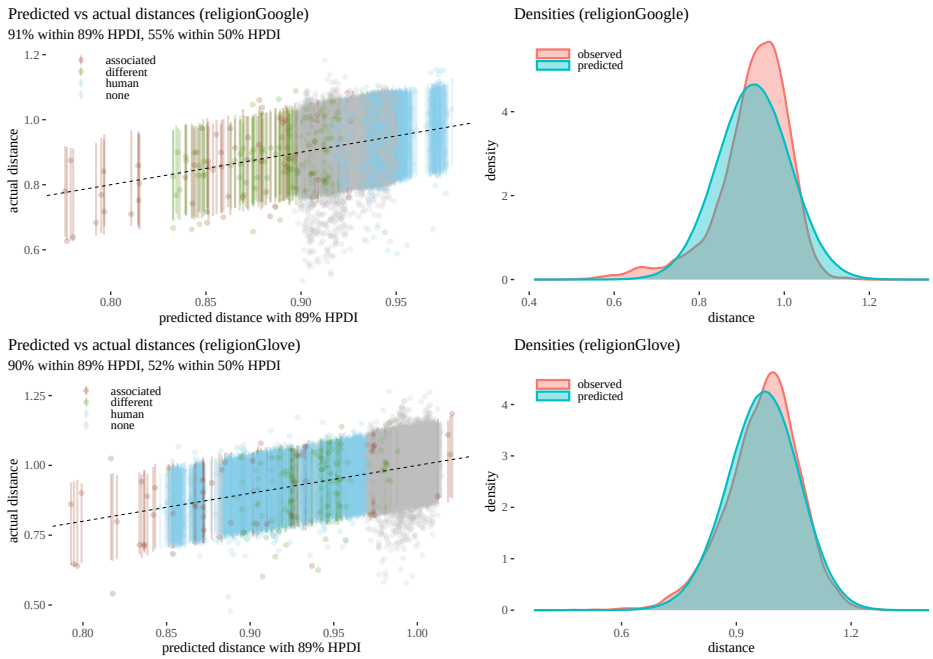
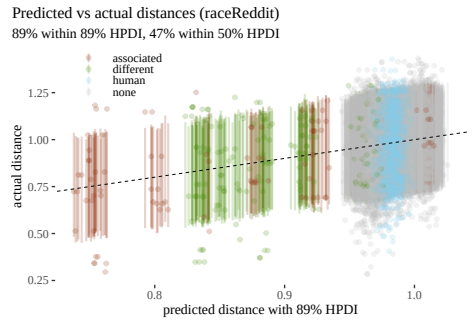
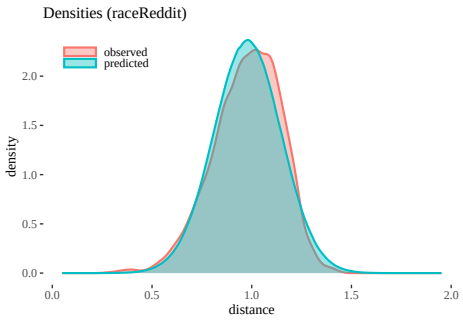


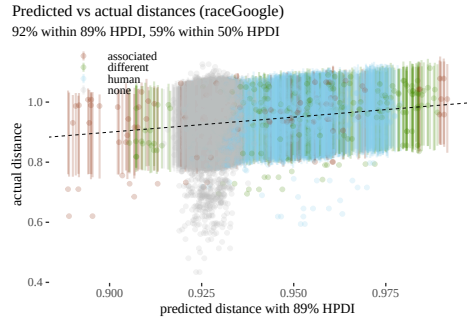
Figure 29: dsds

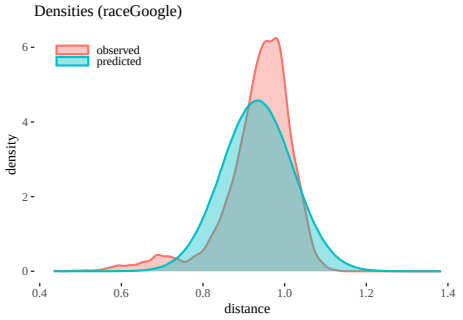
A.2. Posterior predictive checks

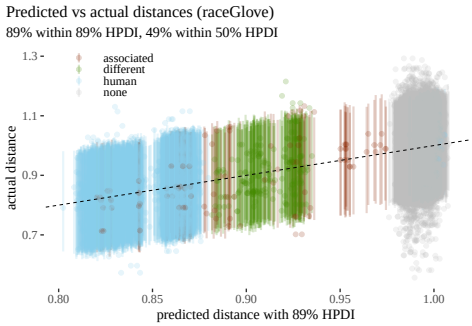


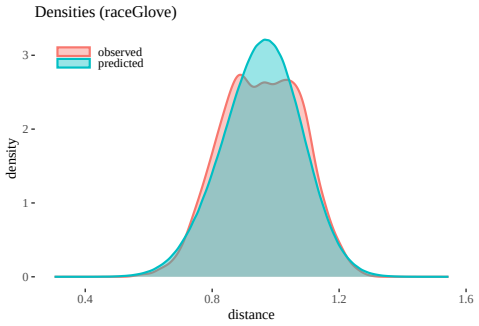

Predicted vs actual distances (raceReddit)
89% within 89% HPDI, 47% within 50% HPDI
associated
different
human
none
actual distance
predicted distance with 89% HPDI

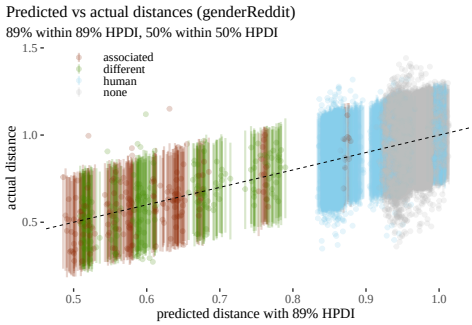

Densities (raceReddit)
observed
predicted
density
distance

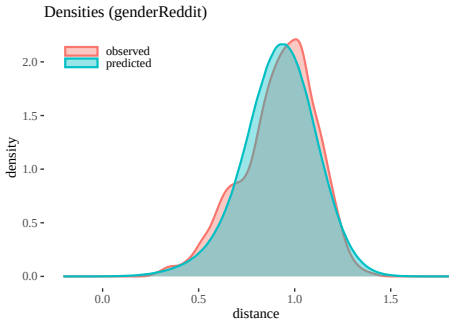

Predicted vs actual distances (raceGoogle)
92% within 89% HPDI, 59% within 50% HPDI
associated
different
human
none
actual distance
predicted distance with 89% HPDI

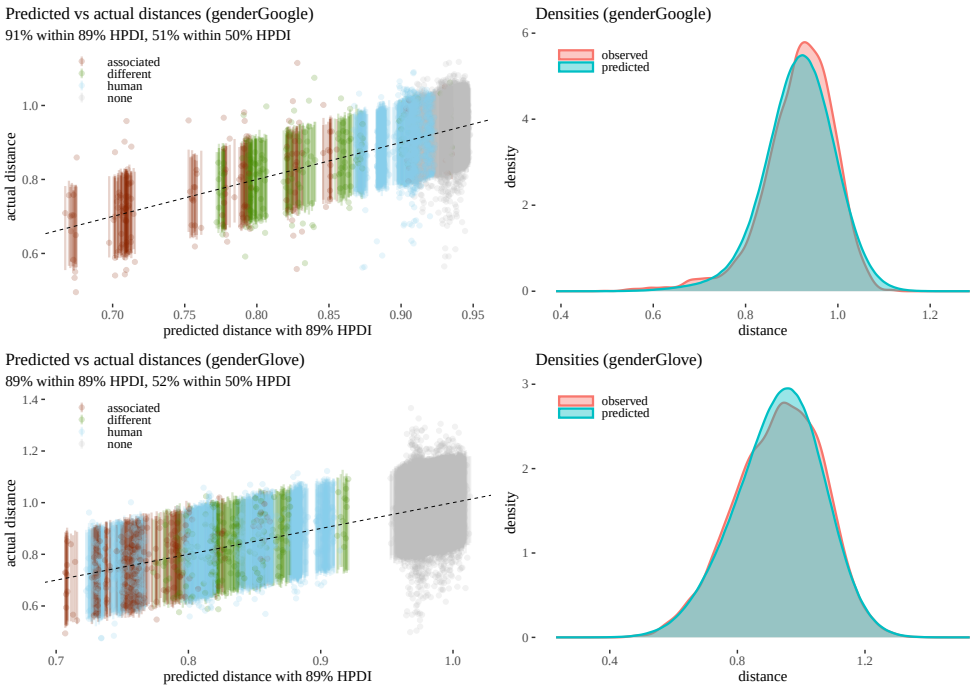

Densities (raceGoogle)
observed
predicted
density
distance


Predicted vs actual distances (raceGlove)
89% within 89% HPDI, 49% within 50% HPDI
associated
different
human
none
actual distance
predicted distance with 89% HPDI


Densities (raceGlove)
observed
predicted
density
distance


Predicted vs actual distances (genderReddit)
89% within 89% HPDI, 50% within 50% HPDI
associated
different
human
none
actual distance
predicted distance with 89% HPDI


Densities (genderReddit)
observed
predicted
density
distance



A.3. Word lists

A.3.1. Lists for [12]:

The lists are available here:

- **Gender:** https://github.com/TManzini/DebiasMulticlassWordEmbedding/blob/master/Debiasing/data/vocab/gender_attributes_optm.json
- **Race:** https://github.com/TManzini/DebiasMulticlassWordEmbedding/blob/master/Debiasing/data/vocab/race_attributes_optm.json
- **Religion:** https://github.com/TManzini/DebiasMulticlassWordEmbedding/blob/master/Debiasing/data/vocab/religion_attributes_optm.json

The lists are:

- **Gender:**

Pairs of protected attributes: [“he”, “she”], [“his”, “hers”], [“son”, “daughter”], [“father”, “mother”], [“male”, “female”], [“boy”, “girl”], [“uncle”, “aunt”]

Lists of stereotypes per gender: man: [“manager”, “executive”, “doctor”, “lawyer”, “programmer”, “scientist”, “soldier”, “supervisor”, “rancher”, “janitor”, “firefighter”, “officer”], woman: [“secretary”, “nurse”, “clerk”, “artist”, “homemaker”, “dancer”, “singer”, “librarian”, “maid”, “hairdresser”, “stylist”, “receptionist”, “counselor”] }

- **Race:**

Sets of protected attributes: [“black”, “caucasian”, “asian”], [“african”, “caucasian”, “asian”], [“black”, “white”, “asian”], [“africa”, “america”, “asia”], [“africa”, “america”, “china”], [“africa”, “europe”, “asia”]

Lists of stereotypes per race: “caucasian”: [“manager”, “executive”, “redneck”, “hillbilly”, “leader”, “farmer”], “asian” : [“doctor”, “engineer”, “laborer”, “teacher”], “black” : [“slave”, “musician”, “runner”, “criminal”, “homeless”]

- **Religion:**

Sets of protected attributes: [“judaism”, “christianity”, “islam”], [“jew”, “christian”, “muslim”], [“synagogue”, “church”, “mosque”], [“torah”, “bible”, “quran”], [“rabbi”, “priest”, “imam”]

Lists of stereotypes per race: “jew” : [“greedy”, “cheap”, “hairy”, “liberal”], “christian” : [“judgemental”, “conservative”, “familial”], “muslim” : [“violent”, “terrorist”, “dirty”, “uneducated”]

A.3.2. Our custom word lists

- **Neutral:**

['ballpark', 'glitchy', 'billy', 'dallas', 'rip', 'called', 'outlooks', 'floater', 'rattlesnake', 'exports', 'recursion', 'shortfall', 'corrected', 'solutions', 'diagnostic', 'patently', 'flops', 'approx', 'percents', 'lox', 'hamburger', 'engulfed', 'households', 'north', 'playtest', 'replayability', 'glottal', 'parable', 'gingers', 'anachronism', 'organizing', 'reach', 'shtick', 'eleventh', 'cpu', 'ranked', 'irreversibly', 'ponce', 'velociraptor', 'defects', 'puzzle', 'smasher', 'northside', 'heft', 'observation', 'rectum', 'mystical', 'telltale', 'remnants', 'inquiry', 'indisputable', 'boatload', 'lessening', 'uselessness', 'observes', 'fictitious', 'repatriation', 'duh', 'attic', 'schilling', 'charges', 'chatter', 'pad', 'smurfing', 'worthiness', 'definitive', 'neat', 'homogenized', 'lexicon', 'nationalized', 'earpiece', 'specializations', 'lapse', 'concludes', 'weaving', 'apprentices', 'fri', 'militias', 'inscriptions', 'gouda', 'lift', 'laboring', 'adaptive', 'lecture', 'hogging', 'thorne', 'fud', 'skews', 'epistles', 'tagging', 'crud', 'two', 'rebalanced', 'payroll', 'damned', 'approve', 'reason', 'formally', 'releasing', 'muddled', 'mineral', 'shied', 'capital', 'nodded', 'escrow', 'disconnecting', 'marshals', 'winamp', 'forceful', 'lowes', 'sip', 'pencils', 'stomachs', 'goff', 'cg', 'backyard', 'uprooting', 'merging', 'helpful', 'eid', 'trenchcoat', 'airlift', 'frothing', 'pulls', 'volta', 'guinness', 'viewership', 'eruption', 'peeves', 'goat', 'goofy', 'disbanding', 'relented', 'ratings', 'disputed', 'vitamins', 'singled', 'hydroxide', 'telegraphed', 'mercantile', 'headache', 'muppets', 'petal', 'arrange', 'donovan', 'scrutinized', 'spoil', 'examiner', 'ironed', 'maia', 'condensation', 'receipt', 'solider', 'tattooing', 'encoded', 'compartmentalize', 'lain', 'gov', 'printers', 'hiked', 'resentment', 'revisionism', 'tavern', 'backpacking', 'pestering', 'acknowledges', 'testimonies', 'parlance', 'hallucinate', 'speeches', 'engaging', 'solder', 'perceptive', 'microbiology', 'reconnaissance', 'garlic', 'neutrals', 'width', 'literaly', 'guild', 'despicable', 'dion', 'option', 'transistors', 'chiropractic', 'tattered', 'consolidating', 'olds', 'garmin', 'shift', 'granted', 'intramural', 'allie', 'cylinders', 'wishlist', 'crank', 'wrongly', 'workshop', 'yesterday', 'wooden', 'without', 'wheel', 'weather', 'watch', 'version', 'usually', 'twice', 'tomato', 'ticket', 'text', 'switch', 'studio', 'stick', 'soup', 'sometimes', 'signal', 'prior', 'plant', 'photo', 'path', 'park', 'near', 'menu', 'latter', 'grass', 'clock']

- **Human-related:**

['wear', 'walk', 'visitor', 'toy', 'tissue', 'throw', 'talk', 'sleep', 'eye', 'enjoy', 'blogger', 'character', 'candidate', 'breakfast', 'supper', 'dinner', 'eat', 'drink', "carry", "run", "cast", "ask", "awake", "ear", "nose", "lunch", "coalition", "policies", "restaurant", "stood", "assumed", "attend", "swimming", "trip", "door", "determine", "gets", "leg", "arrival", "translated", "eyes", "step", "whilst", "translation", "practices", "measure", "storage", "window", "journey", "interested", "tries", "suggests", "allied", "cinema", "finding", "restoration", "expression","visitors", "tell", "visiting", "appointment", "adults", "bringing", "camera", "deaths", "filmed", "annually", "plane", "speak", "meetings", "arm", "speaking", "touring", "weekend", "accept", "describe", "everyone", "ready", "recovered", "birthday", "seeing", "steps", "indicate", "anyone", "youtube"]

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR* abs/1607.06520, (2016). Retrieved from http://arxiv.org/abs/1607.06520

[2] Kawin Ethayarajh. 2020. Is your classifier actually biased? Measuring fairness under uncertainty with bernstein bounds. *CoRR* abs/2004.12332, (2020). Retrieved from https://arxiv.org/abs/2004.12332

[3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, (November 2017). DOI:https://doi.org/10.1073/pnas.1720347115

[4] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (April 2018), E3635–E3644. DOI:https://doi.org/10.1073/pnas.1720347115

[5] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. DOI:https://doi.org/10.18653/v1/N19-1061

[6] Jonathan Gordon and Benjamin Durme. 2013. Reporting bias and knowledge acquisition. In *AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, Co-located with CIKM 2013*, 25–30. DOI:https://doi.org/10.1145/2509558.2509563

[7] Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21, 5 (October 2014), 1157–1164. DOI:https://doi.org/10.3758/s13423-013-0572-3

[8] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR* abs/1608.07187, (2016). Retrieved from http://arxiv.org/abs/1608.07187

[9] Gabbrielle Johnson. forthcoming. Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal Moral Philosophy* (forthcoming).

[10] John Kruschke. 2015. *Doing bayesian data analysis (second edition)*. Academic Press, Boston.

[11] Anne Lauscher and Goran Glavas. 2019. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR* abs/1904.11783, (2019). Retrieved from http://arxiv.org/abs/1904.11783

[12] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. Retrieved from https://arxiv.org/abs/1904.04047

[13] Richard McElreath. 2020. *Statistical rethinking: A bayesian course with examples in r and stan, 2nd edition* (2nd ed.). CRC Press. Retrieved from http://xcelab.net/rm/statistical-rethinking/

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. DOI:https://doi.org/10.48550/ARXIV.1301.3781

[15] Richard Morey, Rink Hoekstra, Jeffrey Rouder, Michael Lee, and EJ Wagenmakers. 2015. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* (September 2015).

[16] Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics* 46, 2 (June 2020), 487–497. DOI:https://doi.org/10.1162/coli_a_00379

[17] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 101–115. DOI:https://doi.org/10.1037/1089-2699.6.1.101