Online appendix of the paper Naszodi, A., and Mendonca, F., 2022, "Changing educational homogamy: Shifting preferences or evolving educational distribution?"

July 19, 2022

Appendix A

In this appendix, we present some details about the original Liu–Lu measure, and the generalized Liu–Lu measure.

Liu and Lu (2006) assume that the economy consists of equal number of men and women. This number is also equal to the number of couples to be formed since all of the individuals are assumed to marry a person of the opposite sex eventually.

In this economy, there are four types of matches. Accordingly, the contingency table is: $K = \begin{bmatrix} N_{L,L} & N_{L,H} \\ N_{H,L} & N_{H,H} \end{bmatrix}$. The educational distributions are assumed to be non-degenerate: there is at least one man in each educational group ($N_{L,\cdot} \geq 1$, $N_{H,\cdot} \geq 1$) and there is at least one woman in each educational group ($N_{\cdot,L} \geq 1$, $N_{\cdot,H} \geq 1$).

The *Liu–Lu measure* is given by:

$$\mathrm{LL}(K) = \begin{cases} \frac{N_{H,H}-Q^-}{\min(N_{H,\cdot},N_{\cdot,H})-Q^-} \ , & \text{if} \quad N_{H,H} \geq Q \ , \\ \frac{N_{H,H}-Q^+}{Q^+-\max(0,N_{H,\cdot}-N_{\cdot,L})} \ , & \text{if} \quad N_{H,H} < Q \ , \end{cases} \tag{1}$$

where the interpretations of N (total number of couples), $N_{H,\cdot}$ (number of couples with high educated husbands) and $N_{\cdot,H}$ (number of couples with high educated wives) are the same as before. While $Q = N_{H,\cdot}N_{\cdot,H}/N$ is the expected number of H,H-type couples under the counterfactual of random matching. Furthermore, Q^- is the biggest integer that is smaller than or equal to Q, while Q^+ is the smallest integer that is larger than or equal to Q.

Since highly educated people tend to marry highly educated people, the empirically relevant aggregate matching outcomes are those, where there are more H,H-type couples than would be under random matching. Once $N_{H,H} \geq Q$ is assumed, the Liu–Lu measure simplifies to

$$\mathrm{LL}^{\mathrm{sim}}(K) = \frac{N_{H,H} - Q^-}{\min(N_{H,\cdot}, N_{\cdot,H}) - Q^-} \ . \tag{2}$$

The *statistical interpretation of the simplified Liu–Lu measure* is this. It is the normalized distance between the realized matching outcome K and a benchmark outcome where

individuals are randomly matched. If the number of H,H-type couples equals to its (integer valued) expected value under random matching, i.e., $N_{H,H} = Q^{-}$, the Liu–Lu measure takes the value zero. If sorting maximizes the number of H,H-type couples, the Liu–Lu measure takes its maximum value, 1. The number of H,H-type couples is maximized for any given marginal distributions if H-type men (women) marry low educated women (men) only if no highly educated women (men) remain single. For any other matching outcomes, the simplified Liu–Lu measure takes a value between zero and one.

In the empirical part of this paper, we work with an assorted trait variable that can take three possible values. The corresponding contingency table is a 3-by-3 matrix:

$$B = \begin{bmatrix} N_{L,L} & N_{L,M} & N_{L,H} \\ N_{M,L} & N_{M,M} & N_{M,H} \\ N_{H,L} & N_{H,M} & N_{H,H} \end{bmatrix},$$

where $N_{m,f}$ is the number of couples with husbands' education level $m \in \{L, M, H\}$ and wives' education level $f \in \{L, M, H\}$.

To analyze this case, we apply the generalized Liu–Lu measure. Following Naszodi and Mendonca (2021), we briefly describe the steps of calculating this generalized statistics.

First, the educational trait variable needs to be dichotomized. It can be done in four different ways depending on whether the M-type men and the M-type women are considered to be low or high educated in the dichotomous world. The four dichotomizations of the contingency table B provide a set of contingency tables: $\{K^{H,H}, K^{H,L}, K^{L,H}, K^{H,H}\}$, where the notation $K^{m,f}$ stands for the 2-by-2 contingency table obtained by reclassifying the M-type husbands to m-type and the M-type wives to f-type ($m, f \in \{L, H\}$).

The second step involves the calculation of the original Liu–Lu measure for each of the four 2-by-2 contingency tables. As a result, we obtain the so-called *Liu–Lu matrix characterizing*

marriages:

$$\mathrm{LL}^{\mathrm{gen}}(B) = \begin{bmatrix} \mathrm{LL}(K^{H,H}) & \mathrm{LL}(K^{H,L}) \\ \mathrm{LL}(K^{L,H}) & \mathrm{LL}(K^{L,L}) \end{bmatrix} . \tag{3}$$

So, this is how we generalize Eq. 1 when the assorted trait is trinomial and ordered.

Appendix B

This appendix describes how the number of individuals in the *population* with a given gender, own education level (OE), reservation point (RP), and marital status (MS) are estimated from the number of *users* of the dating site with the same gender, education level, and reservation point. We denote the number of individuals in the population with certain characteristics by $N_{\text{gender}}^{\text{OE, RP, MS}}$, where gender $\in \{w, m, \cdot\}$ (woman, man, either); OE $\in \{L, M, H, \cdot\}$ (low, medium, high, any); RP $\in \{L, M, H, S, \cdot\}$ (low, medium, high, single, any); MS $\in \{0, 1, \cdot\}$ (not in a couple, in a couple, either). While $n_{\text{gender}}^{\text{OE, RP, MS}}$ stands for the number of users with the same characteristics.

We note that there might be a selection into the sample because individuals, who prefer to remain single (i.e., having RP $= S$) are unlikely to sign up for the service of the dating site.[1] Second, the deferred-acceptance algorithm does not match everybody, i.e., some will remain single even out of those who prefer to be matched. We refer to the first group of singles as the *voluntary singles* (populated by $N_{\cdot}^{\cdot,\mathrm{RP}=S,0}$ number of individuals) and the second group of singles as the *involuntary singles* (with population size $N_{\cdot}^{\cdot,\mathrm{RP}\neq S,0}$).

It is assumed that the sample of users of the dating site is the result of a multinomial sampling from the population of individuals with reservation point different from S. Along

[1] Although we explicitly model this type of selection, we do not model others. For instance, we do not complicate the model with taking into account the fact that only those individuals from the population of users are selected into the sample, who indicate their education level, age, and search criterion.

these lines, the preferences of the users described by the distribution of their search criteria specific to their own type and gender are representative for the preferences of the individuals in the population of the same type and same gender with RP $\in \{L, M, H\}$, although being subject to sampling variation.

Let us first concentrate on the highly educated men. Their total number in the population is denoted by $N_m^{H,\cdot\cdot\cdot}$. Suppose that we know how many out of them prefer to remain single. So, the population share of the highly educated voluntary single men, denoted by $P_m^{H,S}$ is also known: $P_m^{H,S} = \frac{N_m^{H,S,0}}{N_m^{H,\cdot,\cdot}}$.

Then, the probability mass function capturing the likelihood of observing $n_m^{H,L}$, $n_m^{H,M}$ and $n_m^{H,H}$ number of highly educated male users with search criteria L, M and H, respectively out of $n_m^{H,\cdot}$ highly educated male users is:

$$\begin{aligned} &Pr(n_m^{H,L}, n_m^{H,M}, n_m^{H,H} | P_m^{H,L}, P_m^{H,M}, P_m^{H,H}, P_m^{H,S}) = \\ &= \frac{n_m^{H,\cdot}!}{n_m^{H,L}! n_m^{H,M}! n_m^{H,H}!} \left(\frac{P_m^{H,L}}{1-P_m^{H,S}}\right)^{n_m^{H,L}} \left(\frac{P_m^{H,M}}{1-P_m^{H,S}}\right)^{n_m^{H,M}} \left(\frac{P_m^{H,H}}{1-P_m^{H,S}}\right)^{n_m^{H,H}} , \end{aligned} \tag{4}$$

where $P_m^{H,L} = \frac{N_m^{H,L,\cdot}}{N_m^{H,\cdot,\cdot}}$ is the population share of the high educated men with reservation point L relative to all the high educated men; $P_m^{H,M} = \frac{N_m^{H,M,\cdot}}{N_m^{H,\cdot,\cdot}}$ is the population share of the high educated men with reservation point M relative to all the high educated men; and $P_m^{H,H} = \frac{N_m^{H,H,\cdot}}{N_m^{H,\cdot,\cdot}}$ is the population share of the high educated men with reservation point H relative to all the high educated men. By definition $P_m^{H,L} + P_m^{H,M} + P_m^{H,H} + P_m^{H,S} = 1$.

The respective concentrated log likelihood (conditional on $P_m^{H,S}$) is given by:

$$\begin{aligned} &l(P_m^{H,L}, P_m^{H,M}, P_m^{H,H} | P_m^{H,S}, n_m^{H,L}, n_m^{H,M}, n_m^{H,H}) = \text{Const} + n_m^{H,L} \ln(P_m^{H,L}) + n_m^{H,M} \ln(P_m^{H,M}) + \\ &+ n_m^{H,H} \ln(P_m^{H,H}) - (n_m^{H,L} + n_m^{H,M} + n_m^{H,H}) \ln(P_m^{H,L} + P_m^{H,M} + P_m^{H,H}) \quad . \end{aligned}$$

The maximum likelihood estimates on the population shares are obtained as:

$$\hat{P}_m^{H,L} = (1 - P_m^{H,S})\, n_m^{H,L} / (n_m^{H,L} + n_m^{H,M} + n_m^{H,H}),$$

$$\hat{P}_m^{H,M} = (1 - P_m^{H,S})\, n_m^{H,M} / (n_m^{H,L} + n_m^{H,M} + n_m^{H,H}),$$

$$\hat{P}_m^{H,H} = (1 - P_m^{H,S})\, n_m^{H,H} / (n_m^{H,L} + n_m^{H,M} + n_m^{H,H}).$$

The estimates on the population shares of individuals other than the highly educated men

can be obtained analogously. As a result, the estimates for the vector of the 18 population shares $\hat{P}^{\mathrm{RP}\neq S} = [\hat{P}_w^{L,L} \ ... \ \hat{P}_m^{H,H}]^T$ can be expressed as a function of the share of voluntary singles from both genders and of various types.

We estimate the respective 6-by-1 vector of parameters $P^{\mathrm{RP}=S} = [P_w^{L,S} \ P_w^{M,S} \ P_w^{H,S} \ P_m^{L,S} \ P_m^{M,S} \ P_m^{H,S}]^T$ by the *generalized method of moments estimator* (GMM) by minimizing the deviation of the share of singles generated by the deferred-acceptance algorithm (denoted by $*$) from the share of singles in the population subject to the constraint that the share of homogamous couples in the model is equal to its counterpart in the population:

$$\hat{P}^{\mathrm{RP}=S} = \underset{P^{\mathrm{RP}=S}}{\operatorname{argmin}} \Big(h(P^{\mathrm{RP}=S})^T h(P^{\mathrm{RP}=S}) \Big) \quad \text{s.t.} \quad \mathrm{SHC}^{\mathrm{agg},*} = \mathrm{SHC}^{\mathrm{agg}} \text{ , where} \tag{5}$$

$$h(P^{\mathrm{RP}=S}) = \left[\frac{N_w^{L,.,0} - N_w^{L,.,0,*}}{N_w^{L,.,.}}, \frac{N_w^{M,.,0} - N_w^{M,.,0,*}}{N_w^{M,.,.}}, \frac{N_w^{H,.,0} - N_w^{H,.,0,*}}{N_w^{H,.,.}}, \frac{N_m^{L,.,0} - N_m^{L,.,0,*}}{N_m^{L,.,.}}, \frac{N_m^{M,.,0} - N_m^{M,.,0,*}}{N_m^{M,.,.}}, \frac{N_m^{H,.,0} - N_m^{H,.,0,*}}{N_m^{H,.,.}} \right]^T$$

is the 6-by-1 vector of moments.

It is worth to note that the moment conditions used by the GMM do not cover the following three moment conditions: $\mathrm{SHC}_L^* = \mathrm{SHC}_L$, $\mathrm{SHC}_M^* = \mathrm{SHC}_M$, $\mathrm{SHC}_H^* = \mathrm{SHC}_H$. In other words, by satisfying the 6 moment conditions of the GMM and also $\mathrm{SHC}^{\mathrm{agg},*} = \mathrm{SHC}^{\mathrm{agg}}$ does not guarantee that the model-implied values of the education level-specific homogamy measures are identical to their observed counterparts.

References

Liu, H., Lu, J., 2006. Measuring the degree of assortative mating. *Economics Letters* 92, 317–322.

Naszodi, A., Mendonca, F., 2021. A new method for identifying the role of marital preferences at shaping marriage patterns. *Journal of Demographic Economics* , 1–27doi:`https://doi.org/10.1017/dem.2021.1`.