ONLINE APPENDIX: GREENWOOD, GUNER, KOCHARKOV AND SANTOS (2014)

A1. Data

The data used for this paper is freely available from the Integrated Public Use Microdata Series (IPUMS) website. The samples used in this study are taken from the 1 percent sample of the Census for the years 1960, 1970, 1980, 1990, 2000 and the American Community Survey (ACS) for the year 2005. The following variables were included for every year: year or the survey (variable name: year), spouse location flag (sploc), number of family members in the household (famsize), number of children in the household (nchild), age (age), sex (sex), marital status (marst), educational attainment (educ), employment status (empstat), total family income (ftotinc), wage and salary income (incwage). Only singles and married couples that are 25 to 54 years old are considered. The adults in these households either live by themselves or with their children, who are less than 19 years old. Households in which there are other members such as grandparents, uncles/aunts, or other unrelated individuals are excluded. Households with subfamilies of any other type are also excluded from the analysis. Finally, widows, widowers and married individuals whose spouses are absent are excluded as well. Income variables are restricted to be non-negative.

There are 560 types of households used in the analysis. Households are broken down into finer categories than are reported in the text. In principle, this doesn't affect the analysis, since the finer classifications can be combined to attain the more aggregated ones. Following a counterfactual experiment, some households are moved into new income percentiles. So, in practice the finer classification allows more accurate re-sorting into the various income percentile when conducting the counterfactual experiments. Households are classified into different types as follows:

1) Marital status: married, never married males, never married females, divorced males, divorced females.

2) Education: less than high school, high school, some college, college, more than college. For married households, both the husband and wife will have one of these educational levels.

3) Market work: work, does not work. For married households both the husband and wife will have one of these levels of labor market activity.

4) Children: no children, 1 child, 2 children, more than 2 children.

Finally, households are divided into 10 deciles. So, for every year, there are 5,600 (i, j)-combinations of household types/deciles.

A2. The Lorenz Curve and Gini Coefficient

Think of a sample of different household types, $i \in \{1, 2, ..., m\}$, situated in different percentiles, $j \in \mathcal{J}$, of the income distribution. Again, j is expressed as a fraction. Define f_{ij} as the fraction of households that are of type-i in income percentile j. Let r_{ij} represent household (i, j)'s income, y_{ij}, *relative* to mean income, y. Each household's income is adjusted to a per-adult-equivalent basis using the OECD modified equivalence scale, which counts the first adult as 1, the second adult as 0.5 and each child as 0.3 adults. Equivalized households incomes are then divided by mean household income across the whole sample.

The share of income earned by percentile j is

$$s_j = \sum_i f_{ij} r_{ij}.$$

The Lorenz curve is derived by plotting the cumulative shares of the population indexed by percentile p,

$$p = \sum_j^p \sum_i^m f_{ij}$$

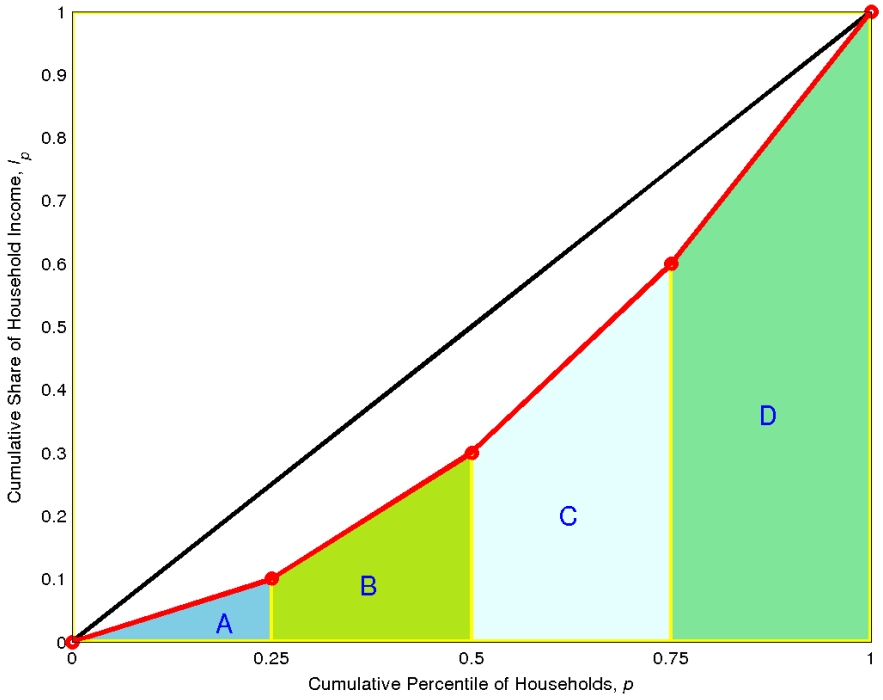

FIGURE A1. THE LORENZ CURVE AND GINI COEFFICIENT

Note: The figure shows the construction of a Lorenz curve when there are four percentiles (quartiles). The Gini coefficient is twice the area between the 45 degree line and the Lorenz curve.

on the x-axis, against the cumulative share of income indexed by percentile p,

$$l_p = \sum_{j}^{p} s_j,$$

on the y-axis. Suppose that the unit interval is split up into n equally sized segments. Then, $j \in \mathcal{J} = \{1/n, \cdots, 1 - 1/n, 1\}$.

Take the example of $n = 4$ (quartiles). The Lorenz curve described above is plotted in Figure 1. The Gini coefficient associated with the Lorenz curve equals twice the area between the Lorenz curve and the 45-degree line. Alternatively, the coefficient can be calculated as equaling $1 - 2\Delta$, where Δ is the area below the Lorenz Curve. In the case of quartiles the area Δ is the summation of the areas of the right triangle A, the right trapezoids B, C, and D. The coordinates on the x-axis are given by 0, $p_1 = 0.25$, $p_2 = 0.5$, $p_3 = 0.75$, and 1.0. The y-axis coordinates of the Lorenz curve are given by 0, $l_{0.25}$, $l_{0.5}$, $l_{0.75}$, and 1.0.

Then, using the formulas for the geometric areas A, B, C, and D, the Gini coefficient, g, can be derived as

$$g = 1 - 2\left(\underbrace{\frac{p_1 l_1}{2}}_{\text{Area } A} + \underbrace{\frac{(l_1 + l_2)(p_2 - p_1)}{2}}_{\text{Area } B} + \underbrace{\frac{(l_2 + l_3)(p_3 - p_2)}{2}}_{\text{Area } C} + \underbrace{\frac{(l_3 + 1)(1 - p_3)}{2}}_{\text{Area } D}\right).$$

After rearranging and canceling out terms, the expression for the Gini coefficient can be simplified to

$$g = (p_1 l_2 - p_2 l_1) + (p_2 l_3 - p_3 l_2) + (p_3 - l_3).$$

The cumulative shares of the population, the p's, are based on quartiles; i.e., $p_1 = 1/4$, $p_2 = 2/4, \cdots$. Thus, above expression can be rewritten as

$$g = \frac{1}{4}\left[(l_2 - 2l_1) + (2l_3 - 3l_2) + (3 - 4l_3)\right].$$

In the more general case of n percentiles, the Gini coefficient equals

$$g = \sum_{p=1/n}^{1-1/n} [p l_{p+1} - (p + 1/n) l_p].$$

The version of this formula for an arbitrary number of income groups of any size and an arbitrary number of sub-populations (types) is presented in Rao (1969).

A3. Counterfactual Experiments

IMPOSING RANDOM MATCHING

Random matching can be imposed on the demographic structure of the U.S. population for each of these years in the sample. Table A1 presents the contingency tables that would occur if matching had been random in 1960 and 2005. Counterfactual Gini coefficients can then be computed. How is this done?

First a bit of notation. Take the distribution of household, $\{f_{ij}\}$. Recall that married households are indexed by the education of the husband, the education of the wife, their labor-force participation, and the number of children in the household. Let the sets $\mathcal{M}_{E_H}$ contain the indices of all married households with a husband who has the educational level, $E_H \in \{HS-, HS, C-, C, C+\}$, where $HS-$ refers to a less-than-high-school educated person, HS refers to someone with a high-school education, $C-$ is some college, C is college, and $C+$ is more-than-college educated. Similarly, the sets $\mathcal{M}_{E_W}$ contain married households with different educational levels for wives, E_W. Furthermore, $\mathcal{M}_{LFP_H}$ ($\mathcal{M}_{LFP_W}$) contain all the married households with a husband's (wife's) labor-force participation status $LFP_{H(W)} \in \{WORK_{H(W)}, \sim WORK_{H(W)}\}$. Finally, the set $\mathcal{M}_{KIDS}$ contain married households with a particular number of children $KIDS \in \{0, 1, 2, 2+\}$. The set of all married households with a particular mix of the education, $\mathcal{M}_{E_H,E_W}$, for the husband and the wife reads

$$\mathcal{M}_{E_H,E_W} = \mathcal{M}_{E_H} \cap \mathcal{M}_{E_W}.$$

Let $\mathcal{M}$ represent the set containing all of the different types of married households. Clearly,

$$\mathcal{M} = \bigcup\nolimits_{E_H,E_W,LFP_H,LFP_W,KIDS} (\mathcal{M}_{E_H} \cap \mathcal{M}_{E_W} \cap \mathcal{M}_{LFP_H} \cap \mathcal{M}_{LFP_W} \cap \mathcal{M}_{KIDS}),$$

where the term in parenthesis is the set of all married households of type (E_H, E_W, LFP_H, $LFP_W, KIDS$).

Here is an example illustrating how the random matching experiment is performed. Take the first element of the matching table in 1960–see Table 1 of the main text. These are the marriages where both the husband and the wife are less-than-high-school educated. In 1960, the fraction of such marriages was 0.32. In terms of the current notation,

$$\frac{\sum_{i \in \mathcal{M}_{HS-,HS-}} \sum_{j=0.1}^{1} f_{ij}^{1960}}{\sum_{i \in \mathcal{M}} \sum_{j=0.1}^{1} f_{ij}^{1960}} = 0.32.$$

Now impose the random matching table entry for type-$(HS-, HS-)$ marriages in 1960; i.e., the first cell of the 1960 panel in Table A1. The fraction of such marriages, if matching in 1960 was random, is 0.21. Denote the counterfactual distribution to be imposed in 1960 by $\widetilde{f}_{ij}^{1960}$. The following equation must hold for the particular marriage group being discussed

$$\frac{\sum_{i\in\mathcal{M}_{HS-,HS-}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}f_{ij}^{1960}}=0.21.$$

The elements in the contingency table refer to the fraction of all married households that a particular type of match between husbands' and wives' educational levels constitutes. The elements in the cells are totals across all income percentiles. The f_{ij}'s refer to the fraction of all households, married and single, that are of type i in income percentile j. Therefore, the cells in the contingency table are aggregated over income percentiles (as well as the other non-educational traits characterizing married households). The ratio of the total number of type-$(HS-, HS-)$ marriages under random matching to what occurs in the data is

$$\frac{\sum_{i\in\mathcal{M}_{HS-,HS-}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}_{HS-,HS-}}\sum_{j=0.1}^{1}f_{ij}^{1960}}=\frac{0.21}{0.32}=0.66.$$

So, under random matching the number of type-$(HS-, HS-)$ marriages is reduced by factor of $0.66 = 0.21/0.31$. Assume that this reduction is spread out evenly across all of the income percentiles, or across all of the j's. Therefore, when undertaking the random matching experiment, $\widetilde{f}_{ij}^{1960}$should be constructed as follows:

$$\widetilde{f}_{ij}^{1960}=\frac{0.21}{0.31}f_{ij}^{1960}, \text{ for } i\in\mathcal{M}_{HS-,HS-} \text{ and all } j.$$

A similar scaling operation is performed for each of the other 24 possible matches. Thus, there is a scaling factor specific to each type of marriage (in the contingency table). For all single and divorced people, keep the original fractions; i.e., $\widetilde{f}_{ij}^{1960}=f_{ij}^{1960}$.

Imposing Random Matching While Holding Fixed Married Female Labor-Force Participation

The impact of random matching on inequality can be interacted with changes in the labor-force participation decisions of married females. The procedure for imposing random matching in 1960 is outlined in the previous section. Suppose that in addition to imposing random matching in 1960, married female labor-force participation is fixed at its 2005 level. How can this be implemented?

The married female labor-force participation rate in 1960 when random matching is imposed is

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}=0.33,$$

while the labor-force participation rate in 2005 is

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}f_{ij}^{2005}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}f_{ij}^{2005}}=0.68.$$

Denote the desired new counterfactual distribution for married households in 1960 by $\widehat{f}_{ij}^{1960}$, for $i\in\mathcal{M}$ and all j. This new counterfactual distribution for 1960 must give the 2005 married female

labor-force participation rate so

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}}=0.68.$$

Bear in mind that the fraction of married people in 1960 does not change in the counterfactual experiments; i.e.,

$$\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}f_{ij}^{1960}=\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}=\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}.$$

Consequently,

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}=\frac{0.68}{0.33}\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}=0.68.$$

Imposing a labor-force participation rate from 2005 onto the 1960 counterfactual distribution of random matching amounts to scaling up all $(i,\,j)$-combinations of married households in which women work. On the other hand, the married households in which women do not work should be scaled down so that the total fraction of married households does not change.

Therefore, the counterfactual distribution, $\{\widehat{f}_{ij}^{1960}\}$, should be constructed in the following way:

$$\widehat{f}_{ij}^{1960}=\frac{0.68}{0.33}\widetilde{f}_{ij}^{1960},\text{ for }i\in\mathcal{M}_{WORK_W}\text{ and all }j,$$

and

$$\widehat{f}_{ij}^{1960}=\frac{1-0.68}{1-0.33}\widetilde{f}_{ij}^{1960},\text{ for }i\in\mathcal{M}_{\sim WORK_W}\text{ and all }j.$$

This way the total fraction of married households stays constant,

$$\begin{aligned}\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}&=\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}+\sum_{i\in\mathcal{M}_{\sim WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}\\&=\frac{0.68}{0.33}\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}+\frac{1-0.68}{1-0.33}\sum_{i\in\mathcal{M}_{\sim WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}\\&=0.68\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}+(1-0.68)\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}\\&=\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}.\end{aligned}$$

As with the previous counterfactual distribution adjustment, keep the original fractions, $\widehat{f}_{ij}^{1960}=f_{ij}^{1960}$, for all single and divorced people.

A4. Standardizing Contingency Tables

Mosteller (1968) suggests that when comparing two contingency tables they should first be standardized so that they both have the same marginal distributions associated with the rows and columns. Take a 5×5 table. It can be standardized so that each element of the two marginal distributions is $1/5$. This can be done by employing the Sinkhorn-Knopp (1967) algorithm, which iteratively scales each row and column. Standardization preserves the core pattern of association in a contingency table. For example, Tan, Kumar and Srivastava (2004) note that such standardization does not affect the odds ratios in a contingency table, a typical measure used to gauge the pattern of association between

variables.

SINKHORN-KNOPP (1967) ALGORITHM

1) Enter an iteration with a contingency table.

2) This contingency table has a marginal distribution associated with the rows (for men) obtained by summing each row along its columns to obtain a total for that row. Divide each row through by 5 times its total. The marginal distribution associated with the rows is now $(1/5, 1/5, 1/5, 1/5, 1/5)$.

3) Compute the marginal distribution associated with the columns (for women) by summing each column along its rows to obtain a total for that column. Divide each column through by its 5 times its total.

4) Recompute the marginal distribution associated with the rows. It has changed following the previous two steps. Check its distance from the desired marginal distribution $(1/5, 1/5, 1/5, 1/5, 1/5)$. If it has reached the desired level of closeness then stop. If not, go back to Step 1.

THE STANDARDIZED TABLES

The two resulting standardized tables for 1960 and 2005 are shown in Table A2. The diagonal elements in the 2005 table are larger than in the 1960 one. Assortative mating has increased.

There is no need to standardize the tables so that each element of the marginal distributions is $1/5$. One can standardize the 1960 table so that its marginal distributions coincide with those in the data for 2005, or vice versa. This results are shown in Table A3. This way the standardized table for 1960 (2005) can be compared with the one from the data for 2005 (1960). Both tables will have the same 2005 (1960) marginal distributions. By comparing the standardized contingency table for 1960 (Table A.3) with the contingency table from the data for 2005 (see Table 1 in the text) it can be seen that assortative mating has increased. Once again, the diagonal elements are larger in the table for 2005. Likewise, a comparison of the standardized table for 2005 (Table A.3) with the one in the data for 1960 (again, see Table 1 in the text) shows an increase in assortative mating.

A5. A Brief Literature Review

The increase in assortative mating in the U.S. has also been examined by Hou and Myles (2008), Lam (1997), Qian and Preston (1993), and Schwartz and Mare (2005) to name a few papers. Siow (2013) documents an increase in educational homogamy, but not a general increase in positive assortative matching. Lam (1997) and Schwartz (2010) discuss the relationship between assortative mating and income inequality. Cancian and Reed (1998, 1999) also focus on the role that married female-labor force participation plays in the relationship between assortative mating and income inequality.

TABLE A1—RANDOM MATCHING CONTINGENCY TABLES: MATING BY EDUCATIONAL CLASS

Marital Sorting by Educational Pairing

1960

Husband	Wife				
	HS-	HS	C-	C	C+
HS-	**0.207**	0.192	0.053	0.026	0.008
HS	0.118	**0.110**	0.031	0.015	0.004
C-	0.045	0.042	**0.012**	0.006	0.002
C	0.030	0.028	0.008	**0.004**	0.001
C+	0.025	0.024	0.007	0.003	**0.001**
Marginal, Wives	0.425	0.396	0.110	0.054	0.016

2005

HS-	**0.006**	0.027	0.020	0.020	0.010
HS	0.024	**0.114**	0.084	0.084	0.041
C-	0.015	0.073	**0.054**	0.053	0.026
C	0.015	0.072	0.053	**0.053**	0.026
C+	0.009	0.043	0.032	0.032	**0.015**
Marginal, Wives	0.070	0.329	0.242	0.241	0.118

Note: Each cell gives the fraction of married households that would lie in the indicated educational pairing between husbands and wives when matching is random.

TABLE A2—STANDARDIZED CONTINGENCY TABLES: ASSORTATIVE MATING BY EDUCATIONAL CLASS

Marital Sorting by Educational Pairing

1960

Marginal Distributions $= (1/5, \cdots, 1/5)$

Husband	Wife				
	HS-	HS	C-	C	C+
HS-	**0.126**	0.043	0.017	0.007	0.007
HS	0.046	**0.079**	0.038	0.019	0.017
C-	0.020	0.045	**0.067**	0.037	0.032
C	0.005	0.023	0.047	**0.081**	0.043
C+	0.002	0.010	0.031	0.055	**0.102**
Marginal Distribution, Wives	1/5	1/5	1/5	1/5	1/5

2005

Marginal Distributions $= (1/5, \cdots, 1/5)$

HS-	**0.146**	0.035	0.014	0.004	0.002
HS	0.035	**0.088**	0.047	0.019	0.011
C-	0.013	0.047	**0.079**	0.038	0.023
C	0.004	0.021	0.039	**0.082**	0.054
C+	0.002	0.010	0.022	0.057	**0.109**
Marginal Distribution, Wives	1/5	1/5	1/5	1/5	1/5

Note: The upper panel shows the contingency table for 1960 when it has been normalized using the Sinkhorn-Knopp algorithm so that each element of marginal distributions over education for men and women equals $1/5$. The lower panel shows the same thing for 2005.

TABLE A3—STANDARDIZED CONTINGENCY TABLES: ASSORTATIVE MATING BY EDUCATIONAL CLASS

Marital Sorting by Educational Pairing

1960

Using the 2005 Marginal Distributions

Husband \ Wife	HS-	HS	C-	C	C+
HS-	**0.029**	0.035	0.011	0.005	0.003
HS	0.030	**0.186**	0.072	0.040	0.019
C-	0.008	0.065	**0.079**	0.048	0.022
C	0.002	0.032	0.055	**0.101**	0.028
C+	0.001	0.010	0.025	0.048	**0.047**
Marginal Distribution, Wives	0.070	0.329	0.242	0.241	0.118

2005

Using the 1960 Marginal Distributions

HS-	**0.354**	0.114	0.015	0.002	0.000
HS	0.054	**0.183**	0.033	0.007	0.001
C-	0.011	0.054	**0.031**	0.008	0.001
C	0.004	0.027	0.017	**0.019**	0.003
C+	0.002	0.017	0.013	0.017	**0.009**
Marginal Distribution, Wives	0.425	0.396	0.110	0.054	0.016

Note: The upper panel shows the contingency table for 1960 when it has been normalized using the Sinkhorn-Knopp algorithm so that the marginal distributions for men and women over education equal what there are in the data for 2005. The lower panel shows the contingency table for 2005 when it has been normalized so that the marginal distributions for men and women equal what there are in the data for 1960.

ADDITIONAL REFERENCES FOR THE ONLINE APPENDIX

Cancian, Maria and Deborah Reed. 1999. "The Impact of Wives' Earnings on Income Inequality: Issues and Estimates." *Demography*, 36, (2): 173–84.

Hou, Feng and John Myles. 2008. "The Changing Role of Education in the Marriage market: Assortative Marriage in Canada and the United States since the 1970s." *Canadian Journal of Sociology*, 33 (2): 337-366.

Lam, David. 1997. "Demographic Variables and Income Inequality." In *Handbook of Population and Family Economics*, edited by Mark R. Rosenzweig and Oded Stark, 1015–1059. Amsterdam, Elsevier North Holland.

Mosteller, Frederick. 1968. "Association and Estimation in Contingency Tables." *Journal of the American Statistical Association*, 63 (321): 1-28.

Qian, Zhenchao and Samuel H. Preston. 1993. "Changes in American Marriage, 1972 to 1987: Availability and Forces of Attraction by Age and Education." *American Sociological Review*, 58 (4): 482-495.

Rao, V. M. 1969. "Two Decompositions of Concentration Ratio." *Journal of the Royal Statistical Society*, Series A (General), 132 (3): 418-425.

Schwartz, Christine R. and Robert D. Mare. 2005. "Trends in Educational Assortative Marriage from 1940 to 2003." *Demography*, 42 (4): 621-646.

Sinkhorn, Richard and Paul Knopp. 1967. "Concerning Nonnegative Matrices and Doubly Stochastic Matrices." *Pacific Journal of Mathematics*, 21 (2): 343-348.

Siow, Aloysius. 2013. "Testing Becker's Theory of Positive Assortative Matching." *Journal of Labor Economics*, forthcoming.

Tan, Pang-Ning, Vipin Kumar, Jaideep Srivastava. 2004. "Selecting the Right Objective Measure for Association Analysis." *Information Systems*, 29 (4): 293–313.