Appendix A Data

This section contains information on the additional embeddings used in this supplement and the various word lists used. These lists will also be provided in a public repository alongside code to reproduce our results.

A.1 Additional Embeddings

A.1.1 Wikipedia 2014+Gigaword 5 GloVe

Trained on Wikipedia data from 2014 and newswire data from the mid 1990s through 2011 [1] using GloVe [2]. Available online at `https://nlp.stanford.edu/projects/glove/`. We refer to these vectors as the Wikipedia vectors.

A.1.2 Common Crawl GloVe

Trained on Common Crawl using GloVe [2]. Also available online at `https://nlp.stanford.edu/projects/glove/`. We refer to these vectors as the Commmon Crawl vectors.

A.2 Group words

Man words *he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews*

Woman words *she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, femen, sisters, aunt, aunts, niece, nieces*

White last names *harris, nelson, robinson, thompson, moore, wright, anderson, clark, jackson, taylor, scott, davis, allen, adams, lewis, williams, jones, wilson, martin, johnson*

Hispanic last names *ruiz, alvarez, vargas, castillo, gomez, soto, gonzalez, sanchez, rivera, mendoza, martinez, torres, rodriguez, perez, lopez, medina, diaz, garcia, castro, cruz*

Asian last names *cho, wong, tang, huang, chu, chung, ng, wu, liu, chen, lin, yang, kim, chang, shah, wang, li, khan, singh, hong*

Russian last names *gurin, minsky, sokolov, markov, maslow, novikoff, mishkin, smirnov, orloff, ivanov, sokoloff, davidoff, savin, romanoff, babinski, sorokin, levin, pavlov, rodin, agin*

Chinese last names *chung, liu, wong, huang, ng, hu, chu, chen, lin, liang, wang, wu, yang, tang, chang, hong, li*

Islam words *allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad*

Christianity words *baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church*

Starting with a breakdown of ethnicity by last name compiled by [3][1], we identify 20 last names for each Whites, Asians, and Hispanics as follows: 1) Start with list of top 50 last names by percent of that ethnicity, conditioned on being top 5000 surnames overall, as well as the top 50 last names by total number in that ethnicity (i.e., multiplied count of that last name by percent in that ethnicity). 2) Choose the 20 names that appeared most on average in the Google Books/COHA vectors over time (with a minimum number for each time period). This second step ensures that an accurate ethnicity vector is identified each time period, with minimal distortions. Russian last names are collated from various sources online.

[1] available `https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv`

A.3 Neutral Words

Occupations *janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper, accountant, physicist, gardener, dentist, weaver, blacksmith, psychologist, supervisor, mathematician, surveyor, tailor, designer, economist, mechanic, laborer, postmaster, broker, chemist, librarian, attendant, clerical, musician, porter, scientist, carpenter, sailor, instructor, sheriff, pilot, inspector, mason, baker, administrator, architect, collector, operator, surgeon, driver, painter, conductor, nurse, cook, engineer, retired, sales, lawyer, clergy, physician, farmer, clerk, manager, guard, artist, smith, official, police, doctor, professor, student, judge, teacher, author, secretary, soldier*

Professional Occupations[2] *statistician, auctioneer, photographer, geologist, accountant, physicist, dentist, psychologist, supervisor, mathematician, designer, economist, postmaster, broker, chemist, librarian, scientist, instructor, pilot, administrator, architect, surgeon, nurse, engineer, lawyer, physician, manager, official, doctor, professor, student, judge, teacher, author*

Occupations with Human Stereotype Scores from [4] *teacher, author, mechanic, broker, baker, surveyor, laborer, surgeon, gardener, painter, dentist, janitor, athlete, manager, conductor, carpenter, housekeeper, secretary, economist, geologist, clerk, doctor, judge, physician, lawyer, artist, instructor, dancer, photographer, inspector, musician, soldier, librarian, professor, psychologist, nurse, sailor, accountant, architect, chemist, administrator, physicist, scientist, farmer*

Adjectives from [5, 6] *headstrong, thankless, tactful, distrustful, quarrelsome, effeminate, fickle, talkative, dependable, resentful, sarcastic, unassuming, changeable, resourceful, persevering, forgiving, assertive, individualistic, vindictive, sophisticated, deceitful, impulsive, sociable, methodical, idealistic, thrifty, outgoing, intolerant, autocratic, conceited, inventive, dreamy, appreciative, forgetful, forceful, submissive, pessimistic, versatile, adaptable, reflective, inhibited, outspoken, quitting, unselfish, immature, painstaking, leisurely, infantile, sly, praising, cynical, irresponsible, arrogant, obliging, unkind, wary, greedy, obnoxious, irritable, discreet, frivolous, cowardly, rebellious, adventurous, enterprising, unscrupulous, poised, moody, unfriendly, optimistic, disorderly, peaceable, considerate, humorous, worrying, preoccupied, trusting, mischievous, robust, superstitious, noisy, tolerant, realistic, masculine, witty, informal, prejudiced, reckless, jolly, courageous, meek, stubborn, aloof, sentimental, complaining, unaffected, cooperative, unstable, feminine, timid, retiring, relaxed, imaginative, shrewd, conscientious, industrious, hasty, commonplace, lazy, gloomy, thoughtful, dignified, wholesome, affectionate, aggressive, awkward, energetic, tough, shy, queer, careless, restless, cautious, polished, tense, suspicious, dissatisfied, ingenious, fearful, daring, persistent, demanding, impatient, contented, selfish, rude, spontaneous, conventional, cheerful, enthusiastic, modest, ambitious, alert, defensive, mature, coarse, charming, clever, shallow, deliberate, stern, emotional, rigid, mild, cruel, artistic, hurried, sympathetic, dull, civilized, loyal, withdrawn, confident, indifferent, conservative, foolish, moderate, handsome, helpful, gentle, dominant, hostile, generous, reliable, sincere, precise, calm, healthy, attractive, progressive, confused, rational, stable, bitter, sensitive, initiative, loud, thorough, logical, intelligent, steady, formal, complicated, cool, curious, reserved, silent, honest, quick, friendly, efficient, pleasant, severe, peculiar, quiet, weak, anxious, nervous, warm, slow, dependent, wise, organized, affected, reasonable, capable, active, independent, patient, practical, serious, understanding, cold, responsible, simple, original, strong, determined, natural, kind*

Larger adjective list, mostly from [7] *disorganized, devious, impressionable, circumspect, impassive, aimless, effeminate, unfathomable, fickle, unprincipled, inoffensive, reactive, providential, resentful, bizarre, impractical, sarcastic, misguided, imitative, pedantic, venomous, erratic, insecure, resourceful, neurotic, forgiving, profligate, whimsical, assertive, incorruptible, individualistic, faithless, disconcerting, barbaric, hypnotic, vindictive, observant, dissolute, frightening, complacent, boisterous, pretentious, disobedient, tasteless, sedentary, sophisticated, regimental, mellow, deceitful, impulsive, playful, sociable, methodical, willful, idealistic, boyish, callous, pompous, unchanging, crafty, punctual, compassionate, intolerant, challenging, scornful, possessive, conceited, imprudent, dutiful, lovable, disloyal, dreamy, appreciative, forgetful, unrestrained, forceful, submissive, predatory, fanatical, illogical, tidy, aspiring, studious, adaptable, conciliatory, artful, thoughtless, deceptive, frugal, reflective, insulting, unreliable, stoic, hysterical, rustic, inhibited, outspoken, unhealthy, ascetic, skeptical, painstaking, contemplative, leisurely, sly, mannered, outrageous, lyrical, placid, cynical, irresponsible,*

[2] These were hand-coded from the overall list of occupations; follow-on work should study this more systematically.

vulnerable, arrogant, persuasive, perverse, steadfast, crisp, envious, naive, greedy, presumptuous, obnoxious, irritable, dishonest, discreet, sporting, hateful, ungrateful, frivolous, reactionary, skillful, cowardly, sordid, adventurous, dogmatic, intuitive, bland, indulgent, discontented, dominating, articulate, fanciful, discouraging, treacherous, repressed, moody, sensual, unfriendly, optimistic, clumsy, contemptible, focused, haughty, morbid, disorderly, considerate, humorous, preoccupied, airy, impersonal, cultured, trusting, respectful, scrupulous, scholarly, superstitious, tolerant, realistic, malicious, irrational, sane, colorless, masculine, witty, inert, prejudiced, fraudulent, blunt, childish, brittle, disciplined, responsive, courageous, bewildered, courteous, stubborn, aloof, sentimental, athletic, extravagant, brutal, manly, cooperative, unstable, youthful, timid, amiable, retiring, fiery, confidential, relaxed, imaginative, mystical, shrewd, conscientious, monstrous, grim, questioning, lazy, dynamic, gloomy, troublesome, abrupt, eloquent, dignified, hearty, gallant, benevolent, maternal, paternal, patriotic, aggressive, competitive, elegant, flexible, gracious, energetic, tough, contradictory, shy, careless, cautious, polished, sage, tense, caring, suspicious, sober, neat, transparent, disturbing, passionate, obedient, crazy, restrained, fearful, daring, prudent, demanding, impatient, cerebral, calculating, amusing, honorable, casual, sharing, selfish, ruined, spontaneous, admirable, conventional, cheerful, solitary, upright, stiff, enthusiastic, petty, dirty, subjective, heroic, stupid, modest, impressive, orderly, ambitious, protective, silly, alert, destructive, exciting, crude, ridiculous, subtle, mature, creative, coarse, passive, oppressed, accessible, charming, clever, decent, miserable, superficial, shallow, stern, winning, balanced, emotional, rigid, invisible, desperate, cruel, romantic, agreeable, hurried, sympathetic, solemn, systematic, vague, peaceful, humble, dull, expedient, loyal, decisive, arbitrary, earnest, confident, conservative, foolish, moderate, helpful, delicate, gentle, dedicated, hostile, generous, reliable, dramatic, precise, calm, healthy, attractive, artificial, progressive, odd, confused, rational, brilliant, intense, genuine, mistaken, driving, stable, objective, sensitive, neutral, strict, angry, profound, smooth, ignorant, thorough, logical, intelligent, extraordinary, experimental, steady, formal, faithful, curious, reserved, honest, busy, educated, liberal, friendly, efficient, sweet, surprising, mechanical, clean, critical, criminal, soft, proud, quiet, weak, anxious, solid, complex, grand, warm, slow, false, extreme, narrow, dependent, wise, organized, pure, directed, dry, obvious, popular, capable, secure, active, independent, ordinary, fixed, practical, serious, fair, understanding, constant, cold, responsible, deep, religious, private, simple, physical, original, working, strong, modern, determined, open, political, difficult, knowledge, kind

Competence Adjectives[3] *precocious, resourceful, inquisitive, sagacious, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, luminous, venerable, imaginative, shrewd, thoughtful, sage, smart, ingenious, clever, brilliant, logical, intelligent, apt, genius, wise*

Physical Appearance Adjectives[4] *alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong*

Terrorism related words *terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target, conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold, wreckage, aggression, slaughter, execute, overthrow, casualties, massacre, retaliation, proliferation, militia, hostility, debris, acid, execution, militant, rocket, guerrilla, sacrifice, enemy, soldier, terrorist, missile, hostile, revolution, resistance, shoot*

Outsider Adjectives *devious, bizarre, venomous, erratic, barbaric, frightening, deceitful, forceful, deceptive, envious, greedy, hateful, contemptible, brutal, monstrous, calculating, cruel, intolerant, aggressive, monstrous*

Princeton Trilogy Stereotypes from [8, 9, 10] For 1 analysis, all scores in tables used (where enough embedding data exists). For the second, only those for which scores over time were available were used.

All static scores in tables *courteous, deceitful, meditative, artistic, conservative, reserved, nationalistic, loyal, ignorant, superstitious, quiet, sly, religious, traditional, industrious*

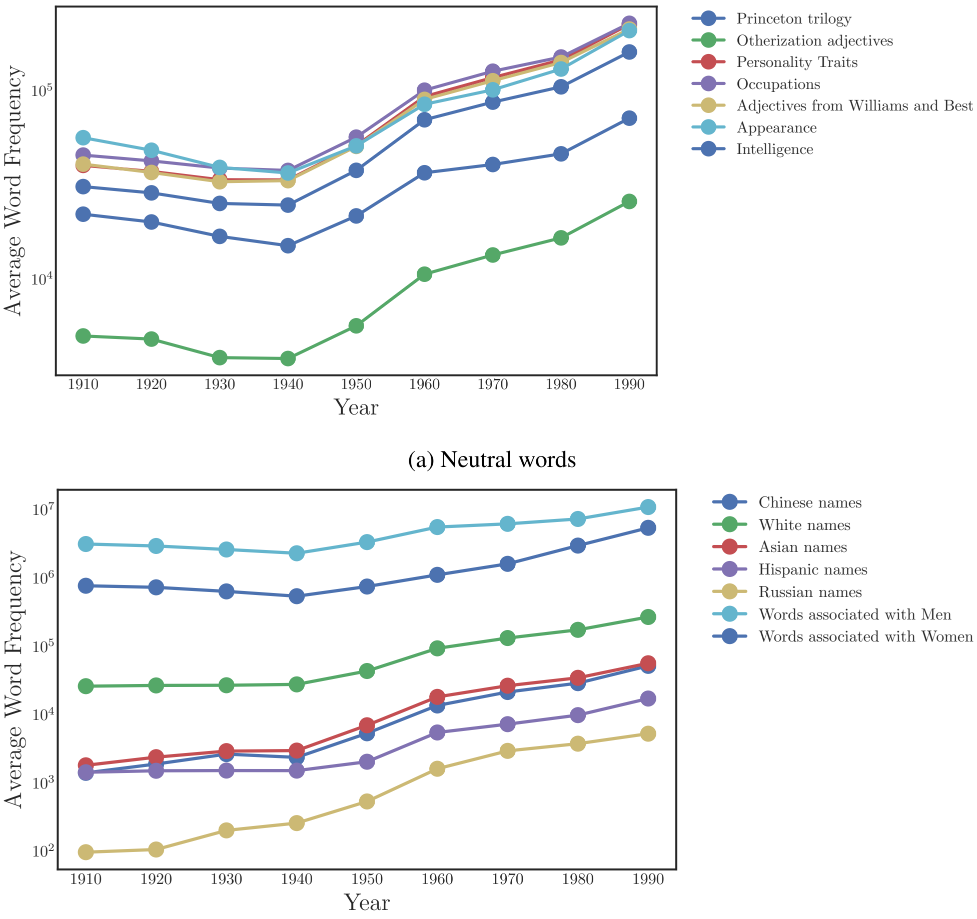
Scores over time *deceitful, meditative, conservative, reserved, loyal, ignorant, superstitious, quiet, sly, religious, traditional, industrious*

[3] mostly from `https://www.e-education.psu.edu/writingrecommendationlettersonline/node/151`,`https://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-intelligent-or-wise-people`

A.4 Embedding Quality

Here we show that the quality of the average vector constructed for each group does not appreciably change over time and thus cannot explain the overtime trends.

Figure A.1 shows the average frequency of words in each list across time in the COHA embeddings. Counts in general increase for all lists throughout time as datasets get larger. However, vector quality remains about the same; Figure A.2 shows the average variance on each dimension across time for each group. Except for Russian names (which are used for a plot in the appendix), these variances remain relatively steady throughout time, with small potential decreases as the size of the training datasets increase, indicating that average vector quality does not change appreciably.



(a) Neutral words

(b) Group words

Figure A.1: Average counts of words in a list in the COHA embeddings over time

[4] mostly from http://usefulenglish.ru/vocabulary/appearance-and-character, http://www.sightwordsgame.com/parts-of-speech/adjectives/appearance/, http://www.stgeorges.co.uk/blog/physical-appearance-adjectives-the-bald-and-the-beautiful

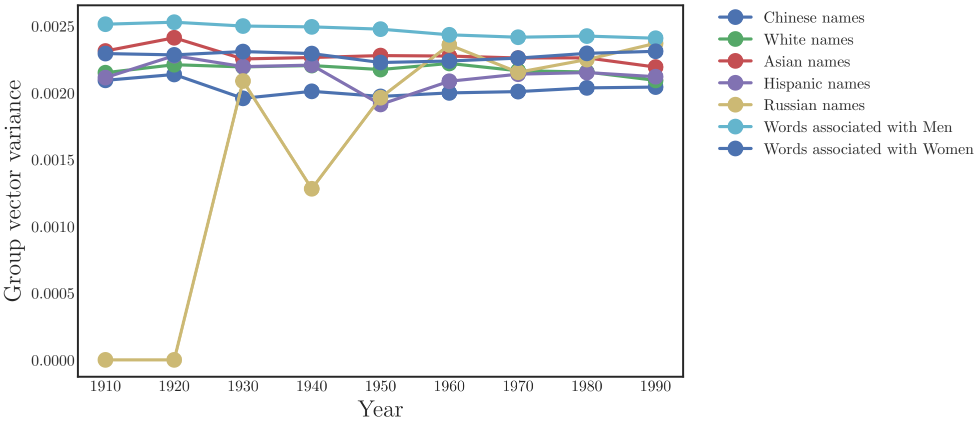


Figure A.2: Average variance in each embedding dimension over time for each group in the COHA embeddings

A.5 Similarity metrics

In the main exposition, we use the relative norm bias metric, $\sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2$. Using relative cosine similarity instead, $\sum_{v_m \in M} v_m \cdot v_2 - v_m \cdot v_1$, makes no difference – the metrics have Pearson correlation $> .95$ in general (note that Pearson correlation of 1 is a perfectly linear relationship). The following table shows their correlation for a few embedding/neutral word/group combinations, and the pattern holds generally.

Table A.1: Pearson correlation of two possible bias metrics

Embedding	Neutral Words	Groups	Pearson correlation
Google News	Occupations	Men, Women	.998
Google News	Personality Traits	Men, Women	.998
SGNS 1990	Occupations	Men, Women	.997
SGNS 1990	Personality Traits	Men, Women	.994
Google News	Occupations	Whites, Asians	.973
Google News	Personality Traits	Whites, Asians	.993
SGNS 1990	Occupations	Whites, Asians	.991
SGNS 1990	Personality Traits	Whites, Asians	.999

We thus primarily report results using only the relative norm difference bias metric.

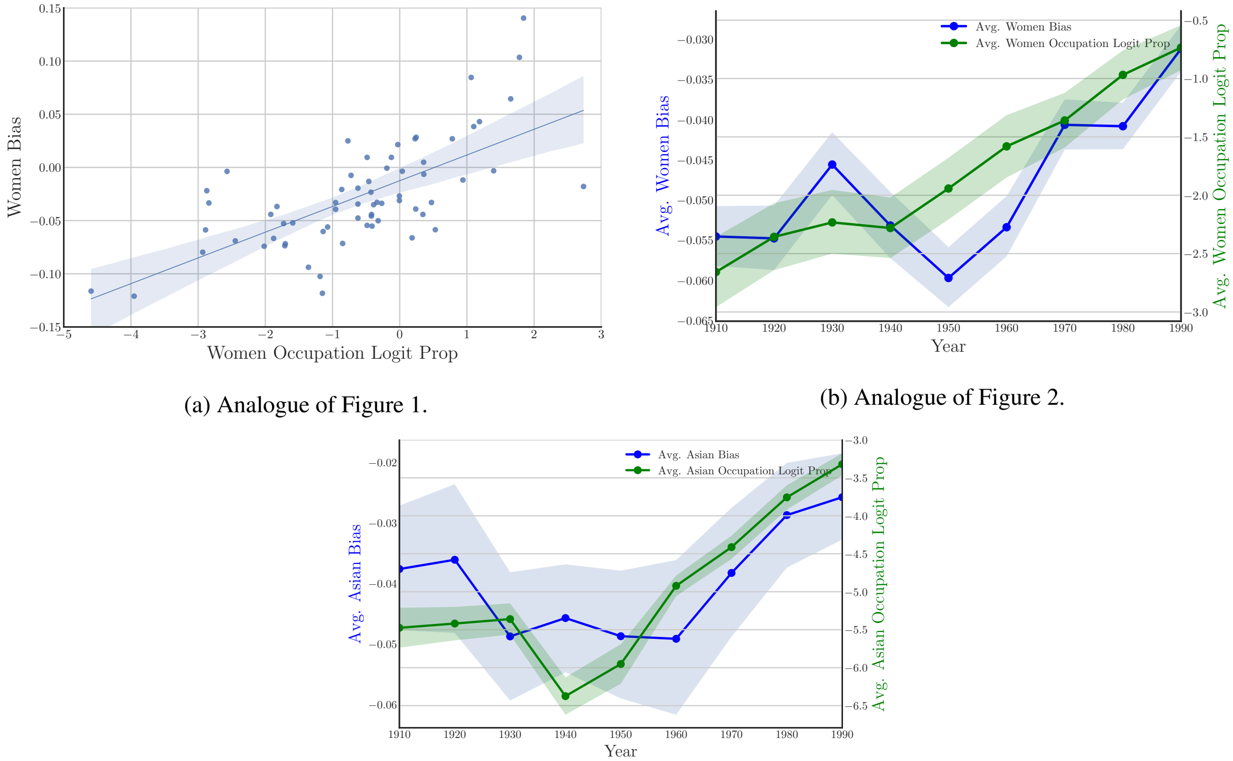
A.6 Occupation percentage transformation

In the main exposition, we use the relative percent difference of participation statistics in an occupation as a representation of the census data. [11], whose preprocessing steps we follow for the rest of the census data, instead uses the following transformation (for gender) occupation statistics:

$$\text{logit-prop}(p) = \log \frac{p}{1-p} \tag{1}$$
$$\text{where } p = \% \text{ of women in occupation}$$

For ethnicity, an analogous metric, the conditional logit-proportion in Equation (2) can be used.

$$\text{cond-logit-prop}(\text{group } 1, \text{group } 2) = \log \frac{p}{1-p} \tag{2}$$
$$\text{where } p = \frac{\% \text{ of group } 1}{\% \text{ of group } 1 + \ \% \text{ of group } 2}$$



(a) Analogue of Figure 1.

(b) Analogue of Figure 2.

(c) Analogue of Figure 3.

Figure A.3: Analogue of occupation verification figures using logit-proportion for occupation data.

The results do not qualitatively change when these transformations are used instead, as our (linear) models already are rough estimates for any true relationship. For brevity, we reproduce just the relevant figures from the main exposition using these logit proportion metrics. Statistical tests and equivalent plots from the rest of this appendix can also be reproduced, and results are qualitatively similar.

Appendix B Gender analysis

B.1 Additional Validation Analysis

B.1.1 Occupations

Table B.1 shows the top occupations and adjectives by gender in the Google News embedding.

Below, we first show the regression table corresponding to Figure 1, and then show the same plot for other embeddings, as well as the subset of occupations corresponding to *professional* occupations.

Occupations		Adjectives	
Man	Woman	Man	Woman
carpenter	nurse	honorable	maternal
mechanic	midwife	ascetic	romantic
mason	librarian	amiable	submissive
blacksmith	housekeeper	dissolute	hysterical
retired	dancer	arrogant	elegant
architect	teacher	erratic	caring
engineer	cashier	heroic	delicate
mathematician	student	boyish	superficial
shoemaker	designer	fanatical	neurotic
physicist	weaver	aimless	attractive

Table B.1: Top occupations and adjectives by gender in the Google News embedding.

Dep. Variable:	Women Bias	**R-squared:**	0.499
Model:	OLS	**Adj. R-squared:**	0.491
Method:	Least Squares	**F-statistic:**	63.65
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	3.53e-11
Time:	19:59:00	**Log-Likelihood:**	128.37
No. Observations:	66	**AIC:**	-252.7
Df Residuals:	64	**BIC:**	-248.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. $%$ Difference	0.0007	9.13e-05	7.978	0.000	0.001	0.001
const	-0.0118	0.005	-2.487	0.015	-0.021	-0.002

Omnibus:	1.265	**Durbin-Watson:**	1.652
Prob(Omnibus):	0.531	**Jarque-Bera (JB):**	0.829
Skew:	0.268	**Prob(JB):**	0.661
Kurtosis:	3.119	**Cond. No.**	56.7

Table B.2: OLS Regression Results corresponding to Figure 1, regressing women occupation percent difference with word vector relative distance in Google News vectors in 2015.

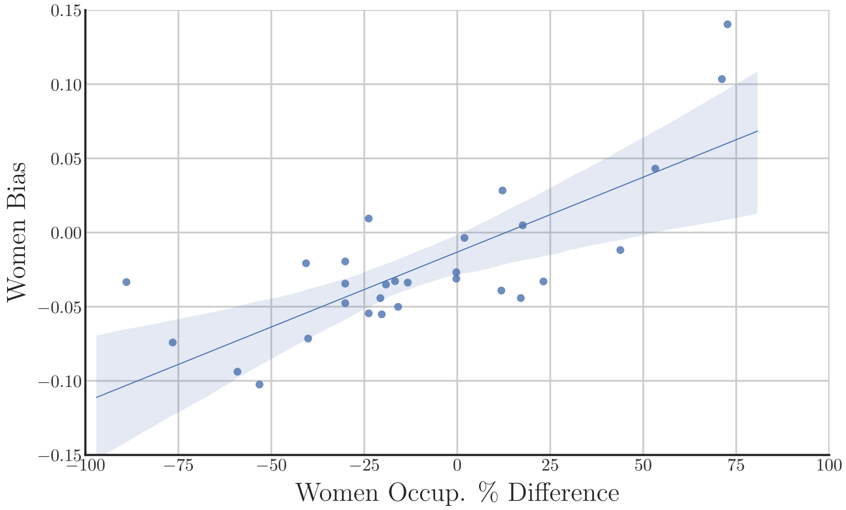


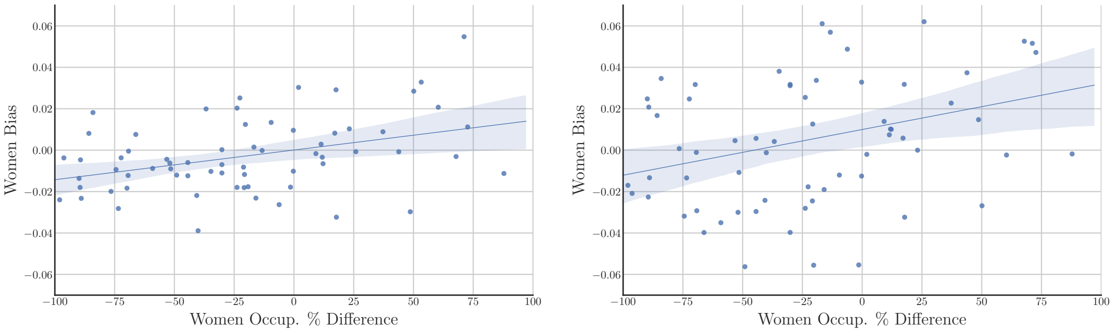
Figure B.1: Percent woman in an occupation vs relative norm distance from *professional* occupations in Google News vectors. $p < 10^{-5}$, r-squared$= .595$. Regression bias coefficient confidence interval: $(-.026, 0)$. Note that there is little difference in model from all occupations.

Dep. Variable:	Women Bias	**R-squared:**	0.595
Model:	OLS	**Adj. R-squared:**	0.580
Method:	Least Squares	**F-statistic:**	39.72
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	9.57e-07
Time:	19:59:11	**Log-Likelihood:**	58.614
No. Observations:	29	**AIC:**	-113.2
Df Residuals:	27	**BIC:**	-110.5
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0010	0.000	6.303	0.000	0.001	0.001
const	-0.0132	0.006	-2.072	0.048	-0.026	-0.000

Omnibus:	3.544	**Durbin-Watson:**	1.514
Prob(Omnibus):	0.170	**Jarque-Bera (JB):**	2.644
Skew:	0.739	**Prob(JB):**	0.267
Kurtosis:	3.034	**Cond. No.**	40.9

Table B.3: OLS Regression Results corresponding to Figure B.1, regressing women occupation percent difference with word vector relative distance in Google News vectors in 2015 for *professional* occupations.



(a) Common Crawl GloVe. $p = .001$, r-squared$= .149$.

(b) Wikipedia GloVe. $p = .008$, r-squared$= .105$.

Figure B.2: Percent woman in an occupation vs relative norm distance from occupations to the respective gender words in Common Crawl and Wikipedia.

Dep. Variable:	Women Bias	**R-squared:**	0.149
Model:	OLS	**Adj. R-squared:**	0.136
Method:	Least Squares	**F-statistic:**	11.19
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.00138
Time:	19:59:16	**Log-Likelihood:**	178.31
No. Observations:	66	**AIC:**	-352.6
Df Residuals:	64	**BIC:**	-348.2
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0001	4.28e-05	3.345	0.001	5.77e-05	0.000
const	6.248e-05	0.002	0.028	0.978	-0.004	0.004

Omnibus:	1.013	**Durbin-Watson:**	2.405
Prob(Omnibus):	0.603	**Jarque-Bera (JB):**	0.578
Skew:	0.217	**Prob(JB):**	0.749
Kurtosis:	3.148	**Cond. No.**	56.7

Table B.4: OLS Regression Results corresponding to Figure B.2a, regressing women occupation percent difference with word vector relative distance in Common Crawl vectors.

Dep. Variable:	Women Bias	**R-squared:**	0.105
Model:	OLS	**Adj. R-squared:**	0.091
Method:	Least Squares	**F-statistic:**	7.546
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.00780
Time:	19:59:22	**Log-Likelihood:**	136.80
No. Observations:	66	**AIC:**	-269.6
Df Residuals:	64	**BIC:**	-265.2
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	8.03e-05	2.747	0.008	6.02e-05	0.000
const	0.0100	0.004	2.395	0.020	0.002	0.018

Omnibus:	1.316	**Durbin-Watson:**	2.481
Prob(Omnibus):	0.518	**Jarque-Bera (JB):**	1.073
Skew:	0.057	**Prob(JB):**	0.585
Kurtosis:	2.386	**Cond. No.**	56.7

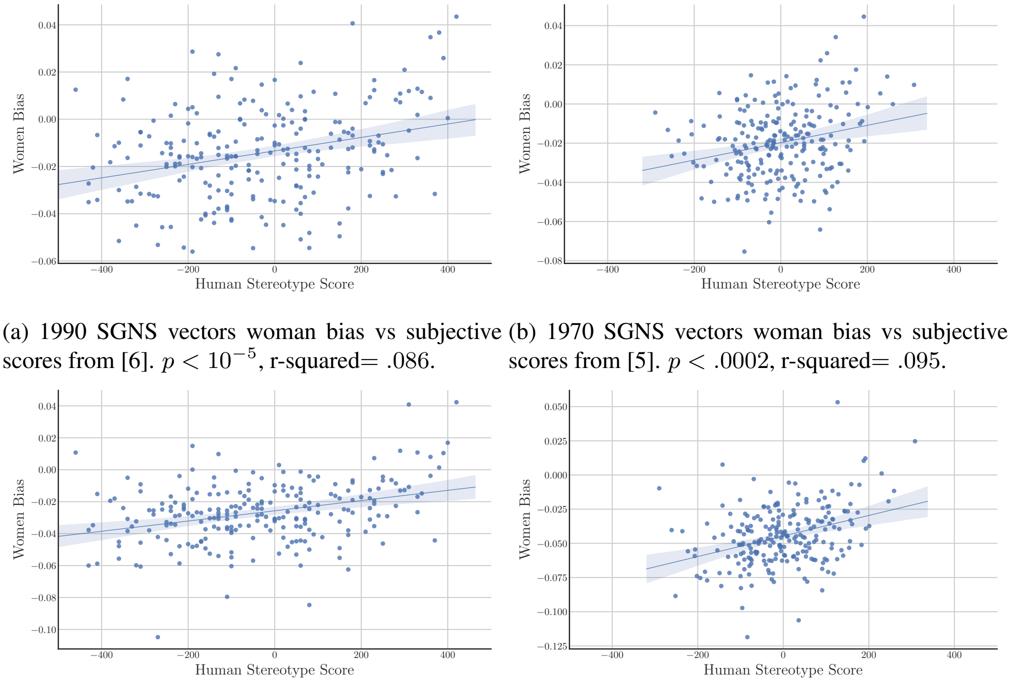
Table B.5: OLS Regression Results corresponding to Figure B.2b, regressing women occupation percent difference with word vector relative distance in Wikipedia vectors.

Table B.6: Most Man and Woman *residuals* when regressing embedding bias vs occupation gender proportion. The more Woman (Man) a residual the more biased the embedding is toward Women (Men) above what occupation proportions would suggest.

Man	Woman
secretary	nurse
mechanic	housekeeper
musician	gardener
architect	clerk
janitor	librarian
carpenter	sailor
broker	judge
geologist	artist
accountant	dancer
economist	painter

B.1.2 Adjectives

This subsection contains the supplementary material related to adjectives. In particular, we show the scatter plots and fit information corresponding the stereotype scores from [5] and [6] to the SGNS and SVD embeddings from the respective decades.



(a) 1990 SGNS vectors woman bias vs subjective scores from [6]. $p < 10^{-5}$, r-squared= .086.

(b) 1970 SGNS vectors woman bias vs subjective scores from [5]. $p < .0002$, r-squared= .095.

(c) 1990 SVD vectors woman bias vs subjective scores from [6]. $p < 10^{-6}$, r-squared= .116.

(d) 1970 SVD vectors woman bias vs subjective scores from [5]. $p < 10^{-7}$, r-squared= .127.

Figure B.3: Human stereotype score vs relative norm distance from adjectives to the respective gender words in 1970 and 1990, using both SGNS and SVD vectors

Note the stereotype scores used in the plots are linear transformations of the scores reported in the original papers, done to standardize figures. Given a score r from 1990 [6], the transformed score $r' = 500 - 10r$. Given a score from 1970 [5], the transformed score $r' = 500 - r$.

Dep. Variable:	Women Bias	**R-squared:**	0.095
Model:	OLS	**Adj. R-squared:**	0.087
Method:	Least Squares	**F-statistic:**	11.87
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.25e-05
Time:	19:59:27	**Log-Likelihood:**	592.34
No. Observations:	230	**AIC:**	-1179.
Df Residuals:	227	**BIC:**	-1168.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	2.842e-05	6.16e-06	4.612	0.000	1.63e-05	4.06e-05
counts	4.005e-09	2.63e-09	1.525	0.129	-1.17e-09	9.18e-09
const	-0.0143	0.001	-10.492	0.000	-0.017	-0.012

Omnibus:	0.981	**Durbin-Watson:**	1.856
Prob(Omnibus):	0.612	**Jarque-Bera (JB):**	1.095
Skew:	0.129	**Prob(JB):**	0.578
Kurtosis:	2.783	**Cond. No.**	5.69e+05

Table B.7: Regression table associated with Figure B.3a

Dep. Variable:	Women Bias	**R-squared:**	0.062
Model:	OLS	**Adj. R-squared:**	0.053
Method:	Least Squares	**F-statistic:**	7.464
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.000725
Time:	19:59:29	**Log-Likelihood:**	601.86
No. Observations:	230	**AIC:**	-1198.
Df Residuals:	227	**BIC:**	-1187.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	4.434e-05	1.15e-05	3.863	0.000	2.17e-05	6.7e-05
counts	2.932e-10	4.64e-09	0.063	0.950	-8.85e-09	9.43e-09
const	-0.0198	0.001	-15.407	0.000	-0.022	-0.017

Omnibus:	0.434	**Durbin-Watson:**	1.827
Prob(Omnibus):	0.805	**Jarque-Bera (JB):**	0.223
Skew:	0.051	**Prob(JB):**	0.894
Kurtosis:	3.113	**Cond. No.**	3.03e+05

Table B.8: Regression table associated with Figure B.3b

Dep. Variable:	Women Bias	**R-squared:**	0.116
Model:	OLS	**Adj. R-squared:**	0.108
Method:	Least Squares	**F-statistic:**	14.86
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	8.59e-07
Time:	19:59:32	**Log-Likelihood:**	600.10
No. Observations:	230	**AIC:**	-1194.
Df Residuals:	227	**BIC:**	-1184.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	3.197e-05	5.96e-06	5.366	0.000	2.02e-05	4.37e-05
counts	2.304e-09	2.54e-09	0.907	0.365	-2.7e-09	7.31e-09
const	-0.0262	0.001	-19.937	0.000	-0.029	-0.024

Omnibus:	12.746	**Durbin-Watson:**	1.810
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	28.422
Skew:	-0.160	**Prob(JB):**	6.73e-07
Kurtosis:	4.692	**Cond. No.**	5.69e+05

Table B.9: Regression table associated with Figure B.3c

Dep. Variable:	Women Bias	**R-squared:**	0.127
Model:	OLS	**Adj. R-squared:**	0.119
Method:	Least Squares	**F-statistic:**	16.48
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	2.08e-07
Time:	19:59:34	**Log-Likelihood:**	571.07
No. Observations:	230	**AIC:**	-1136.
Df Residuals:	227	**BIC:**	-1126.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	7.525e-05	1.31e-05	5.734	0.000	4.94e-05	0.000
counts	-1.464e-09	5.3e-09	-0.276	0.783	-1.19e-08	8.99e-09
const	-0.0445	0.001	-30.277	0.000	-0.047	-0.042

Omnibus:	16.771	**Durbin-Watson:**	1.840
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36.244
Skew:	0.317	**Prob(JB):**	1.35e-08
Kurtosis:	4.839	**Cond. No.**	3.03e+05

Table B.10: Regression table associated with Figure B.3d

B.2 Amazon Mturk Stereotypes and Embedding bias

Dep. Variable:	Women Bias	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.439
Method:	Least Squares	**F-statistic:**	34.69
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	5.72e-07
Time:	20:59:25	**Log-Likelihood:**	80.806
No. Observations:	44	**AIC:**	-157.6
Df Residuals:	42	**BIC:**	-154.0
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0007	0.000	5.890	0.000	0.000	0.001
const	-0.0140	0.006	-2.240	0.030	-0.027	-0.001

Omnibus:	1.880	**Durbin-Watson:**	1.589
Prob(Omnibus):	0.391	**Jarque-Bera (JB):**	1.453
Skew:	0.445	**Prob(JB):**	0.484
Kurtosis:	2.960	**Cond. No.**	52.9

Table B.11: Women embedding bias vs occupation percent difference for occupations for which census data, MTurk data, and embedding bias is available.

Dep. Variable:	Women Bias	**R-squared:**	0.655
Model:	OLS	**Adj. R-squared:**	0.647
Method:	Least Squares	**F-statistic:**	79.88
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	2.87e-11
Time:	20:59:29	**Log-Likelihood:**	90.999
No. Observations:	44	**AIC:**	-178.0
Df Residuals:	42	**BIC:**	-174.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Stereotype Score	0.0611	0.007	8.938	0.000	0.047	0.075
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.12: Women embedding bias vs MTurk Stereotype score for occupations for which census data, MTurk data, and embedding bias is available.

We next perform an additional joint regression, with the crowdsource scores and the occupation percent difference as covariates and the embedding bias as the outcome. The crowdsource scores remain significantly associated with the embedding bias while the occupation percent difference do not (at $p < 10^{-5}$, versus $p = .203$ for occupation percentage, $r^2 = .669$, intercept confidence interval $(-.281, .027)$). This result indicates that the embedding bias is more closely aligned with human stereotypes than with actual occupation participation.

Dep. Variable:	Women Bias	**R-squared:**	0.669
Model:	OLS	**Adj. R-squared:**	0.653
Method:	Least Squares	**F-statistic:**	41.42
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.44e-10
Time:	20:59:29	**Log-Likelihood:**	91.880
No. Observations:	44	**AIC:**	-177.8
Df Residuals:	41	**BIC:**	-172.4
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	0.000	1.294	0.203	-0.000	0.000
Stereotype Score	0.0517	0.010	5.179	0.000	0.032	0.072
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.13: Women embedding bias vs occupation percent difference and MTurk stereotype data for occupations for which census data, MTurk data, and embedding bias is available.

Dep. Variable:	Stereotype Score	**R-squared:**	0.537
Model:	OLS	**Adj. R-squared:**	0.526
Method:	Least Squares	**F-statistic:**	48.78
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.53e-08
Time:	20:59:27	**Log-Likelihood:**	-29.168
No. Observations:	44	**AIC:**	62.34
Df Residuals:	42	**BIC:**	65.90
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0106	0.002	6.984	0.000	0.008	0.014
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.14: MTurk Stereotype score vs occupation percent difference and MTurk stereotype data for occupations for which census data, MTurk data, and embedding bias is available.

B.3 Dynamic Analysis

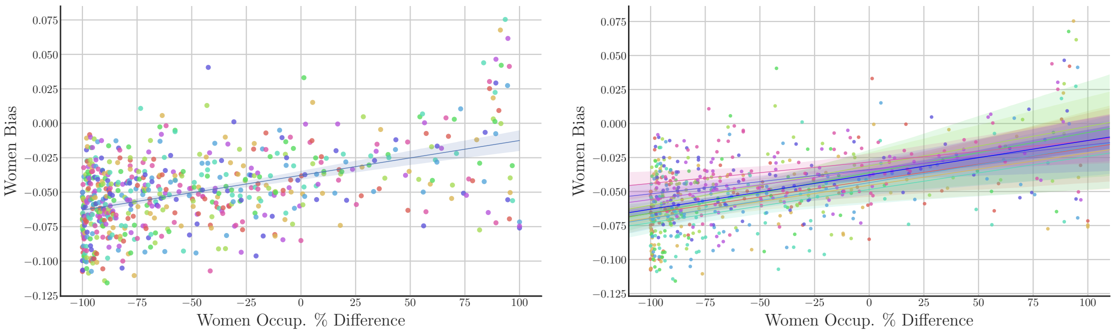
This section contains additional information for gender bias associated with the "dynamic" or over-time analysis. We first present additional information such as regression tables and plots with SVD embeddings for robustness. We then show the occupations and adjectives most associated with men and women, respectively, for each decade.

B.3.1 Model from occupation percentage to embedding bias over time

In this section, we show that the relationship between occupation relative percentage to embedding bias is approximately *consistent* over time, i.e. a *single* linear model performs well across datasets/time. This consistency allows us to extract meaning from trends in the embedding associations over time.

We first train a single model for all (occupation percentage, embedding bias) pairs across time. We compare its performance to a single model where there are additional term for each year. Next, we compare its performance to models trained separately for each year, showing that the single model both has similar parameters and performance to such separate models. Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation). Finally, we repeat the entire analysis using the SVD embeddings.

We note that the separate models and the joint model are similar both in their parameters and their performance.



(a) Single model trained for all time

(b) Separate model trained for each decade

Figure B.4: Scatter plot for occupation proportion vs embedding bias all years of SVD embeddings, along with individually trained regression lines.

Dep. Variable:	Women Bias	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	196.5
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	4.17e-39
Time:	17:12:19	**Log-Likelihood:**	1445.6
No. Observations:	638	**AIC:**	-2887.
Df Residuals:	636	**BIC:**	-2878.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.8e-05	14.020	0.000	0.000	0.000
const	-0.0378	0.001	-28.576	0.000	-0.040	-0.035

Omnibus:	7.243	**Durbin-Watson:**	1.744
Prob(Omnibus):	0.027	**Jarque-Bera (JB):**	7.296
Skew:	0.218	**Prob(JB):**	0.0260
Kurtosis:	3.291	**Cond. No.**	97.4

Table B.15: Single model for Embedding bias vs occupation percentage difference across years

Next, here is the regression with an individual term for each year.

Dep. Variable:	Women Bias	**R-squared:**	0.298
Model:	OLS	**Adj. R-squared:**	0.288
Method:	Least Squares	**F-statistic:**	29.59
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	4.45e-43
Time:	17:12:19	**Log-Likelihood:**	1472.5
No. Observations:	638	**AIC:**	-2925.
Df Residuals:	628	**BIC:**	-2880.
Df Model:	9		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	1.77e-05	13.342	0.000	0.000	0.000
const	-0.0346	0.001	-29.916	0.000	-0.037	-0.032
yr_1910.0	-0.0076	0.003	-2.813	0.005	-0.013	-0.002
yr_1920.0	-0.0109	0.003	-4.068	0.000	-0.016	-0.006
yr_1930.0	0.0006	0.003	0.222	0.824	-0.005	0.006
yr_1940.0	-0.0053	0.003	-1.815	0.070	-0.011	0.000
yr_1950.0	-0.0143	0.003	-5.275	0.000	-0.020	-0.009
yr_1960.0	-0.0101	0.003	-3.735	0.000	-0.015	-0.005
yr_1970.0	0.0033	0.003	1.226	0.221	-0.002	0.009
yr_1980.0	0.0011	0.003	0.395	0.693	-0.004	0.006
yr_1990.0	0.0085	0.003	3.099	0.002	0.003	0.014

Omnibus:	23.364	**Durbin-Watson:**	1.746
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27.127
Skew:	0.404	**Prob(JB):**	1.29e-06
Kurtosis:	3.608	**Cond. No.**	8.37e+17

Table B.16: Embedding bias vs occupation percentage difference with additional term for each year

Next, we show the performance and model values when models are trained separately for each embedding year.

Model Year	r^2	coefficient p-value	coefficient value	intercept p-value	intercept value
1910	0.2332	$1.508e-05$	$0.0002408 \pm 5.181e-05$	$4.565e-15$	-0.04197 ± 0.004226
1920	0.2236	$2.398e-05$	$0.0002385 \pm 5.275e-05$	$1.3e-16$	-0.04547 ± 0.004215
1930	0.2865	$8.952e-07$	$0.0002896 \pm 5.386e-05$	$1.904e-10$	-0.03109 ± 0.004195
1940	0.205	0.0002197	$0.000296 \pm 7.525e-05$	$9.189e-08$	-0.03622 ± 0.005962
1950	0.2099	$5.191e-05$	$0.0002437 \pm 5.651e-05$	$9.101e-18$	-0.04853 ± 0.004224
1960	0.2821	$1.586e-06$	$0.0002639 \pm 5.031e-05$	$2.247e-18$	-0.04347 ± 0.003672
1970	0.1827	0.0001801	$0.0001795 \pm 4.536e-05$	$2.905e-16$	-0.03378 ± 0.003176
1980	0.2567	$7.626e-06$	$0.0002158 \pm 4.454e-05$	$6.358e-19$	-0.03429 ± 0.002792
1990	0.1231	0.002905	$0.0001572 \pm 5.088e-05$	$1.706e-14$	-0.02832 ± 0.002913

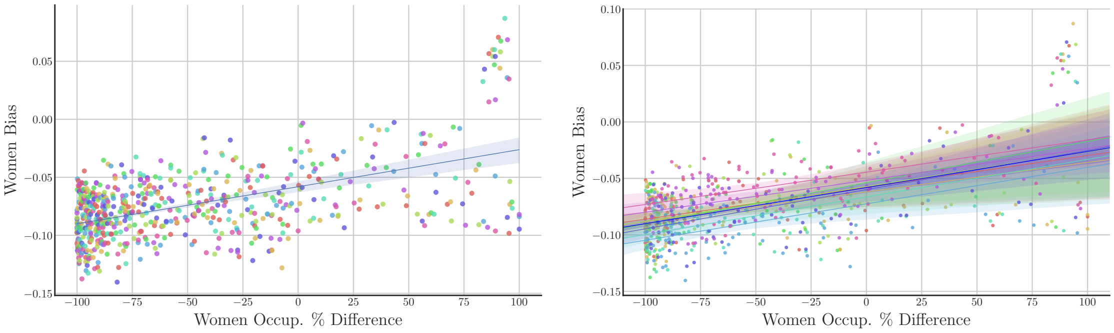
Table B.17: Performances of models trained separately for each decade.

Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation).

Year	MSE using own model	MSE using model from other years
1910	0.000587	0.0005998
1920	0.0006938	0.000743
1930	0.0006222	0.0006474
1940	0.0008265	0.0008325
1950	0.0006428	0.0007494
1960	0.0005713	0.0006107
1970	0.0004408	0.0005077
1980	0.0003628	0.000389
1990	0.0004424	0.0006109

Table B.18: Comparative performance of model trained using a given year's data versus one using all other years

SVD embeddings We repeat the above using SVD embeddings.



(a) Single model trained for all time

(b) Separate model trained for each decade

Figure B.5: Scatter plot for occupation proportion vs embedding bias all years of SVD embeddings, along with individually trained regression lines.

Dep. Variable:	Women Bias	**R-squared:**	0.292
Model:	OLS	**Adj. R-squared:**	0.290
Method:	Least Squares	**F-statistic:**	261.7
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	1.46e-49
Time:	17:44:46	**Log-Likelihood:**	1386.6
No. Observations:	638	**AIC:**	-2769.
Df Residuals:	636	**BIC:**	-2760.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.98e-05	16.178	0.000	0.000	0.000
const	-0.0581	0.001	-40.112	0.000	-0.061	-0.055

Omnibus:	37.968	**Durbin-Watson:**	1.608
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	57.798
Skew:	0.461	**Prob(JB):**	2.81e-13
Kurtosis:	4.151	**Cond. No.**	97.4

Table B.19: Single model for Embedding bias vs occupation percentage difference across years

Next, here is the regression with an individual term for each year.

Dep. Variable:	Women Bias	**R-squared:**	0.353
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	38.02
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	6.19e-54
Time:	17:44:46	**Log-Likelihood:**	1415.4
No. Observations:	638	**AIC:**	-2811.
Df Residuals:	628	**BIC:**	-2766.
Df Model:	9		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.94e-05	15.873	0.000	0.000	0.000
const	-0.0528	0.001	-41.662	0.000	-0.055	-0.050
yr_1910.0	-0.0036	0.003	-1.218	0.224	-0.009	0.002
yr_1920.0	-0.0067	0.003	-2.277	0.023	-0.012	-0.001
yr_1930.0	-0.0044	0.003	-1.511	0.131	-0.010	0.001
yr_1940.0	-0.0046	0.003	-1.444	0.149	-0.011	0.002
yr_1950.0	-0.0150	0.003	-5.066	0.000	-0.021	-0.009
yr_1960.0	-0.0204	0.003	-6.924	0.000	-0.026	-0.015
yr_1970.0	-0.0088	0.003	-2.998	0.003	-0.015	-0.003
yr_1980.0	0.0019	0.003	0.645	0.519	-0.004	0.008
yr_1990.0	0.0089	0.003	2.952	0.003	0.003	0.015

Omnibus:	74.864	**Durbin-Watson:**	1.536
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	152.277
Skew:	0.687	**Prob(JB):**	8.58e-34
Kurtosis:	4.960	**Cond. No.**	8.37e+17

Table B.20: Embedding bias vs occupation percentage difference with additional term for each year

Next, we show the performance and model values when models are trained separately for each embedding year.

Model Year	r^2	coefficient p-value	coefficient value	intercept p-value	intercept value
1910	0.252	$6.057e-06$	$0.0002816 \pm 5.759e-05$	$2.572e-19$	-0.05791 ± 0.004697
1920	0.1995	$7.468e-05$	$0.000261 \pm 6.205e-05$	$1.333e-19$	-0.06195 ± 0.004958
1930	0.2769	$1.471e-06$	$0.000308 \pm 5.866e-05$	$9.807e-20$	-0.05716 ± 0.004568
1940	0.3064	$3.071e-06$	$0.0003647 \pm 7.084e-05$	$1.026e-13$	-0.05381 ± 0.005612
1950	0.3167	$2.678e-07$	$0.0003328 \pm 5.843e-05$	$6.926e-24$	-0.06643 ± 0.004367
1960	0.3159	$2.78e-07$	$0.0003227 \pm 5.675e-05$	$2.783e-27$	-0.07251 ± 0.004141
1970	0.3701	$1.441e-08$	$0.0003472 \pm 5.414e-05$	$8.662e-25$	-0.05991 ± 0.003791
1980	0.3071	$6.469e-07$	$0.0002819 \pm 5.136e-05$	$7.529e-25$	-0.05174 ± 0.003219
1990	0.2841	$2.029e-06$	$0.0002859 \pm 5.502e-05$	$6.646e-22$	-0.04449 ± 0.00315

Table B.21: Performances of models trained separately for each decade.

Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation).

Year	MSE using own model	MSE using model from other years
1910	0.0007251	0.0007358
1920	0.0009599	0.0009724
1930	0.0007379	0.000741
1940	0.0007324	0.0007397
1950	0.000687	0.0007683
1960	0.0007269	0.0009379
1970	0.0006279	0.000639
1980	0.0004824	0.0005458
1990	0.0005174	0.0007325

Table B.22: Comparative performance of model trained using a given year's data versus one using all other years

B.3.2 Average Occupation bias over time

The following is the analogue of Figure 2 with the SVD embeddings.

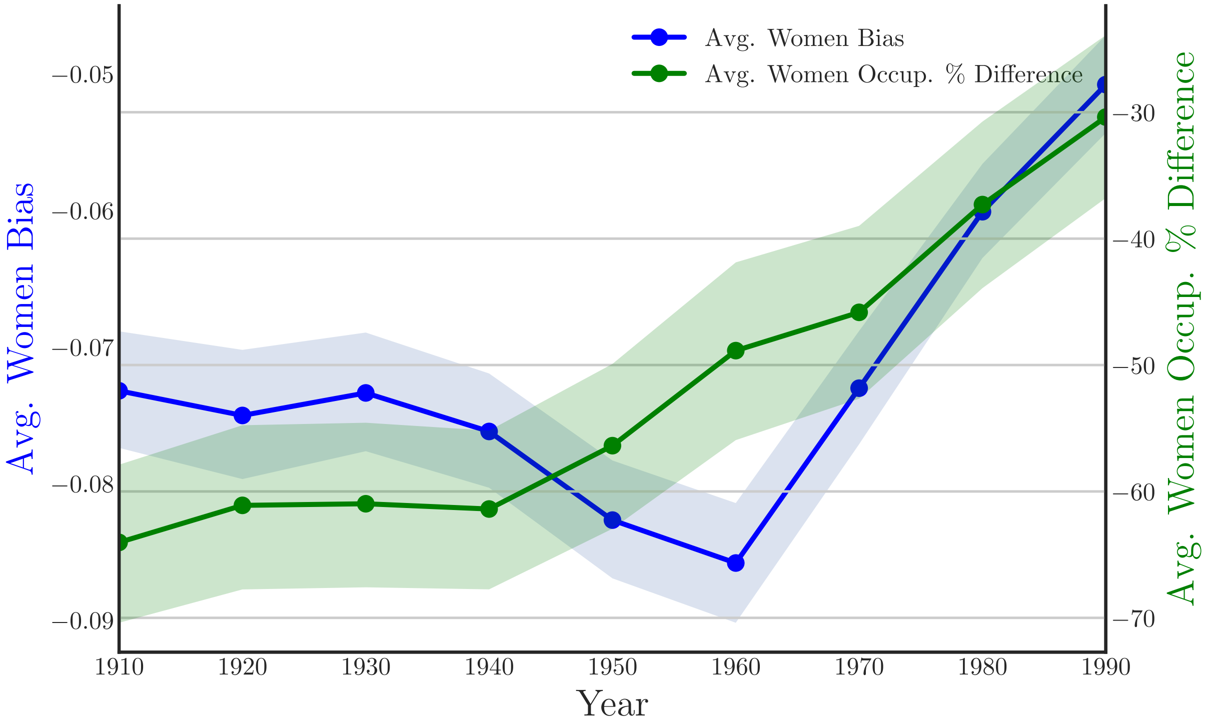


Figure B.6: Gender bias over time in COHA dataset in occupations vs the average relative percentage. In blue is the woman bias in the SVD embeddings, while the in green is the average relative percentage women in the occupation.

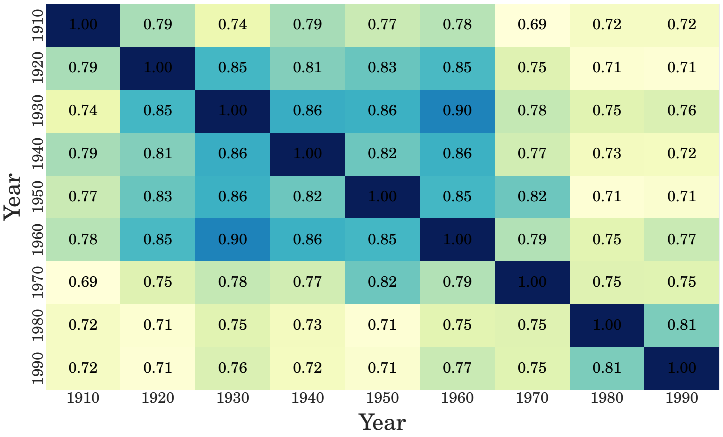
B.3.3 Cross-time correlation plots

Here, we give supporting information for the cross time correlation plot for gender bias over time, Figure 4. After describing and reporting results from a statistical test for phase shifts, we provide the same plot but with occupations, and with SVD embeddings.

We first run a test to determine whether the changes in how adjectives are associated with women significantly change between corresponding decades. In particular, we do the following: from the correlation heatmap in the figure, we take the differences between each adjacent column (excluding the changes between the diagonal elements and their neighbors); these differences quantify the change in the associations more robustly than just looking at the correlations between adjacent embeddings would. Then, we test whether a given set of differences is distributed differently than the rest through a Kolmogorov-Smirnov 2 sample test, which quantifies the difference between 2 empirical distribution functions; for each transition interval (difference between 2 columns), we run a K-S test between the differences in that interval and those between every other interval.

Transition Interval	Test statistic	p value
1910-1920	0.449	0.1206
1920-1930	0.3265	0.4475
1930-1940	0.449	0.1206
1940-1950	0.5306	0.03958
1950-1960	0.3469	0.3714
1960-1970	0.8571	7.173e-05
1970-1980	0.551	0.0291
1980-1990	0.3265	0.4475

Table B.23: Kolmogorov-Smirnov tests for phase change for Figure 4.



Year	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.79	0.74	0.79	0.77	0.78	0.69	0.72	0.72
1920	0.79	1.00	0.85	0.81	0.83	0.85	0.75	0.71	0.71
1930	0.74	0.85	1.00	0.86	0.86	0.90	0.78	0.75	0.76
1940	0.79	0.81	0.86	1.00	0.82	0.86	0.77	0.73	0.72
1950	0.77	0.83	0.86	0.82	1.00	0.85	0.82	0.71	0.71
1960	0.78	0.85	0.90	0.86	0.85	1.00	0.79	0.75	0.77
1970	0.69	0.75	0.78	0.77	0.82	0.79	1.00	0.75	0.75
1980	0.72	0.71	0.75	0.73	0.71	0.75	0.75	1.00	0.81
1990	0.72	0.71	0.76	0.72	0.71	0.77	0.75	0.81	1.00

Figure B.7: Pearson correlation in embedding bias scores for occupations over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.3673	0.3041
1920-1930	0.4082	0.1962
1930-1940	0.2857	0.6203
1940-1950	0.5918	0.01519
1950-1960	0.2653	0.7109
1960-1970	0.5306	0.03958
1970-1980	0.4898	0.07071
1980-1990	0.8163	0.0001858

Table B.24: Kolmogorov-Smirnov tests for phase change for Figure B.7.

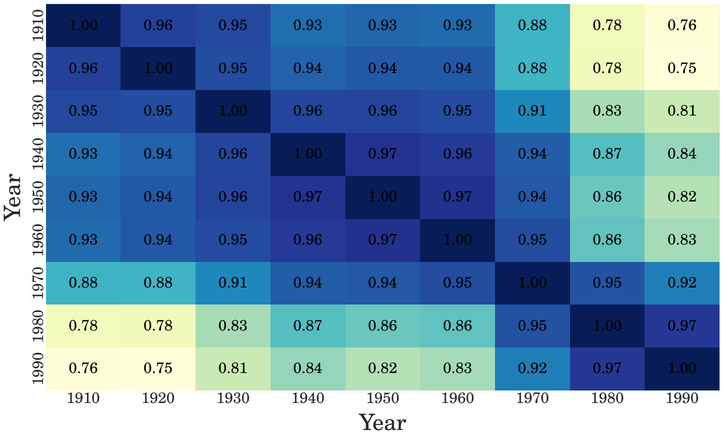


Figure B.8: Pearson correlation in SVD embedding bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.6735	0.003603
1920-1930	0.3265	0.4475
1930-1940	0.4286	0.1547
1940-1950	0.7551	0.0007097
1950-1960	0.7755	0.0004593
1960-1970	0.6939	0.002443
1970-1980	0.8776	4.38e-05
1980-1990	0.5306	0.03958

Table B.25: Kolmogorov-Smirnov tests for phase change for Figure B.8.

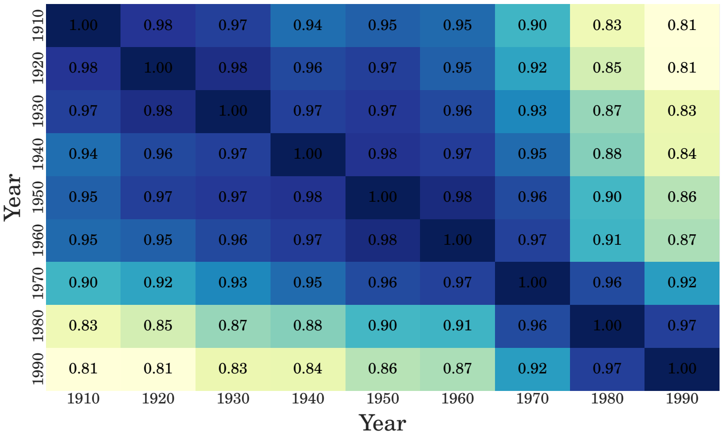


Figure B.9: Pearson correlation in SVD embedding bias scores for occupations over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.5306	0.03958
1920-1930	0.449	0.1206
1930-1940	0.449	0.1206
1940-1950	0.3673	0.3041
1950-1960	0.6122	0.01079
1960-1970	0.6122	0.01079
1970-1980	0.9592	5.423e-06
1980-1990	0.6531	0.005254

Table B.26: Kolmogorov-Smirnov tests for phase change for Figure B.9.

B.3.4 Tables with top occupations and adjectives for each gender

This subsection contains the top occupations and adjectives for each gender for each decade. We caution that due to the noisy nature of embeddings, these tables must be analyzed in the aggregate rather than focusing on individual associations.

1910	1920	1930	1940	1950	1960	1970	1980	1990
mathematician	accountant	engineer	surveyor	architect	lawyer	architect	architect	architect
soldier	surveyor	architect	architect	engineer	architect	engineer	auctioneer	mathematician
architect	architect	lawyer	engineer	mathematician	surveyor	judge	judge	surveyor
surveyor	lawyer	surveyor	smith	lawyer	soldier	economist	surveyor	engineer
administrator	mathematician	manager	sheriff	sheriff	engineer	soldier	sheriff	pilot
lawyer	sheriff	pilot	lawyer	postmaster	pilot	author	author	lawyer
judge	engineer	author	scientist	surveyor	scientist	surveyor	engineer	author
scientist	statistician	scientist	author	scientist	economist	administrator	broker	judge
author	mason	mathematician	economist	author	author	mason	inspector	soldier
economist	scientist	accountant	mason	soldier	mason	mathematician	police	blacksmith

Table B.27: Most Man occupations in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse
attendant	housekeeper	housekeeper	attendant	housekeeper	attendant	dancer	dancer	housekeeper
housekeeper	attendant	attendant	janitor	attendant	dancer	housekeeper	attendant	midwife
cashier	dancer	dancer	housekeeper	dancer	housekeeper	attendant	housekeeper	dentist
cook	teacher	janitor	midwife	cook	photographer	conductor	midwife	student
bailiff	supervisor	midwife	dentist	gardener	midwife	dentist	statistician	dancer
porter	cook	clerical	cook	cashier	dentist	statistician	student	supervisor
operator	doctor	dentist	clerical	midwife	janitor	baker	conductor	bailiff
supervisor	dentist	cook	clergy	musician	cook	clerical	dentist	physician
clergy	mechanic	teacher	sailor	sailor	porter	sailor	supervisor	doctor

Table B.28: Most Woman occupations in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
honorable	regimental	honorable	honorable	knowledge	gallant	honorable	honorable	honorable
gallant	honorable	trusting	conservative	gallant	honorable	wise	loyal	regimental
regimental	stoic	courageous	ambitious	honorable	sage	knowledge	petty	unreliable
skillful	political	gallant	shrewd	directed	regimental	gallant	gallant	skillful
disobedient	sage	confident	regimental	regimental	knowledge	insulting	lyrical	gallant
faithful	ambitious	adventurous	knowledge	efficient	wise	trusting	honest	honest
wise	reserved	experimental	destructive	sage	conservative	honest	faithful	loyal
obedient	progressive	efficient	misguided	wise	honest	providential	obedient	wise
obnoxious	unprincipled	predatory	gallant	faithful	adventurous	modern	wise	directed
steadfast	shrewd	modern	petty	creative	efficient	regimental	hostile	courageous

Table B.29: Most Man adjectives in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
charming	charming	charming	delicate	delicate	sweet	attractive	maternal	maternal
placid	relaxed	delicate	placid	sweet	charming	maternal	attractive	morbid
delicate	delicate	soft	sweet	charming	soft	charming	masculine	artificial
passionate	amiable	hysterical	gentle	transparent	relaxed	sweet	impassive	physical
sweet	hysterical	transparent	soft	placid	attractive	caring	emotional	caring
dreamy	placid	sweet	warm	childish	placid	venomous	protective	emotional
indulgent	soft	relaxed	charming	soft	delicate	silly	relaxed	protective
playful	gentle	shy	childish	colorless	maternal	neat	charming	attractive
mellow	attractive	maternal	irritable	tasteless	indulgent	delicate	naive	soft
sentimental	sweet	smooth	maternal	agreeable	gentle	sensitive	responsive	tidy

Table B.30: Most Woman adjectives in each decade in the SGNS embeddings.

B.3.5 Trend analysis for competence and physical related adjectives

The following contains statistical tests associated with Section *Individual words whose biases changed over time*.

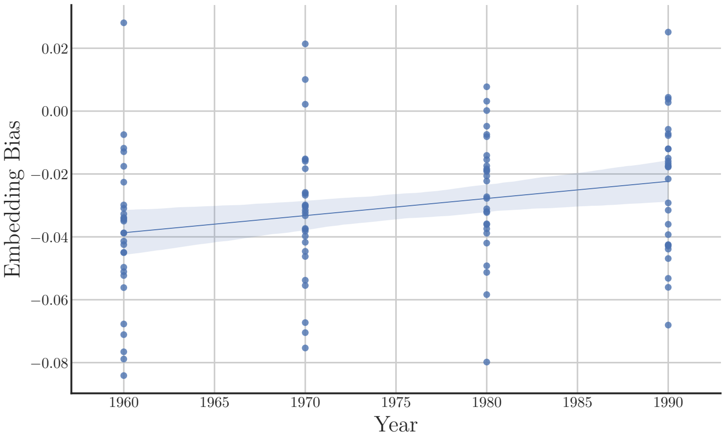


Figure B.10: Embedding bias of competence related words by decade after 1960.

Dep. Variable:	Embedding Bias	**R-squared:**	0.074
Model:	OLS	**Adj. R-squared:**	0.066
Method:	Least Squares	**F-statistic:**	8.844
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	0.00361
Time:	23:45:30	**Log-Likelihood:**	271.25
No. Observations:	112	**AIC:**	-538.5
Df Residuals:	110	**BIC:**	-533.1
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Year	0.0005	0.000	2.974	0.004	0.000	0.001
const	-1.1062	0.362	-3.058	0.003	-1.823	-0.389

Omnibus:	1.434	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.488	**Jarque-Bera (JB):**	0.943
Skew:	0.092	**Prob(JB):**	0.624
Kurtosis:	3.410	**Cond. No.**	3.49e+05

Table B.31: Regression table associated with Figure B.10.

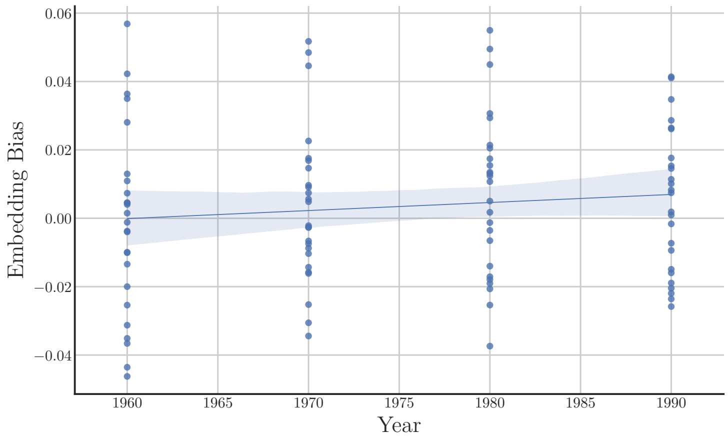


Figure B.11: Embedding bias of appearance related words by decade after 1960.

Dep. Variable:	Embedding Bias	**R-squared:**	0.013
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	1.247
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	0.267
Time:	23:45:32	**Log-Likelihood:**	233.94
No. Observations:	100	**AIC:**	-463.9
Df Residuals:	98	**BIC:**	-458.7
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Year	0.0002	0.000	1.117	0.267	-0.000	0.001
const	-0.4612	0.416	-1.108	0.270	-1.287	0.365

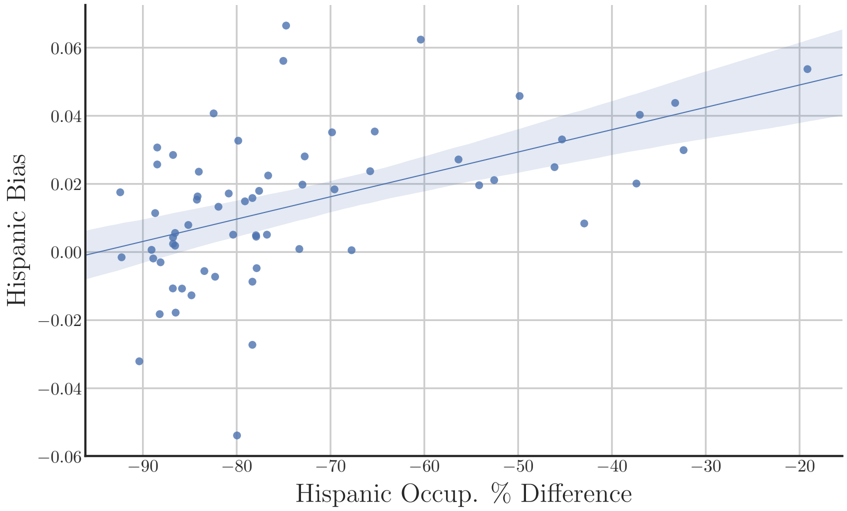
Omnibus:	2.269	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.322	**Jarque-Bera (JB):**	2.208
Skew:	0.296	**Prob(JB):**	0.332
Kurtosis:	2.577	**Cond. No.**	3.49e+05

Table B.32: Regression table associated with Figure B.11.

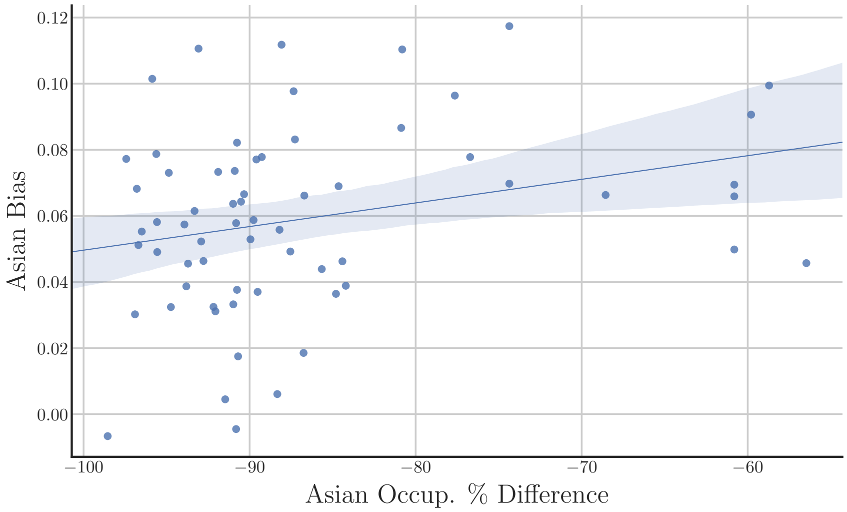
Appendix C Ethnic groups

This section provides additional information and figures related to Section *Quantifying ethnic stereotypes*.

C.1 Snapshot Analysis



(a) Percent difference of Hispanics (compared to Whites) in an occupation vs relative norm distance from occupations to the respective gender words in Google News vectors. More positive indicates more Hispanic associated, for both proportion of occupations and for relative distance. $p < 10^{-5}$ and with r-squared= .279.



(b) Percent difference of Asians (compared to Whites) in an occupation vs relative norm distance from occupations to the respective gender words in Google News vectors. More positive indicates more Asian associated, for both proportion of occupations and for relative distance. $p = .041$ and with r-squared= .065.

Dep. Variable:	Hispanic Bias	**R-squared:**	0.279
Model:	OLS	**Adj. R-squared:**	0.267
Method:	Least Squares	**F-statistic:**	23.93
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	7.42e-06
Time:	23:47:48	**Log-Likelihood:**	163.37
No. Observations:	64	**AIC:**	-322.7
Df Residuals:	62	**BIC:**	-318.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Hispanic Occup. % Difference	0.0006	0.000	4.892	0.000	0.000	0.001
const	0.0517	0.008	6.380	0.000	0.036	0.068

Omnibus:	4.172	**Durbin-Watson:**	2.261
Prob(Omnibus):	0.124	**Jarque-Bera (JB):**	4.783
Skew:	-0.065	**Prob(JB):**	0.0915
Kurtosis:	4.333	**Cond. No.**	239.

Table C.1: Regression table corresponding to Figure C.1a

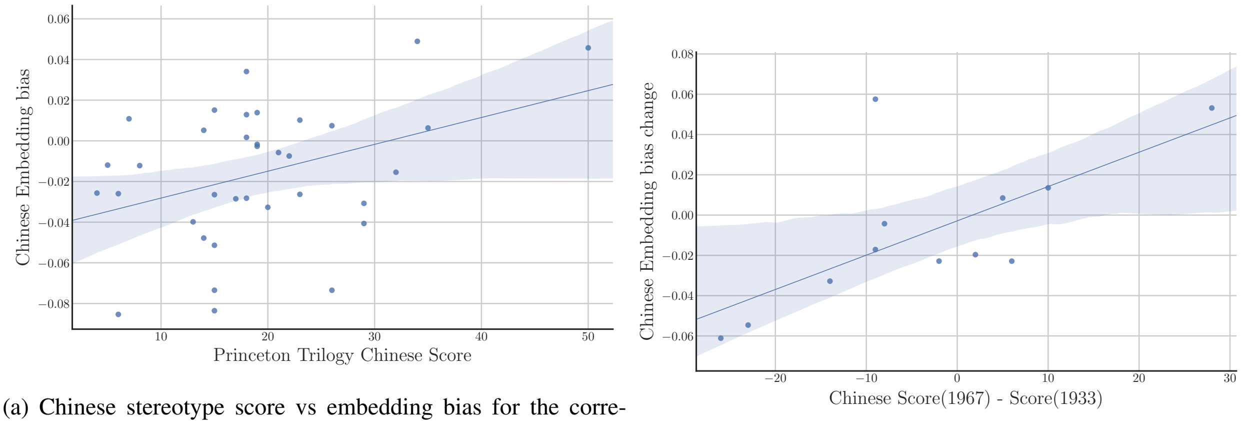
Dep. Variable:	Asian Bias	**R-squared:**	0.065
Model:	OLS	**Adj. R-squared:**	0.050
Method:	Least Squares	**F-statistic:**	4.338
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.0414
Time:	23:47:45	**Log-Likelihood:**	141.23
No. Observations:	64	**AIC:**	-278.5
Df Residuals:	62	**BIC:**	-274.1
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Asian Occup. % Difference	0.0006	0.000	2.083	0.041	2.41e-05	0.001
const	0.1087	0.024	4.521	0.000	0.061	0.157

Omnibus:	0.103	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.950	**Jarque-Bera (JB):**	0.157
Skew:	-0.089	**Prob(JB):**	0.925
Kurtosis:	2.834	**Cond. No.**	594.

Table C.2: Regression table corresponding to Figure C.1b

C.2 Princeton Trilogy



(a) Chinese stereotype score vs embedding bias for the corresponding decade across all three trilogy studies. All stereotypes (which are present in the embeddings) in the Chinese portion of Table 1 in [10] are included.

(b) Change in stereotypes between 1933 and 1969

Figure C.2: Plots associated with Princeton Trilogy validation. Full word lists are in Section A.2.

Dep. Variable:	Chinese Embedding bias	**R-squared:**	0.146
Model:	OLS	**Adj. R-squared:**	0.120
Method:	Least Squares	**F-statistic:**	5.644
Date:	Fri, 23 Feb 2018	**Prob (F-statistic):**	0.0235
Time:	21:24:48	**Log-Likelihood:**	72.702
No. Observations:	35	**AIC:**	-141.4
Df Residuals:	33	**BIC:**	-138.3
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Princeton Trilogy Chinese Score	0.0013	0.001	2.376	0.023	0.000	0.002
const	-0.0413	0.012	-3.489	0.001	-0.065	-0.017

Omnibus:	1.808	**Durbin-Watson:**	1.846
Prob(Omnibus):	0.405	**Jarque-Bera (JB):**	1.693
Skew:	-0.468	**Prob(JB):**	0.429
Kurtosis:	2.465	**Cond. No.**	48.0

Table C.3: Regression Table associated with Figure C.2a.

Dep. Variable:	Chinese Embedding bias change	**R-squared:**	0.472
Model:	OLS	**Adj. R-squared:**	0.419
Method:	Least Squares	**F-statistic:**	8.931
Date:	Fri, 23 Feb 2018	**Prob (F-statistic):**	0.0136
Time:	21:24:47	**Log-Likelihood:**	26.887
No. Observations:	12	**AIC:**	-49.77
Df Residuals:	10	**BIC:**	-48.80
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Chinese Score(1967) - Score(1933)	0.0017	0.001	2.988	0.014	0.000	0.003
const	-0.0029	0.008	-0.343	0.739	-0.021	0.016

Omnibus:	18.181	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13.350
Skew:	1.914	**Prob(JB):**	0.00126
Kurtosis:	6.471	**Cond. No.**	15.1

Table C.4: Regression Table associated with Figure C.2b.

C.3 Dynamic Analysis

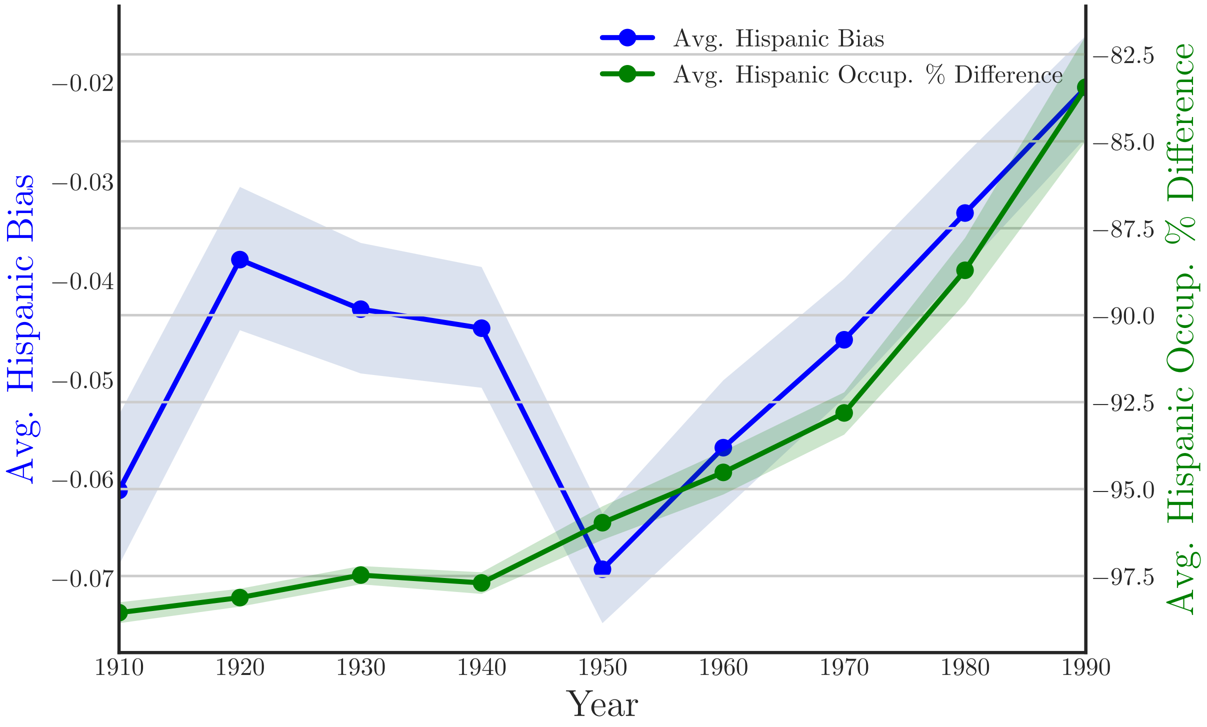


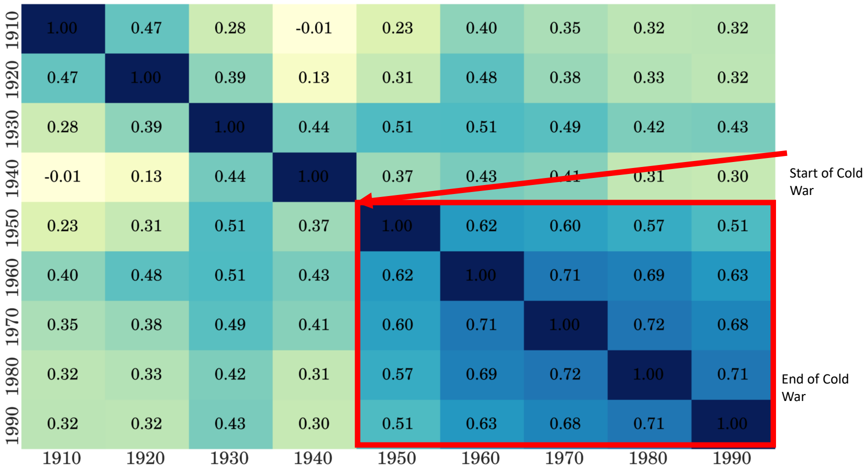
Figure C.3: Ethnic (Hispanic vs White) bias over time in COHA dataset in occupations vs the average percent difference. In blue is the relative Hispanic bias in the SGNS embeddings, while in green is the average percent difference of Hispanics in each occupation.

C.4 Cross-time Correlation plots

Here, we give supporting information for the cross time correlation plot for Asian bias over time. We first provide the results of the KolmogorovSmirnov tests for Figure 5 (as we performed and described in Appendix Section B.3.3). We then show the same plots and test results for Russian and Hispanic associations, respectively.

Transition Interval	Test Statistic	p-value
1910-1920	0.2653	0.7109
1920-1930	0.3469	0.3714
1930-1940	0.3265	0.4475
1940-1950	0.2653	0.7109
1950-1960	0.6122	0.01079
1960-1970	0.3673	0.3041
1970-1980	0.7551	0.0007097
1980-1990	0.4898	0.07071

Table C.5: KolmogorovSmirnov tests for phase change for Figure 5.



	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.47	0.28	-0.01	0.23	0.40	0.35	0.32	0.32
1920	0.47	1.00	0.39	0.13	0.31	0.48	0.38	0.33	0.32
1930	0.28	0.39	1.00	0.44	0.51	0.51	0.49	0.42	0.43
1940	-0.01	0.13	0.44	1.00	0.37	0.43	0.41	0.31	0.30
1950	0.23	0.31	0.51	0.37	1.00	0.62	0.60	0.57	0.51
1960	0.40	0.48	0.51	0.43	0.62	1.00	0.71	0.69	0.63
1970	0.35	0.38	0.49	0.41	0.60	0.71	1.00	0.72	0.68
1980	0.32	0.33	0.42	0.31	0.57	0.69	0.72	1.00	0.71
1990	0.32	0.32	0.43	0.30	0.51	0.63	0.68	0.71	1.00

Start of Cold War

End of Cold War

Figure C.4: Pearson correlation in SGNS embedding Russian bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.2653	0.7109
1920-1930	0.5102	0.05321
1930-1940	0.5102	0.05321
1940-1950	0.7347	0.001084
1950-1960	0.4286	0.1547
1960-1970	0.5102	0.05321
1970-1980	0.5306	0.03958
1980-1990	0.7143	0.001637

Table C.6: KolmogorovSmirnov tests for phase change for Figure C.4.

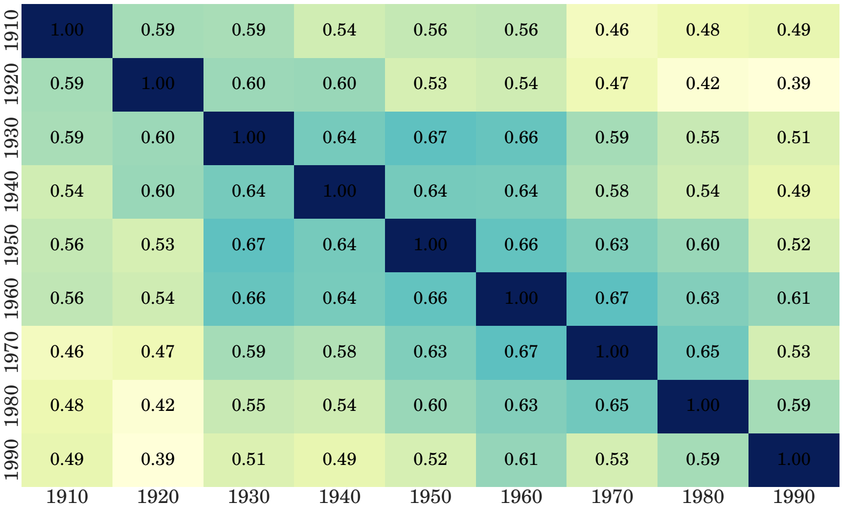


Figure C.5: Pearson correlation in SGNS embedding Hispanic bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.1837	0.9729
1920-1930	0.7143	0.001637
1930-1940	0.5102	0.05321
1940-1950	0.2653	0.7109
1950-1960	0.5102	0.05321
1960-1970	0.4286	0.1547
1970-1980	0.3265	0.4475
1980-1990	0.1429	0.9989

Table C.7: KolmogorovSmirnov tests for phase change for Figure C.5.

1910	1920	1930	1940	1950	1960	1970	1980	1990
irresponsible	mellow	hateful	solemn	disorganized	imprudent	cynical	superstitious	inhibited
envious	relaxed	unchanging	reactive	outrageous	pedantic	solemn	upright	passive
barbaric	haughty	oppressed	outrageous	pompous	irresponsible	mellow	providential	dissolute
aggressive	tense	contemptible	bizarre	unstable	inoffensive	discontented	unstable	haughty
transparent	hateful	steadfast	fanatical	effeminate	sensual	dogmatic	forceful	complacent
monstrous	venomous	relaxed	assertive	unprincipled	venomous	aloof	appreciative	forceful
hateful	stubborn	cruel	unprincipled	venomous	active	forgetful	dry	fixed
cruel	pedantic	disorganized	barbaric	disobedient	inert	dominating	reactive	active
greedy	transparent	brutal	haughty	predatory	callous	disconcerting	fixed	sensitive
bizarre	compassionate	intolerant	disconcerting	boisterous	inhibited	inhibited	sensitive	hearty

Table C.8: Top Asian (vs White) Adjectives over time by relative norm difference.

References

[1] Parker R, Graff D, Kong J, Chen K, Maeda K (2011) English Gigaword Fifth Edition LDC2011t07.

[2] Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds. Moschitti A, Pang B. (Association for Computational Linguistics, Doha, Qatar), pp. 1532–1543.

[3] Chalabi M, Flowers A (2014) Dear Mona, Whats The Most Common Name In America?

[4] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings in *Advances in Neural Information Processing Systems 29*, eds. Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R. (Curran Associates, Inc.), pp. 4349–4357.

[5] Williams JE, Best DL (1977) Sex Stereotypes and Trait Favorability on the Adjective Check List. *Educational and Psychological Measurement* 37(1):101–110.

[6] Williams JE, Best DL (1990) *Measuring sex stereotypes: A multination study, Rev. ed.*, Measuring sex stereotypes: A multination study, Rev. ed. (Sage Publications, Inc, Thousand Oaks, CA, US).

[7] Gunkel P (1987) 638 Primary Personality Traits.

[8] Katz D, Braly K (1933) Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology* 28(3):280.

[9] Gilbert GM (1951) Stereotype persistence and change among college students. *The Journal of Abnormal and Social Psychology* 46(2):245.

[10] Karlins M, Coffman TL, Walters G (1969) On the fading of social stereotypes: Studies in three generations of college students. *Journal of personality and social psychology* 13(1):1.

[11] Levanon A, England P, Allison P (2009) Occupational Feminization and Pay: Assessing Causal Dynamics Using 19502000 U.S. Census Data. *Social Forces* 88(2):865–891.