Exploring Human Gender Stereotypes with Word Association Test

Yupei Du , Yuanbin Wu , and Man Lan

Department of Computer Science and Technology,
East China Normal University
{yupeidu.cs}@gmail.com {ybwu,mlan}@cs.ecnu.edu.cn

Abstract

Word embeddings have been widely used to study gender stereotypes in texts. One key problem regarding existing bias scores is to evaluate their validities: do they really reflect true bias levels? For a small set of words (e.g. occupations), we can rely on human annotations or external data. However, for most words, evaluating the correctness of them is still an open problem. In this work, we utilize word association test, which contains rich types of word connections annotated by human participants, to explore how gender stereotypes spread within our minds. Specifically, we use random walk on word association graph to derive bias scores for a large amount of words. Experiments show that these bias scores correlate well with bias in the real world. More importantly, comparing with word-embedding-based bias scores, it provides a different perspective on gender stereotypes in words.[1]

1 Introduction

Consciously or not, gender stereotypes can be found in many places in our life. From proportions of female in certain occupations, to students' choices on their majors, from the stories people been told in books and movies, to the decisions they make every day, researchers try to build different stereotype models, to understand patterns behind various gender-related social and psychological phenomena.

Here, we investigate gender bias using the lens of language, especially, the words. Recent studies have shown that, by learning from large amounts of texts, low dimensional word embeddings (Mikolov et al., 2013) can exhibit human gender bias: the vector of "engineer" is more close to "he" than "she" (Bolukbasi et al., 2016). What's more insightful is that the revealed bias correlates well with census data, and its trends on historical texts matches women's movements in 1960-1970s (Garg et al., 2017). Thus word embeddings might be a reasonable tool in the study of gender stereotypes.

In this work, we try to resolve two problems regarding existing word-embedding-based bias scores. First, most embeddings are trained on word co-occurrence information in texts: "engineer" co-occurs with "he" more often than "she" in corpus. Since co-occurrence is only one of the various relations among words, it is natural to ask whether other types of relations can also reflect gender bias. Second, the validity of word-embedding-based bias scores are often examined using a small set of occupation words (i.e., how they correlate with occupation statistics in the real world) or a small set of human annotated words. How to evaluate validities of bias scores on the remaining large amount of words is still an open problem.

In order to answer these questions, we depart from existing models and explore a new tool for quantifying gender bias of words. Our method utilizes a recent result of large-scale word association test (De Deyne et al., 2018). In this test, participants are given a *cue* word and asked to write down most related words (*responses*) in their minds. Comparing with the co-occurrence relation, which subjects to contexts and topics of a document, word association test bears a more direct interaction with human brains. Therefore, it may provide a more detailed profile on human gender stereotypes.

We build a word association graph based on records of cue-response pairs. The vertices represent words, and an edge indicates whether two words are associated in some tests. With this

[1] Our code is publicly available at https://github.com/Yupei-Du/bias-in-wat.

graph, we can study how gender bias propagates among words using various graph-based algorithms. For example, a random walk starting from "he" and "she" can be applied to model the stochastic behaviors in real association tests, and the probabilities of reaching a word can be seen as its distances to "he" and "she". One advantage of graph-based models is that the derived bias scores are characterized by the intrinsic structure of the word association graph. Therefore, if we roughly assume the word association graph is an externalization of concept networks in human brains, as properties of the graph, the bias scores can be seen as a benchmark to help examining validities of other models.

In empirical evaluations, we compare the word-embedding-based bias scores with our word-association-based bias scores. First, regarding the real world census data and human annotations, word-association-based scores show stronger correlations than vanilla word-embedding-based scores. Second, as embeddings could also be applied for building graphs, we investigate such graphs and look into their differences with the word association graph. It turns out that both kinds of graphs exhibit "small world" properties, but they have different hub words and connecting mechanisms. We then run the same stereotype propagation algorithm and compare the derived gender bias scores. Third, as a case study, we use word-association-based scores as a benchmark to see the effectiveness of existing de-bias algorithms on word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018). The results suggest that it is hard for both de-bias algorithms to remove gender stereotypes in word embeddings completely, which matches the conclusion in (Gonen and Goldberg, 2019).

2 Word Association Test

Word association test is a simple and sometimes entertaining game in which participants are asked to respond with the first several words that come out in their mind (the response) after being presented with a word (the cue)[2]. Table 1 lists some examples of the test. It is considered to be one of the most straightforward approaches for gaining insight into our semantic knowledge (Steyvers and Tenenbaum, 2005), and is also a common approach to measure words' meanings in our mind (so called mental lexicons (Jackendoff and Jackendoff, 2002)). For studies of gender stereotypes, word association test provides a way to inspect how a gender-specific word (e.g., "he", "she") connecting with other words, and the patterns of the connections may help us to probe gender stereotypes in our brains.

Cue	R1	R2	R3
way	path	via	method
extra	plus	special	additional
i	you	me	eye
come	go	closer	on
than	then	there	though
son	daughter	sun	boy
mind	brain	cognition	thinking

Table 1: Examples of word association test records in SWOWEN datasets (De Deyne et al., 2018). R1, R2, R3 are participants' responses.

Several collections of English word association test records are publicly available (Nelson et al., 2004; Moss et al., 1996; Kiss et al., 1973). In this work, we focus on SWOWEN (De Deyne et al., 2018), which is currently the largest English word association database. The dataset is collected from 2011 to 2018, and contains more than 12000 cues and responses from over 90000 participants. The collection process begins with a small number of words as a word pool. Every time a participant starts a new test, a cue word is sampled randomly from the word pool. Once a new word not in the word pool is associated, it is added into the pool. This snowball sampling method helps to include both frequent and less-frequent cues at the same time.

3 Word Association Graph

Based on a table of word association test results, we try to figure out a complete picture on how gender biases are embedded in the test. We build a word association graph $G = (V, E)$ whose vertices V are cues and responses, edges E encode associations in the test, and weights of edges represent the strength of associations. Specifically, we take the following steps.

Pre-Processing Our pre-processing steps are mostly based on De Deyne et al. (2018)[3]. These

[2] https://en.wikipedia.org/wiki/Word_Association

[3] Raw word association test results and pre-process scripts are available in https://github.com/SimonDeDeyne/SWOWEN-2018. We re-implement

procedures consist of spell-check, unifying cue forms, removing untrustworthy participants, and Americanize all non-American spellings. Through these steps, we reduce word association records from 88722 participants, 12292 cues, and 4069086 responses to 83863 participants, 12217 cues, and 3665100 responses.

Counting Association Pairs We set all preprocessed cues and responses as vertices. An edge is added if it corresponds to a cue-response pair (directions of pairs are ignored for simplicity), and we count the occurrence numbers for edges as their weights. The final graph G contains 12217 nodes and 1283047 edges.

Next, we will study how the gender bias propagates in G. More concretely, we investigate how gender information transit from a set of *gender-specific words* (i.e., words related with a gender by definition, such as "he", "she", "man", "woman") to other words. We obtain bias scores for *every* word based on the propagation and the graph structure, then we use them as a benchmark for evaluating validities of other models.

4 Stereotype Propagation

In the literature of psychology, Dell (1986) proposes a mental process in which words can spread its meaning by word associations. This theory suggests that once a gender-specific word in word association graph is activated, its influence will propagate along graph edges and form gender stereotypes. If a gender-neutral word (e.g. *nurse*) has stronger connections with gender-specific words in one gender (e.g. feminine words) than the other one, this gender-neutral word will receive imbalanced gender information, leading to unnecessary gender tendency. We thus tend to relate *nurse* more tightly with a woman than a man.

To characterize stereotypes of words, we simulate this mental process by spreading gender information from a set of commonly used gender-specific words $\mathcal{L}$ to the whole graph, using a random walk approach (Zhou et al., 2003). After enough iterations, gender information received by each word will eventually converge. Then we can calculate bias scores by seeing how much gender information they have received. We call this method *stereotype propagation*.

First, we build the set of gender-specific words

their R codes with Python.

he / she	brother / sister
father / mother	man / woman
son / daughter	boy / girl
husband / wife	uncle / aunt
grandfather / grandmother	gentleman / lady

Table 2: Gender-specific word pairs in $\mathcal{L}$.

$\mathcal{L}$. In order to balance influence from different genders and avoid their heterogeneity, we collect them pair-by-pair. Specifically, $\mathcal{L} = \{(w_m^k, w_f^k)\}_{k=1}^{|\mathcal{L}|}$, where w_m^k (a *masculine word*) and w_f^k (a *feminine word*) are semantically related but different in gender. Table 2 is the $\mathcal{L}$ we use by default.

For each word, we use a 2-dimension vector (b_m, b_f) to record the amount of gender information they will receive from masculine words (b_m) and feminine words (b_f). Let $P \in \mathbb{R}^{|V|\times 2}$ be the matrix containing vectors of all words, and P_0 be the initial state of P. For the gender-specific words in $\mathcal{L}$, we set their vectors in P_0 with $(1, 0)$ (masculine words) and $(0, 1)$ (feminine words). Vectors of other words are initialized with $(0, 0)$.

Next, following (Zhou et al., 2003), we iteratively update all words' masculine and feminine information until reaching a stationary P^*,

$$P_{t+1} = \alpha T P_t + (1-\alpha) P_0, \tag{1}$$

where $T \in \mathbb{R}^{|V|\times|V|}$ is a normalization of G's adjacent matrix S (entries are edge weights),

$$T = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}, \;\; D = \mathrm{diag}_i\Big(\sum_{j=1}^{N} S_{ij}\Big),$$

and α is a hyper parameter in range (0, 1), controlling the balance between *global* and *local* consistency. When α is small, we will pay more attention to keep the initial vectors of words (i.e., global consistency). When α gets larger, we will emphasize the smoothness between connected words: close words should have similar labels (i.e., local consistency).

We have a closed-form solution of Equation 1

$$P^* = (1-\alpha)(I - \alpha T)^{-1} P_0. \tag{2}$$

It is worth noting that the solution only depends on the graph structure (T), gender-specific words (P_0), and a hyper parameter α. We think that by minimizing the influence of parameters (only P_0

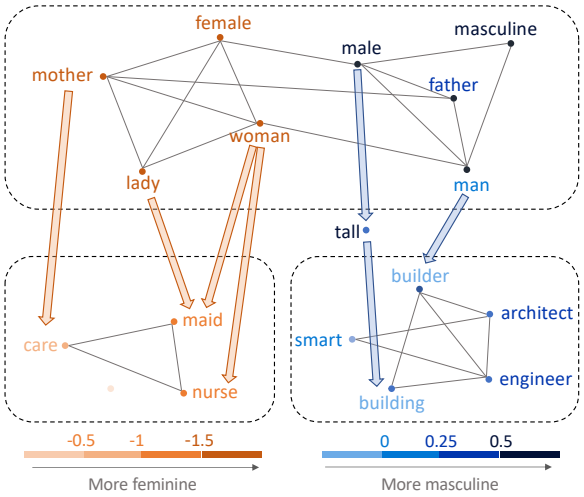

Figure 1: A snapshot on how stereotypes propagate in word association graph. Darker color on nodes means higher bias (visualize b). Colored arrows indicate how gender information propagate.

and α), vectors in P^* could be more robustly tied to the intrinsic properties of the word association graph and thus can be applied to (in)validate other bias scores.

Finally, we define a word's bias score b, from vectors in P^*,

$$b = \log \frac{\bar{b}_m}{\bar{b}_f},$$

where

$$\bar{b}_m = \frac{e^{b_m}}{e^{b_m} + e^{b_f}}, \quad \bar{b}_f = \frac{e^{b_f}}{e^{b_m} + e^{b_f}}.$$

We snapshot a sub-graph of G with P^* in Figure 1 to show how stereotype propagates. We first observe that words in the association graph are clustered. Those clusters are connected by some hubs (i.e., high degree words), through which gender information is able to spread quickly. As a result, words in a cluster usually have a same gender polarity. In Section 5.4, we will further show that G enjoys some key small-word properties. Thus, inherently, people may be "lazy" at carefully bounding gender biases. Another finding in Figure 1 is that "building" bears a male bias since both buildings and men are "tall". This kind of connections are hard to observe in co-occurrences (word embeddings). We will also discuss details in Section 5.4.

5 Empirical Evaluation and Application

In this section, we first evaluate the word-association-based bias scores by examining whether it correlates to other real-world data. Then, we perform various controlled comparisons on settings of the stereotype propagation, including different graph structures, gender specific word sets $\mathcal{L}$ and propagation algorithms. Finally, as a case study, we use our proposed scores as a benchmark to evaluate the performances of two recent de-bias methods for word embeddings.

5.1 Alignments with Real-world Data

In this section, we perform two experiments to validate the feasibility and reliability of studying gender stereotypes using the word association test.

In the first experiment, we examine whether the bias detected from word association test can correlate with gender stereotypes in the real world. We adopt two data sources here, namely the gender proportion of occupations in U.S. census data (*census data* in Table 3) [4] (Ruggles et al., 2015), and manually annotated professions stereotypes by U.S. based crowd workers (*human judgments* in Table 3) (Bolukbasi et al., 2016). We measure the strengths of correlations by Pearson's r[5] and significance p in Table 3.

Our second experiments compare average bias scores of words between different concepts using *Implicit Association Test* from (Caliskan et al., 2017). We adopt three pairs of concepts, namely *career vs. family*, *math vs. arts*, and *science vs. arts*[6]. Words in each pair are thought to have different gender tendencies. For example, we often relate *career* with a man but relate *family* with a woman. Words we use are shown in Table 4. For each pair, we report Cohen's d [7] and significance p.

For comparison, we also conduct a similar analysis for word-embedding-based stereotype scores. We take two widely used settings, namely *we-cos* (Bolukbasi et al., 2016) and *we-norm* (Garg et al.,

[4]Follow Levanon et al. (2009), we choose the log proportions, log-prop(p) = $\log \frac{p}{1-p}$, where p = percentage of women in an occupation, similar with the way we calculate words' stereotype scores.

[5]Pearson's r is a measure of the linear correlation between two variables. It has a value between 1 and -1, where 1 means total positive linear correlation, 0 means no linear correlation, and -1 means total negative linear correlation.

[6]Three words are not included in our word association graph, we thus substitute them by similar words, they are *cousins* to *cousin*, *equations* to *equation* and *computation* to *compute*.

[7]Cohen's d is an effect size used to indicate the standardized difference between two means. Conventional small, medium, and large values of d are 0.2, 0.5, and 0.8 respectively.

Method	Correlation Analysis				Implicit Association Test					
	Census data		Human Judgments		career vs. family		math vs. arts		science vs. arts	
	r	p	r	p	d	p	d	p	d	p
we-cos	0.58	10^{-7}	0.57	$< 10^{-10}$	**1.42**	.00	**1.30**	.01	1.37	.01
we-norm	**0.59**	10^{-7}	0.57	$< 10^{-10}$	1.18	.02	**1.30**	.01	1.37	.01
G	0.58	10^{-7}	**0.66**	$< 10^{-10}$	0.89	.10	0.98	.06	**1.55**	.00
G^e_{global}	0.04	.60	-0.01	.85	-0.20	.60	-0.08	.88	-0.31	.51
G^e_{local}	-0.08	.49	0.03	.72	-0.07	.87	-0.13	.74	-0.01	.98

Table 3: Validation of stereotype scores. r is the Pearson's r in correlation analysis, d is the effect size, and p is the significance. G, G^e_{global}, G^e_{local} represent stereotype propagation on different graphs. α is set to 0.99 by default.

concept	**words**
career	executive, professional, corporation, salary, office, business, career
family	home, parents, children, family, cousin, marriage, wedding, relatives
maths	math, algebra, geometry, calculus, equation, compute, numbers, addition
arts	poetry, art, dance, literature, novel, symphony, drama
science	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy

Table 4: Words in each concept.

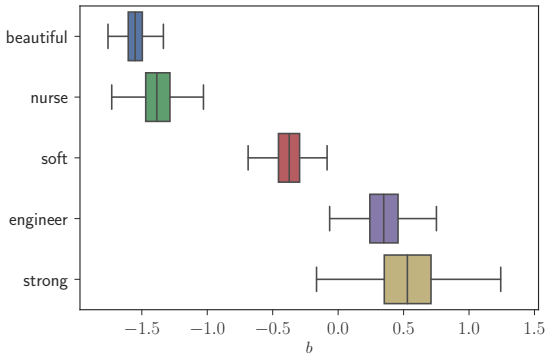

Figure 2: Examples of words' bias scores with confidence intervals (95%).

2017). They are both obtained by comparing a word w's embedding similarities with masculine and feminine words in gender word pairs $\mathcal{L}$.

The bias scores in "we-cos" are defined as [8],

$$b = \frac{1}{|\mathcal{L}|}\sum_{k=1}^{|\mathcal{L}|}\Big(\frac{w^\intercal w_m^k}{||w||\cdot||w_m^k||} - \frac{w^\intercal w_f^k}{||w||\cdot||w_f^k||}\Big), \quad (3)$$

and in "we-norm", the scores are

$$b = -\frac{1}{|\mathcal{L}|}\sum_{k=1}^{|\mathcal{L}|}\big(||w - w_m^k||^2 - ||w - w_f^k||^2\big), \quad (4)$$

where w_m, w_f is the embeddings of masculine and feminine words in $\mathcal{L}$. We re-train *word2vec* embeddings (Mikolov et al., 2013) on a Wikipedia dump.

Table 3 lists the correlation and implicit association test results. We find that

- Word-association-based bias scores have strong performances in both experiments. In correlation analysis, they exhibit similar or stronger relationships with real-world bias comparing to word-embedding-based bias scores. In implicit association test, our method still show large effects (Cohen's $d > 0.8$). It means that word association test is reliable in detecting human gender stereotypes.

- More importantly, although both word-embedding-based and word-association-based bias scores align well with real-world bias, gender stereotypes discovered from texts and human brains are different. The key evidence is that the word-association-based bias scores align much better on human annotated words.[9] Therefore, it is valuable and helpful to introduce a different perspective for lexical-level bias analyses.

5.2 The Influence of $\mathcal{L}$

To test the robustness of the proposed scores, we are interested in how $\mathcal{L}$ influences the results. We apply a bootstrap sampling strategy on the entire $\mathcal{L}$. Specifically, we sample 8 gender word pairs each time from $\mathcal{L}$. For all 45 possible combinations of the 8 gender words, we compute their bias scores to get confidence intervals w.r.t. to the full $\mathcal{L}$. We assume bias scores of a word to follow Gaussian distribution for easy calculation of confidence interval. We set the confidence level at 95%. Figure 2 illustrates some examples of words' bias scores with confidence intervals. We further

[8] We reuse symbol w as word w's embedding vector.

[9] Significantly higher correlation with human judgments than word-embedding-based methods (standard score Z=2.05).

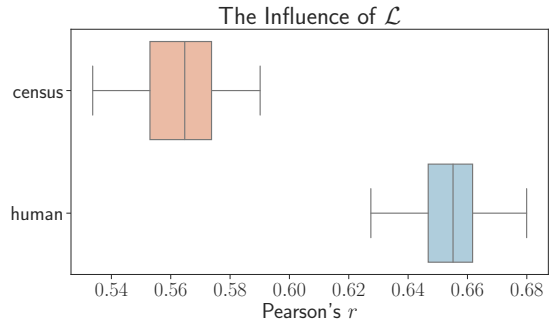

Figure 3: Confidence intervals (95%) of correlations between stereotype propagation results and real-world bias. Census stands for the census data dataset and human stands for the human judgments dataset.

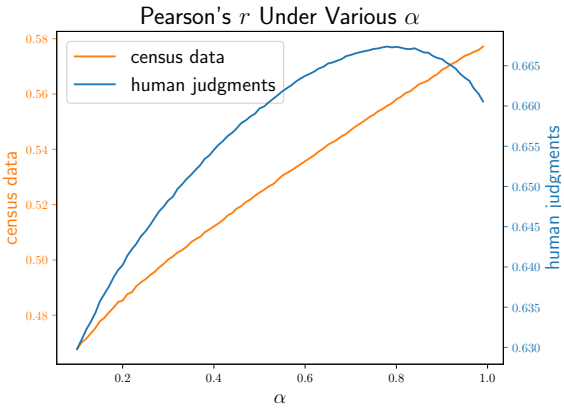

Figure 4: Correlation analysis under various α.

calculate the confidence intervals of correlations between our results and gender stereotypes in the real world in Figure 3. The results show that the words' bias scores are stable with respect to reasonable settings of $\mathcal{L}$.

5.3 Variants of Stereotype Propagation

There are many variants of the random walk method in previous literature (Zhou et al., 2003; Zhu et al., 2003; Velikovich et al., 2010; Vicente et al., 2017). We experiment with another approach in (Zhu et al., 2003), to test whether bias scores of words are insensitive towards random walk algorithms. In this method, we consider only local consistency (without the second term of Equation 1). As a result, we find high consistency between their bias scores (Pearson's $r = 0.789$, $p = 0.0$).

We also examine the influence of hyper parameter α. Here, we conduct similar correlation analysis in Section 5.1 under various α (Figure 4). We find that stereotype propagation has a stable performance regarding different α, while a proper α (close to 1, allowing information to spread more freely) do helps.

5.4 Comparing with Word Embedding Graphs

Word embeddings can also be applied to build graphs on words: edges and weights can be defined based on similarities among word vectors. Previous work Hamilton et al. (2016) has adopted these kind of approaches to study the propagation of words' sentiment. Therefore, it is natural to compare the graphs built from word embeddings with the word association graph here.

5.4.1 Word Embedding Graphs

We take two strategies, namely local and global, to construct two word embedding graphs G^e_{local} and G^e_{global}. The local strategy is to connect a word with its k nearest neighbors (same as Hamilton et al. (2016)). The global strategy is to find e most similar word pairs in all possible word pairs as graph edges. To ensure that information will not be interrupted during propagation, we rectify both $G^e_{local}, G^e_{global}$ to their largest connected components [10]. To make a fair comparison, we make the two word embedding graphs have the same average node degree with the word association graph (105), hence, we set $k = 105$ and $e = 1235178$. We use the same word2vec embeddings in Section 5.

We run stereotype propagation on both embedding based graphs. As reported in Figure 3, unfortunately, both of them can hardly capture gender bias of words. To understand the reasons, we further compare the two types of graphs in detail.

5.4.2 Graph Properties

We summarize some descriptive statistics of the three graphs in Table 5, where L is the average shortest path length between every possible word pair. D is the longest shortest path between two words in the graph. C is the clustering coefficient, which is the average proportion of two arbitrary neighbors of a random word being themselves neighbors. When $C = 0$, no words have neighbors that are also each others' neighbors and when $C = 1$, the graph should be a fully connected graph. It measures how tightly words are

[10] We won't have the same problem in the word association graph, because a word will be added only when it is associated from another word.

	Definition	G^e_{local}	G^e_{global}	G
$\|V\|$	number of words	11870	11758	12217
K	average in degree	104.69	105.05	105.02
L	average shortest path length	2.89	3.43	2.12
D	diameter of the graph	5	10	3
C	clustering coefficient	0.27	0.40	0.24

Table 5: Descriptive statistics of graphs. G is the word association graph.

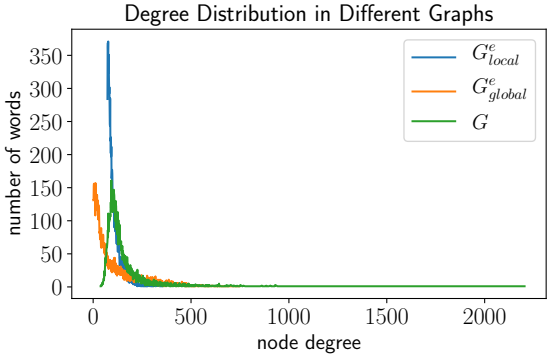

Figure 5: Distributions of node degrees.

organized as clusters.

Small world graphs are defined to be graphs where the distance between two arbitrary nodes follows a logarithmic distribution ($L \propto \log|V|$). In these graphs, node degree distributions follow power laws (Jeong et al., 2000), which means the degree of most nodes' are small, while a few nodes (*hubs*) have much larger degree number. Most nodes in the graph are organized as clusters connected by hubs, resulting in a relatively large clustering coefficient. These attributes work together to accelerate information flow within graphs.

From Table 5, we find that all three graphs exhibit small world properties. First, all three graphs have a quite small L and D ($L = 2.89, D = 5$, $L = 3.43, D = 10$, and $L = 2.12, D = 3$ for local, global, and word association graph respectively, while $\log 10000 = 9.21$). Second, their clustering coefficient C are relatively large (> 0.2), while that of a typical random network is 10^{-3} (Steyvers and Tenenbaum, 2005). Third, as shown in in Figure 5, the distributions of node degrees have shapes of power laws (e.g., long tails). It's worth noting that in word association graph, words with an extremely small number (< 10) of degrees are rare. We think that one reason might be the strategy token in data collection. The word association test records are collected using a snowball sampling method (Section 2), and words with few degrees are new comers in the test, which have not been sampled as cue words for enough times. Another difference is that the curve of word association graph has a significant longer tail which may imply that it has more powerful hubs. This feature allows gender information to spread more quickly (L is smaller) than the word embedding based graphs.

5.4.3 Types of Graph Edge

We also give a finer analysis on the two types of graphs to show their different ways on connecting gender specific words and other words. We examine *gender bias paths* which are all shortest paths [11] connecting words in $\mathcal{L}$ and other words in all three graphs.

According to semantic network theory (Collins and Loftus, 1975), there are three types of connections in human word association, namely *semantic priming* (Traxler, 2011), *neighborhood effects* and *frequency effects* (Andrews, 1989). Later, *co-occurrence* relation has also been found to hold a correlation with word association tests (Spence and Owens, 1990).

Our main observation is that, regarding the gender bias paths, we can find all of these connection types in bias paths of the word association graph. However, neighborhood effects and frequency effects are missed in word embedding graphs. This observation reveals that word association graph has a richer variety of connections, which may help the stereotype propagation. Examples of different connection types in both kind of graphs are exhibited in Table 6. We discuss each edge types in the following.

Semantic priming refers to the effect that word associations can happen between semantically related words. This is observed in bias paths on both word association graph and word embedding graphs, like *potential-maybe* in word association graph and *really-actually* in word embedding graphs. A more profound semantic relationship is hierarchical, including generic to specific and partitive relationships, such as *energy-carbohydrate* and *girl-person*.

Neighborhood effects refer to word associations between words that are highly confusable due to overlapping features, like spelling and pro-

[11] If several paths are equally short, we pick the one with highest cumulative weight.

Type	Word Association Graph	Word Embedding Graph
semantic priming	boy - *energy* - *carbohydrate* girl - *potential* - *maybe*	boy - kid - *really* - *actually* *girl* - *person*
neighborhood effects	*son* - *some* she - *pronoun* - *pro* - *projection*	X
frequency effects	husband - married - *a* - few girl - lady - *the* - article	X
co-occurrence	father - *respect* - *others* mother - *birth* - *day*	he - *even* - *more* she - *cook* - *pickle*

Table 6: Examples of different connection types between words. These words are shown in italics. Blue words refer to masculine word while red words refer to feminine word.

nunciation. Representative examples include *pronoun-pro-programmer* (spelling) and *son* to *some* (pronunciation). We manually check more than 100 bias paths on word embedding graphs, and didn't find this kind of connections, indicating its inability in modeling such effects.

Frequency effects suggest that human tends to associate commonly used words with higher possibility, making hubs in word association graph more likely to be frequent words. We pick top 2% words with largest node degrees as the representatives of hub words, and count the average occurrence of them in Wikipedia. As a result, the hubs of word association graph (297868) appears much more frequently than hubs in word embedding graphs (12931 of G^e_{local}, 3373 of G^e_{local}). It implies a stronger frequency effects in the word association graph.

To investigate how frequency effects influence stereotype propagation, we check the composition of these hub words. It turns out that most gender-specific words (except for *he*, *she*, *daughter*, *aunt*) in $\mathcal{L}$ are contained in the hub words of the word association graph, while none of them lie within those of the embedding-based graphs. Therefore, there is no wonder that $\mathcal{L}$ in the word association graph could distribute its gender information to other words in the graph more efficiently, leading to better performance in detecting gender bias of words.

Co-occurrence Due to the fact that most commonly used word embedding models are trained on word co-occurrence (Mikolov et al., 2013; Pennington et al., 2014), we observe lots of connections with such type in bias paths on word embedding graphs, which also widely exists in those on the word association graph, such as *birth-day* and *even-more*.

To summarize, frequency effects, neighborhood effects, and more concentrated hubs may work together to help spread gender stereotypes. It shows the value of introducing word association test into gender stereotype research.

5.5 Case Study: Do De-bias Methods Really Remove Gender Bias?

Some works have proposed approaches to reduce gender stereotypes in word embeddings. They can mainly be divided into two categories: post-processing step (Bolukbasi et al., 2016) and training goal modification (Zhao et al., 2018). Recent work (Gonen and Goldberg, 2019) has argued that those de-bias procedures only cover up but do not remove stereotypes. In this section, we utilize the bias scores from the word association graph to examine whether those de-bias methods work.

We experiment with de-bias methods of both categories, *Hard-Debias* refers to the method in (Bolukbasi et al., 2016) and *Gn-Glove* refers to the method in (Zhao et al., 2018). For each method, we compare it with an unde-biased version. For Hard-Debias, we compare with embeddings before these procedures [12] and for Gn-Glove, we compare with the original Glove (Pennington et al., 2014) offered in its project homepage [13].

For stereotype in word embeddings, we use the same measurements in Section 5, we-cos and we-norm. We examine whether those word embeddings still contain human-like stereotypes by con-

[12] We use the same word2vec embeddings in Section 5.

[13] https://github.com/uclanlp/gn_glove

		Original		De-biased	
		r	p	r	p
Hard-Debias	we-norm	0.36	.00	0.06	10^{-9}
	we-cos	0.35	.00	0.05	10^{-8}
Gn-Glove	we-norm	0.38	.00	0.34	$< 10^{-10}$
	we-cos	0.38	.00	0.34	$< 10^{-10}$

Table 7: Correlation analysis between word association bias scores and bias scores from de-biased word embeddings. r is Pearson's r and p is the significant value.

ducting a correlation analysis between their word-embedding-based bias scores and our word association bias scores. We expected that word association bias scores would have proper alignments with original word embeddings but poor alignments with de-biased versions if those de-bias procedures have done their jobs.

After removing a set of gender-specific words whose gender information are not removed in Hard-Debias, we extract the words appears both in our word association graph and the word embeddings vocabulary, 11311 words in total, as the word list for this case study.

Experiment results are shown in Table 7. We have three main observations:

- Both methods work to some extent, since Pearson's rs of both de-biased versions decrease compared with the original version.
- Neither Hard-Debias nor Gn-Glove removes gender stereotypes in word embeddings entirely, because bias scores derived from de-biased embeddings still exhibit significant correlations with the word association bias scores.
- Hard-Debias outperforms Gn-Glove. As we can see, Gn-Glove can hardly remove gender bias, which matches the conclusion in (Gonen and Goldberg, 2019).

6 Related Work

Studying gender stereotypes by language analysis (Hamilton and Trolier, 1986; Basow, 1992; Wetherell and Potter, 1993; Holmes and Meyerhoff, 2008; Coates, 2015) is now an import research topic. Although traditional methods like survey (Williams and Best, 1990) are quite effective, they are expensive and time-consuming. Therefore, Garg et al. (2017) use word embeddings as a quantitative lens to investigate historical trends of gender stereotypes. However, due to the limitations in word embedding training, existing methods have constrained effects in this field.

Word association test, as a product of psychology research, has been proven reliable on studying human implicit memory (Cañs, 1990; Playfoot et al., 2016). Different from word embeddings, word association test not only have a correlation with the co-occurrence of words (Spence and Owens, 1990) but also can reflect richer relationships like hierarchical relationships (Nuopponen, 2014), making it an excellent material in studying human stereotypes.

7 Conclusion

We utilize a recent large-scale word association test to explore how gender stereotypes propagate within our mind by constructing a word association graph from it, and derive gender bias scores of words. Experiments suggest that our approach is effective in detecting human stereotypes, and is tied robustly to graph structure. Therefore, these bias scores could be used to validate other methods. What's more, our results indicate that gender bias learned from large-scale texts is different from that within our minds, introducing a new perspective for lexical-level stereotype-related research.

Acknowledgments

The authors wish to thank the reviewers for their helpful comments and suggestions, Qi Zheng, Tao Ji, Yuekun Yao, Changzhi Sun and Xinyue Chen for their useful comments on writing, Hoiyan Mak, Qing Cai, Bing Li, and Yang Yang for contributive discussions on psychology-related topics. This research is (partially) supported by STCSM (18ZR1411500) and the Fundamental Research Funds for the Central Universities. The corresponding author is Yuanbin Wu.

References

Sally Andrews. 1989. Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):802.

Susan A Basow. 1992. *Gender: Stereotypes and roles*. Thomson Brooks/Cole Publishing Co.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

José J Cañs. 1990. Associative strength effects in the lexical decision task. *The Quarterly Journal of Experimental Psychology*, 42(1):121–145.

Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, pages 1–20.

Gary S Dell. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *CoRR*, abs/1711.08412.

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. *CoRR*, abs/1903.03862.

David L Hamilton and Tina K Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 595–605.

Janet Holmes and Miriam Meyerhoff. 2008. *The handbook of language and gender*, volume 25. John Wiley & Sons.

Ray Jackendoff and Ray S Jackendoff. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. 2000. The large-scale organization of metabolic networks. *Nature*, 407(6804):651.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165.

Asaf Levanon, Paula England, and Paul Allison. 2009. Occupational feminization and pay: Assessing causal dynamics using 1950–2000 us census data. *Social Forces*, 88(2):865–891.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Helen Moss, Lianne Older, and Lianne JE Older. 1996. *Birkbeck word association norms*. Psychology Press.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Anita Nuopponen. 2014. Tangled web of concept relations. concept relations for iso 1087-1 and iso 704. In *Terminology and Knowledge Engineering 2014*, pages 10–p.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

David Playfoot, Teodor Balint, Vibhuti Pandya, Averil Parkes, Mollie Peters, and Samantha Richards. 2016. Are word association responses really the first words that come to mind? *Applied Linguistics*, 39(5):607–624.

Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. Integrated public use microdata series: Version 6.0 [dataset]. minneapolis: University of minnesota.

Donald P Spence and Kimberly C Owens. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5):317–330.

Mark Steyvers and Joshua B Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Matthew J Traxler. 2011. *Introduction to psycholinguistics: Understanding language science*. John Wiley & Sons.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan T. McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 777–785.

Iñaki San Vicente, Rodrigo Agerri, and German Rigau. 2017. Q-wordnet PPV: simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *CoRR*, abs/1702.01711.

Margaret Wetherell and Jonathan Potter. 1993. *Mapping the language of racism: Discourse and the legitimation of exploitation*. Columbia University Press.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev*. Sage Publications, Inc.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4847–4853.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 321–328.

Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 912–919.