AMERICAN SOCIOLOGICAL REVIEW

OFFICIAL JOURNAL OF THE AMERICAN SOCIOLOGICAL ASSOCIATION

ONLINE SUPPLEMENT
to article in
AMERICAN SOCIOLOGICAL REVIEW, 2014, VOL. 79

Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004

Jeffrey A. Smith
University of Nebraska-Lincoln

Miller McPherson
Duke University

Lynn Smith-Lovin
Duke University

Part A. Alternative Model of Absolute Homophily Change

This supplement presents a formal test of absolute homophily change. The raw homophily results are presented as a simple logistic regression. Units in the analysis are the respondent-confidant pairs from 1985 and 2004. The dependent variable is year, equal to 1 if the dyad is from 2004, and 0 if the dyad is from 1985. The independent variables are the sociodemographic distances between the respondent and the named confidant. The models thus predict the probability of a respondent-confidant pair appearing in 1985 versus 2004, as a function of sociodemographic distance. A positive coefficient suggests demographically distant confidants are more likely to exist in 2004 than in 1985. The model does not take into account the sociodemographic distance expected by chance, and simply captures the demographic similarity between confidants over time. The model differs from Table 1 in the main text because it conditions the change in one dimension on changes in another. It also provides results using all confidants and only non-kin confidants. The respondent-confidant pairs are nested within respondents, and we take these dependencies into account when calculating the standard errors. Specifically, we adjust for the complex survey design of the data when running the glm.

Table S1. Logistic Regression Predicting Year as a Function of Demographic Distance

	All Ties	Non-kin Ties
Intercept	–.545***	–.603***
	(.082)	(.110)
Different Race	.730***	.636**
	(.171)	(.195)
Different Religion	.202*	.188
	(.087)	(.119)
Different Sex	.128*	–.031
	(.056)	(.113)
Age Difference	–.002	–.009
	(.003)	(.007)
Education Difference	–.015	.003
	(.017)	(.027)
N	6,515	2,806

Note: Standard errors are in parentheses. Units in the analysis are respondent-confidant pairs nested within respondents. The estimation routine accounts for the dependence in the data when producing the standard errors.
**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

Part B. Case-Control Results Using Randomly Selected Confidant

Table S2. Case-Control Logistic Regression for Randomly Selected Confidant, Univariate Analysis

Variable	Intercept	Dimension	Year	Dimension x Year	*N* (dyads)
All Ties					
Different Race	–16.894***	–1.966***	–.132**	.289	1,134,804
	(.029)	(.143)	(.043)	(.192)	
Different Religion	–16.652***	–1.389***	.000	–.245	1,134,804
	(.035)	(.092)	(.049)	(.129)	
Different Sex	–17.276***	–.265***	–.349***	.233*	1,134,804
	(.038)	(.072)	(.058)	(.112)	
Age Difference	–16.330***	–.053***	–.219**	–.005	1,134,804
	(.050)	(.004)	(.076)	(.006)	
Education Difference	–16.749***	–.188***	–.178**	–.037	1,134,804
	(.046)	(.017)	(.062)	(.026)	
Non-kin Ties					
Different Race	–17.630***	–1.711***	–.264***	.176	440,756
	(.057)	(.152)	(.062)	(.216)	
Different Religion	–17.523***	–.971***	–.119	–.328*	440,756
	(.064)	(.097)	(.078)	(.149)	
Different Sex	–17.647***	–.861***	–.446***	.167	440,756
	(.060)	(.096)	(.065)	(.152)	
Age Difference	–16.380***	–.091***	–.210*	–.020	440,756
	(.072)	(.006)	(.100)	(.011)	
Education Difference	–17.340***	–.219***	–.371***	–.017	440,756
	(.071)	(.022)	(.084)	(.036)	

Note: Standard errors are in parentheses; they were calculated using bootstrap estimates. Standard errors are equal to the standard deviation of the coefficients across 1,000 iterations and are thus not dependent on the number of dyads. For each iteration, we took a random sample of respondents from each year and reran the case-control logistic regression.

**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

Table S3. Case-Control Logistic Regression for Randomly Selected Confidant, Multivariate Analysis

	All Ties		Non-kin Ties	
Variables	Model 1	Model 2	Model 3	Model 4
Intercept	–14.395***	–14.470***	–14.226***	–14.423***
	(.065)	(.080)	(.086)	(.103)
Different Race	–1.724***	–1.893***	–1.558***	–1.608***
	(.096)	(.146)	(.111)	(.155)
Different Religion	–1.468***	–1.374***	–1.078***	–.969***
	(.065)	(.092)	(.073)	(.099)
Different Sex	–.148**	–.252***	–.776***	–.844***
	(.057)	(.073)	(.077)	(.097)
Age Difference	–.054***	–.052***	–.097***	–.088***
	(.003)	(.004)	(.005)	(.006)
Education Difference	–.175***	–.151***	–.197***	–.183***
	(.014)	(.019)	(.019)	(.023)
Different Race x Year		.292		.089
		(.196)		(.219)
Different Religion x Year		–.198		–.254
		(.131)		(.153)
Different Sex x Year		.236*		.167
		(.115)		(.155)
Age Difference x Year		–.005		–.021*
		(.006)		(.011)
Education Difference x Year		–.059*		–.038
		(.027)		(.037)
Year	.026	.127	–.081	.184
	(.046)	(.119)	(.064)	(.151)
N (respondents)	3001	3001	3001	3001
N (dyads)	1,134,804	1,134,804	440,756	440,756
–2 x Log-likelihood	29260.14	29229.34	16482.39	16460.78
AIC	29274.14	29253.34	16496.39	16484.78
BIC (*N* based on dyads)	29357.73	29396.64	16573.36	16616.73

Note: Standard errors are in parentheses; they were calculated using bootstrap estimates. Standard errors are equal to the standard deviation of the coefficients across 1,000 iterations and are thus not dependent on the number of dyads. For each iteration, we took a random sample of respondents from each year and reran the case-control logistic regression.

**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

Part C. Results of Network Simulation Approach to Testing Homophily Change

This supplement presents the results of an alternative test to our case-control method. Here, we replicated the analysis using simulated networks as a means of constructing the "by chance" comparisons. In the main analysis, we generated our chance expectations by randomly pairing respondents in the GSS together for each year. We assumed the probability of randomly pairing two people together follows a binomial distribution with probability based on the population weights. Construction of the controls, and thus chance expectations, is only constrained on the distribution of demographic characteristics in the population. It is implicitly not constrained on (1) the volume of ties; (2) the degree distribution (i.e., ties per person); and (3) differential degree (i.e., some groups have more ties than others).

We reconsider those assumptions in this supplementary analysis. We made particular choices in measuring chance expectations; we could have made alternative choices. It is important to consider how our results would have differed under different assumptions. Such choices are easier to represent through network simulation, where one generates random networks and uses that to calculate chance expectations for homophily. The construction of the controls in the article is a particular version of this. Specifically, you can think of the random pairing process as creating a baseline network with N x (N – 1)/2 ties. The network is conditioned on the demographic composition of the population, and everyone has the same number of ties. Here, we extend the analysis to constrain the "simulated" network on edges (or volume), degree distribution, and differential degree.

Analytic Strategy

We began by taking a bootstrap sample of respondents in the GSS for 1985 and 2004. We drew the same number of respondents as in the original sample. We then generated networks for 1985 and 2004 using ERGM (exponential random graph models); specifically using the statnet package in R (Handcock et al. 2008). We began by generating networks constrained on the empirically observed degree distribution (i.e., NUMGIVEN in the data). This also implicitly constrains the baseline network on total volume. We then seeded the network with the sampled respondents. The demographic characteristics of the sampled respondents were mapped onto nodes in the network with the same degree as the respondent (see Smith 2012). This maintains the correlation between demographic characteristics and degree. Thus, highly educated people in the simulated network will have high degree if the sampled respondents with high degree are highly educated. This seeding process also ensures that the generated network will reflect the demographic composition in the data. Thus, the simulation will generate a network that represents random mixing in the population, given the degree distribution, differential degree, and the demographic composition of the population. We repeated this process for both 1985 and 2004. In each case, the simulated networks are size 10,000 (it is impossible to simulate a network of the true size, 200 million or so).

We took a sample of ego networks from the simulated network the same size as the original GSS sample for that year—thus mimicking the true sampling process. We then took all *ij* pairs from the ego networks drawn from the simulated network and calculated the demographic distance between *i* and *j* (e.g., racial or religious matching). We compared the demographic distance in the observed ego network data to the demographic distance from the simulated network, capturing chance expectations.

For race, religion, and sex (the categorical variables), we compared the odds of a tie matching demographically in the observed data to the odds of a tie matching in the simulated network. We report how many times the 1985 ratio (log($\text{odds}_{\text{observed}}/\text{odds}_{\text{chance}}$)) is larger than the 2004 ratio, indicating a decrease in in-group bias (relative to chance). This analysis mirrors a simple CUG (conditional uniform graph) test, and we report results for 1,000 bootstrap samples. We also report a second, alternative summary measure, based on the ratio of frequency counts: log((# Observed Ties Matching)/(# Observed Ties Mismatching)). This ratio is calculated net of chance expectations, based on the simulated network, and compared across 1985 and 2004. We again report how many times the 1985 ratio is larger than the 2004 ratio.

For the continuous measures, age and education, we calculated the ratio: log(# Ties $\text{Observed}_{\text{Distance=x}}$/#Ties $\text{Chance}_{\text{Distance=x}}$).The ratio compares the number of ties in the observed ego networks to the number of ties in the simulated network at a given education or age distance, x. We then see how much an increase in demographic distance lowers the ratio of observed to chance frequency counts. Larger decreases, on average, mean stronger effects of increasing demographic distance. Formally, we focus on the marginal (or average) effect of increasing demographic distance by calculating the ratio as demographic distance increases by 1 and then averaging over those marginal effects. Again, we report how many times the 1985 ratio is larger than the 2004 ratio. Absolutely larger 1985 values mean homophily decreased.

Results

The results presented here mirror the results reported in the main text. There is a significant decrease in gender homophily, where the odds ratios are larger in 1985 than in 2004. The age and education results show no statistically discernible differences across years (although both lean toward an increase in homophily, as in Table 2 in the main text). Religion shows a significant increase in homophily using one summary measure but not the other, mirroring results in the main text, which show a significant increase under some specifications but not others. The religion results remain inconsistent, while pointing to a possible increase in homophily. The only major difference is with race. Here the results indicate a possible decrease in homophily, although the odds ratio results are more inconsistent across samples. The race interaction is, however, never significant in the results reported in the main text.

The racial differences result from the conditioning on differential degree. Because Whites have more ties on average than non-Whites, White-White ties are more frequent in the controls when degree is allowed to vary across demographic groups. More generally, homophily will appear weaker (relative to chance expectations) when degree differences are taken into account. This process is somewhat more exaggerated in 2004 than in 1985. This means that more of the racial matching can be explained by degree differences in 2004; or, once one “controls” for the differences in degree by demographic group, there is a larger decrease in in-group bias for race.

Looking over all the evidence, we do not believe there has been a decrease in racial homophily relative to chance. This is the only set of results that show a decrease in racial homophily and they are highly conditioned, controlling for both differential degree and the degree distribution. Additionally, the results are particularly dependent on the degree information in the data, and we know the 2004 degree information is problematic (given the inflated number of isolates) (see also note 15 in the main text).

Table S4. Comparing Observed Homophily to Homophily in Simulated Networks, 1985 to 2004

	Homophily Measured by Odds Ratio	Homophily Measured by Frequency Ratios	Homophily Measured by Frequency Ratios: Continuous Version
	Number of Samples with Homophily Increase: 1985 > 2004		
Race	74	1	NA
Religion	999	834	NA
Sex	1	1	NA
Age	NA	NA	940
Education	NA	NA	728

Note: Values correspond to the number of bootstrap samples where there is an increase in homophily. We had a total of 1,000 samples, so a value above 975 is strong evidence for an increase in homophily. A count below 25 is strong evidence for a decrease in homophily.

Part D. Case-Control Results Including Isolates in Controls

Table S5. Case-Control Logistic Regression Using All Reported Ties, Including Isolates in Controls, Univariate Analysis

Variable	Intercept	Dimension	Year	Dimension x Year	*N* (dyads)
All Ties					
Different Race	–16.792***	–2.105***	–0.286***	.191	1,130,856
	(.023)	(.115)	(.040)	(.145)	
Different Religion	–16.726***	–1.273***	–.195***	–.272**	1,130,856
	(.029)	(.055)	(.048)	(.084)	
Different Sex	–17.209***	–0.388***	–.500***	.139**	1,130,856
	(.020)	(.029)	(.034)	(.044)	
Age Difference	–16.404***	–.051***	–.438***	–.003	1,130,856
	(.025)	(.002)	(.047)	(.003)	
Education Difference	–16.735***	–.199***	–.438***	–.013	1,130,856
	(.030)	(.009)	(.053)	(.017)	

Note: Standard errors are in parentheses; they were calculated using bootstrap estimates. Standard errors are equal to the standard deviation of the coefficients across 1,000 iterations and are thus not dependent on the number of dyads. For each iteration, we took a random sample of respondents from each year and reran the case-control logistic regression.
**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

Table S6. Case-Control Logistic Regression Using All Reported Ties, Including Isolates in Controls, Multivariate Analysis

	All Ties	
Variables	Model 1	Model 2
Intercept	–14.376***	–14.444***
	(.043)	(.050)
Different Race	–1.941***	–2.038***
	(.076)	(.119)
Different Religion	–1.355***	–1.263***
	(.043)	(.057)
Different Sex	–.325***	–.380***
	(.024)	(.030)
Age Difference	–.051***	–.049***
	(.002)	(.002)
Education Difference	–.173***	–.161***
	(.008)	(.011)
Different Race x Year		.181
		(.152)
Different Religion x Year		–.215*
		(.089)
Different Sex x Year		.142**
		(.047)
Age Difference x Year		–.005
		(.003)
Education Difference x Year		–.034
		(.019)
Year	–.158***	–.045
	(.046)	(.086)
N (respondents)	3,001	3,001
N (dyads)	1,130,856	1,130,856
–2 x Log-likelihood	72634.862	72597.424
AIC	72648.862	72621.424
BIC (*N* based on dyads)	72732.431	72764.686

Note: Standard errors are in parentheses; they were calculated using bootstrap estimates. Standard errors are equal to the standard deviation of the coefficients across 1,000 iterations and are thus not dependent on the number of dyads. For each iteration, we took a random sample of respondents from each year and reran the case-control logistic regression.

**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

Part E. Case-Control Results Using Data from 2010 GSS Survey Experiment

This supplement presents results of an alternative analysis of homophily change. In the main text, the analysis compares homophily rates from the 1985 GSS to the 2004 GSS. Past work shows the 2004 data contained a disproportionate number of isolates, or individuals claiming no close confidants (Paik and Sanchagrin 2013). The GSS embedded an experiment in the 2010 survey to undercover the source and magnitude of this bias. Individuals were asked the same ego network questions as in previous years, but were randomly assigned to three survey conditions: one mimicking the 1985 survey (where the network questions came earlier in the survey); one mimicking the 2004 data (where the network questions came later in the survey, after a battery of voluntary association questions); and one that mimicked neither the 1985 nor 2004 survey.

We exploit this experiment as a way of validating our results on an independently collected dataset. Here, we reran the analysis using the 2010 data. We limited the sample to individuals who received the 1985 survey design. This analysis does not use the 2004 data. The 2010 data are limited by small sample size and scant demographic information (the survey only asks about race and gender), but it still offers an ideal robustness check for the main results—the 2010 data are directly comparable to the 1985 data in terms of survey design.

Table S7 presents results for race and gender. The general findings are the same as with the 2004 data: there is no change in racial homophily but a decrease in gender homophily. The *racial homophily* x *year* coefficient is smaller than with the 2004 data, but our overall conclusions are not affected by the overinflation of isolates found in the 2004 data.

Table S7. Case-Control Logistic Regression Using All Reported Ties, Using 2010 GSS Instead of 2004 GSS, Multivariate Analysis

	All Ties	
Variables	Model 1	Model 2
Different Race	–1.953***	–1.934***
	(.082)	(.098)
Different Sex	–.395***	–.437***
	(.028)	(.032)
Different Race x Year		–.062
		(.175)
Different Sex x Year		.220**
		(.068)
	(.052)	(.070)
N (dyads)	169,970	169,970
–2 x Log-likelihood	44160.067	44150.019
AIC	44168.067	44162.019
BIC	44208.24	44222.279

Note: Standard errors are in parentheses; they were calculated using bootstrap estimates. Standard errors are equal to the standard deviation of the coefficients across 1,000 iterations and are thus not dependent on the number of dyads. For each iteration, we took a random sample of respondents from each year and reran the case-control logistic regression.
**p* < .05; ***p* < .01; ****p* < .001 (two-tailed tests).

References

Handcock, Mark S., Steven M. Goodreau, David R. Hunter, Carter T. Butts, and Martina Morris. 2008. "ERGM: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24:1–29.

Paik, Anthony and Kenneth Sanchagrin. 2013. "Social Isolation in America: An Artifact." *American Sociological Review* 78:339–60.

Smith, Jeffrey A. 2012. "Macrostructure from Survey Data: Generating Whole Systems from Ego Networks." *Sociological Methodology* 42:155–205.