

Synthesizing Familial Linkages for Privacy in Microdata*

Gary Benedetto¹ and Evan Totty¹

September 17, 2020²

Abstract

As the Census Bureau strives to modernize its disclosure avoidance efforts in all of its outputs, synthetic data has become a successful way to provide external researchers a chance to conduct a wide variety of analyses on microdata while still satisfying the legal objective of protecting privacy of survey respondents. Some of the most useful variables for researchers are some of the trickiest to model: relationships between records. These can be family relationships, household relationships, or employer-employee relationships to name a few. This paper describes a method to match synthetic records together in a way that mimics the covariation between related records in the underlying, protected data.

* CED-DA Working Paper. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed in this paper are those of the authors and not necessarily those of the U.S. Census Bureau. All results were reviewed and approved by the Census Bureau Disclosure Review Board: CBDRB-FY20-287.

¹ Affiliation: Center for Enterprise Dissemination, Disclosure Avoidance; U.S. Census Bureau. Address: Center for Enterprise Dissemination, Disclosure Avoidance; U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233. Benedetto: gary.linus.benedetto@census.gov; Totty: evan.totty@census.gov

² Acknowledgements: The authors appreciate the insightful comments on earlier drafts of this paper from Jerome Reiter, Rolando Rodriguez, and Christine Task.

Synthesizing Familial Linkages for Privacy in Microdata

1 Introduction

Data providers face increasing demand from researchers to provide access to the actual microdata they collect, while also facing a more challenging privacy environment where intruders have more access to external data and more sophisticated techniques and computing power to attack privacy. Synthetic data has become one of the leading ways to provide researchers such access while still offering significant privacy protections. Methods for synthesizing attributes of people and firms have improved dramatically over the past couple of decades. However, not all information in microdata takes the form of a classic attribute such as age or income. Some of the most important information researchers want to take advantage of lies in the relationships between records in a data set. Researchers want to know about income mobility across generations, firm effects and co-worker effects in employee wages, how people's behavior impacts their spouses, and peer effects of all kinds. These linkages can be difficult to model, but still can present significant disclosure risk such as very large families, large numbers of divorces and marriages, and the pattern of how many jobs an employee holds over multiple years.

Using the language of graph theory, these linkages are the edges connecting nodes in a graph, and the more standard attributes (such as age and income) are features of the nodes (the people, firms, households, etc. in the data). In this paper, we describe an algorithm we have developed to model such relationships (or edges) that operates smoothly within the standard approach to generating synthetic data. We then use this algorithm to create synthetic familial linkages between individuals in the 2008 Survey of Income and Program Participation and evaluate the quality of the resulting synthetic data.

2 Background on Synthetic Data

Rubin (1993) and Little (1993) were the first to propose that the methods used for multiple imputation of missing data could be used to handle sensitive data as well. Rubin's original idea was that multiple imputation could be used, in essence, to complete the missing survey responses for the entire population from which the original sample had been drawn. Then,

from this population of completed data, the synthetic samples would be drawn. These synthetic samples could be released because they were not actual responses but random draws from an estimated distribution. Data users would not need specialized software or techniques, they could simply run their analyses on the multiple replicates and use the variation in results across replicates to correctly estimate variances of their statistics of interest.

Little proposed imputation to replace original, non-missing values as one of many possible mechanisms to “mask” sensitive values. The idea of replacing values in the original data with multiple imputation techniques is ultimately the direction most synthetic data applications have gone. When all values for all variables are replaced, we call it fully synthetic, and when only a subset of values or variables are replaced, we call it partially synthetic. Reiter (2003) and Reiter (2004) developed the proper formulae to be used in the calculation of variances with synthetic data, both for cases where the original data were complete and where the original data had been completed with multiple imputation prior to data synthesis.

Since then, a number of real world applications of partially and fully synthetic data have been developed, but generally do not include synthetic edges in a graph. Hu, Reiter, and Wang (2018) suggested an approach to modeling these edges using Dirichlet Process mixture models. Using this method, they generate the synthetic graph first (nodes and edges) with a handful of categorical attributes of the nodes (there is also a suggestion for extensions to continuous attributes). Their algorithm allows for modeling of the variables simultaneously at the node (person) and cluster (household) level. In the applications to a household survey, they preserve the general pattern of correlations of person-level attributes within their households quite well. In practice, one would likely use this method to generate the graph first, with a few critical attributes, and then synthesize the rest of the nodal attributes conditional on the attributes of both the node and the cluster to which it belongs (e.g., a person and his/her family/household). We attempt to offer an alternative where the full set of nodes and their attributes have already been synthesized, and all that remains is to synthesize the edges joining the nodes into clusters.

3 Methodology

3.1 One-to-one: Spousal Link

To synthesize one-to-one linkages between synthetic records, we developed a new approach that is similar in concept to predictive mean matching (Little and Rubin, 2002). For ease of explanation, we will describe the process for spouses, but other one-to-one matches (e.g., houses to householders) could use the same process. We assume that we have an internal (in this case, person-level) database, D_0 , with a column (or columns) of unique row identifiers and a set of observed attributes. We also have a separate file with two columns of values of the identifier showing which records from D_0 are connected by the link we are attempting to model (in this case, marriage), which we will call the internal crosswalk. We have already made a public, synthetic database, D_1 , where some (or all) of the attribute data has been synthesized, but we do not have a synthetic crosswalk to identify connections between synthetic records. As part of the set of attributes in D_0 and D_1 , we know which records compose the set of potential wives and husbands in each file (this attribute can be synthesized as well).³

Consider a set of k_w variables for potential wives, x_1, \dots, x_{k_w} , and a set of k_h variables for potential husbands, $y_1 - y_{k_h}$, for which we would like to preserve the correlations between linked records in the internal file when building our synthetic crosswalk. In theory, these can be different variables. If we were matching houses to householders, the two sets of variables could be house characteristics and householder characteristics, respectively. In the case of spouses, these sets of attributes were the same. The observed values of the x and y variables are stored in matrices, X and Y , from D_0 , and matrices, X_s and Y_s , from D_1 . These matrices might be very large and complex in practice, so, for the sake of dimension reduction, we calculate the first k_w^* principal components of X and first k_h^* principal components of Y where $k_w^* \leq k_w$ and $k_h^* \leq k_h$.⁴ In other words, we calculate:

³ We use the terms “husband” and “wife” because of the data we use to test our methodology later in the paper. Our results use the 2008 panel of the Survey of Income and Program Participation (SIPP). The SIPP did not release data with same-sex married couples until its next panel in 2014. Our methodology could be applied to the 2014 SIPP’s gender-neutral links by using the terms “reference person” and “spouse.”

⁴ The appendix provides a summary of methods for variable or data reduction and our motivation for using principal components rather than a variable selection technique.

$X^* = XW$ where W is the $k_w^* \times k_w^*$ matrix of eigenvectors of $X'X$ such that $W'X'XW = \Lambda_w$ and Λ_w is the diagonal matrix of eigenvalues of $X'X$ sorted from greatest to least (i.e., $\lambda_{w11} \geq \lambda_{w22} \geq \lambda_{w33} \dots$)

$Y^* = YH$ where H is the $k_h^* \times k_h^*$ matrix of eigenvectors of $Y'Y$ such that $H'Y'YH = \Lambda_h$ and Λ_h is the diagonal matrix of eigenvalues of $Y'Y$ sorted from greatest to least (i.e., $\lambda_{h11} \geq \lambda_{h22} \geq \lambda_{h33} \dots$)

Using the W and H from above, we also calculate $X_s^* = X_s W$ and $Y_s^* = YH$ on the synthetic data.

Using a non-parametric transform developed by Woodcock and Benedetto (2004), we map each of these principal components independently to approximately standard normal distributions. This transformation involves three steps: (1) estimate the distribution of a variable, x , on a Bayes' bootstrap sample of the internal data using a Kernel Density Estimator; (2) map the variable value into a real number in the interval $[0,1]$ using the estimated cumulative distribution function (CDF), $\hat{F}_x(x)$; (3) map this CDF value into the point on the real line with the same CDF value from the standard Normal distribution using the inverse CDF of the standard Normal, $\Phi^{-1}(\hat{F}_x(x))$. We will denote as $T_{A^*}(A) = \tilde{A}$ the mapping that performs this transformation to all of the columns of matrix, A , independently, estimated on A^* . The transformed random variables, \tilde{x}^* and \tilde{y}^* , represented by the columns of $\tilde{X}^* = T_{X^*}(X^*)$ and $\tilde{Y}^* = T_{Y^*}(Y^*)$ are, by design, a set of $k_w^* + k_h^*$ random variables, each of which is approximately distributed as standard Normal. While it is not necessarily the case that a set of standard Normal random variables follow a multivariate Normal distribution, we proceed with the model assumption that these variables are distributed as a multivariate Normal, $(\tilde{x}^* \tilde{y}^*) \sim N(\mu = [\mu_w \mu_h], \Sigma = \begin{bmatrix} \Sigma_{ww} & \Sigma_{wh} \\ \Sigma_{hw} & \Sigma_{hh} \end{bmatrix})$. We empirically test this assumption in the results section using the Royston Multivariate Normality Test. From the observed spouses in the internal data using the internal crosswalk, we can estimate μ and Σ as $\hat{\mu}$ and $\hat{\Sigma}$. We do this on a Bayes' Bootstrap sample of the internal data so as to account for sample uncertainty and follow proper posterior predictive sampling. We use the same distributions estimated on the internal data to transform the synthetic data: $\tilde{X}_s^* = T_{X^*}(X_s^*)$ and $\tilde{Y}_s^* = T_{Y^*}(Y_s^*)$.

The final step in the process is to link wives to husbands in the synthetic data using what we have estimated so far and the model assumption of multivariate normality. First, we randomly sort the wives from the synthetic data. Second we sequentially move through the randomly sorted set of synthetic wives and draw candidate husbands using the conditional multivariate Normal distribution: $N(\mu_h + \Sigma_{hw}\Sigma_{ww}^{-1}(\tilde{y}^* - \mu_w), \Sigma_{hh} - \Sigma_{hw}\Sigma_{ww}^{-1}\Sigma_{wh})$. Finally, we search through the set of synthetic husbands to find the nearest neighbor to the candidate husband using the Mahalanobis distance measure. When we find the nearest neighbor, we assign that link by creating a record in the synthetic crosswalk, and remove that synthetic husband from the matching pool before proceeding to the next synthetic wife.⁵

3.2 *One-to-many: Parent-Child Link*

The method we use to link one-to-many is very similar to the one-to-one case. For ease of explanation and to match what we did in our empirical tests, we will talk about matching children to mothers; however, this could also be used for other one-to-many matching problems such as employer to employees. In matching children to mothers, not only do we want to preserve correlations between mothers and children, but we also want to preserve correlations between the implied siblings that result from these connections.

The first few steps of getting the principal components and transforming them to standard normal distributions remains the same. Rather than immediately generating all of the candidate children for our synthetic mothers in a single step, we start by drawing a candidate “first” child for a mother according to a sort order chosen by the modeler. For mothers and children, a natural sort order is birthdate; however, if there is no natural sort order, a random sort order will work. Then, using an estimate of the variance/covariance matrix of the children’s variables from the previous child in the sort, we randomly draw a candidate “next” child conditional on the previous candidate child (when a random sort order is used, the variance/covariance matrix would, in essence, be estimated from a random sample of pairs within clusters). We continue to do this until we have the same number of candidate children for

⁵ We chose to sample without replacement so that we are not generating many copies of the same individual. We could also create extra synthetic records to sample from, but this creates quality issues which we discuss later.

the mother as the number of children the mother has (which should be an attribute in the database, and can be synthetic, in D_1). Finally, we match the nearest neighbor from the pool of synthetic children to each candidate child, after which we add that link to the synthetic crosswalk and remove that synthetic child from the pool.

Notice that this process assumes the covariance of children within the family is basically constant across all families. We try to evaluate whether this assumption allows for the appropriate amount of variety in the resulting synthetic families, but it is almost certainly a very inappropriate assumption in the case of matching employers to employees, since different businesses will have very different staffing needs. In such a case, we might want to synthesize, in advance, a variance group for the employer, and then perform the previous steps independently for each variance group.

4 Data

We used the Survey of Income and Program Participation (SIPP) Gold Standard File (GSF) to test our synthetic links methodology. The GSF is a confidential internal U.S. Census Bureau file that links self-reported survey information from SIPP respondents to administrative records with tax and benefit information from the Internal Revenue Service (IRS) and Social Security Administration (SSA). The U.S. Census Bureau provides a synthetic version of the GSF for external researchers known as the SIPP Synthetic Beta (SSB). The most recent version of the SSB (SSB v7.0) creates synthetic familial links using a similar methodology as the one described in this paper. The primary difference is that in this paper we include many variables and then use principal component analysis for data reduction of the variables used to create the links, whereas in SSB v7.0 we there was no data reduction or variable selection. Prior versions of the SSB left the first spousal link unsynthesized. More information on the creation and use of the GSF and SSB can be found in Benedetto et al. (2018).

We extracted a sub-sample of GSF records to test our methodology. The sample is based on individuals from the 2008 panel with non-missing information for race, Hispanic status, foreign born status and time of arrival in the USA, education level, home owner status and home equity. We also limited the sample to individuals who were successfully linked to the administrative

records. This resulted in 54,000 individuals, including 13,500 linked spouse pairs, 2,200 linked moms, and 2,500 linked kids. The variables used for synthesizing the links are shown in Table 1.⁶The years 2009 and 2010 were used because they are the first two full reference years in the 2008 SIPP panel.⁷

Table 1: Variables Used in SIPP Synthetic Familial Linkages

Variable	Description
hispanic	Hispanic ethnicity
foreign_born	Foreign born
time_arrive_usa (and squared)	Date of arrival in USA (if foreign born)
total_der_fica_2009 (and squared)	SSA Detailed Earnings Record – 2009 earnings
total_der_fica_2010	SSA Detailed Earnings Record – 2010 earnings
sipp_birthdate (and squared and cubed)	Birthdate reported in the SIPP
nonwhite	Recode – non-White race
black	Recode – Black race
pos_der_fica_2009	Indicator for positive 2009 DER earnings
pos_der_fica_2010	Indicator for positive 2010 DER earnings
educ_d1-educ_d5	Indicators five-category highest education level
sipp_birthdate X total_der_fica_2009	Interaction term
sipp_birthdate X educ_5cat	Interaction term
total_der_fica_2009 X educ_5cat	Interaction term

Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

We synthesized both the links (edges in the graph) and the variables themselves. The variables, or node attributes, are synthesized using the sequential regression technique outlined in Benedetto et al. (2018). For the sake of differentiating between the effects of synthesizing the nodal attributes and synthesizing the edges, we have four different versions of the microdata that we use in the analysis below: (1) original variables with original links, (2) synthesized variables with synthesized spouse and mother-child links, (3) synthesized variables with original spouse links and no mother-child links, and (4) original variables with synthesized spouse links and original mother-child links.⁸ For each synthesis we generated four replicates.

⁶ Categorical variables were turned into dummy variables for use in the principal components method. In future work we may explore other ways to perform dimension reduction for categorical variables.

⁷ The date variables are calculated as number of days between the date and January 1, 1960 divided by 1,000.

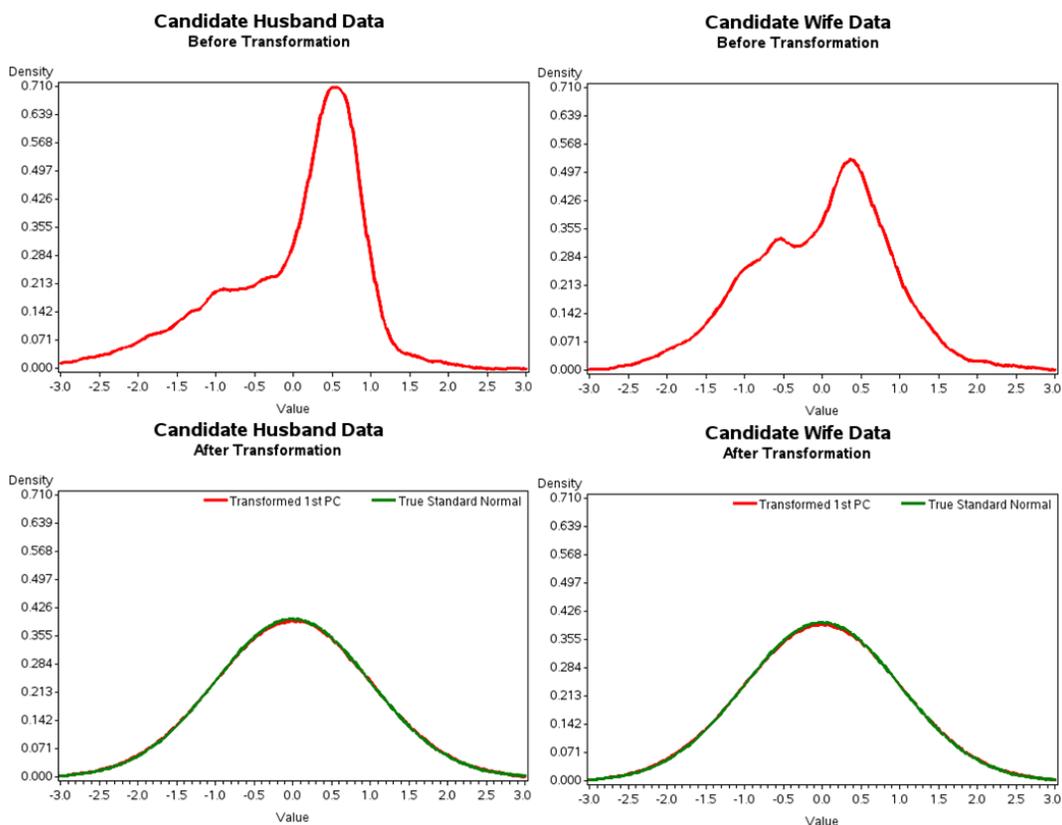
⁸ Mothers in the SIPP data include biological, step, adopted, and foster mothers.

5 Results

5.1 Test for Multivariate Normality

We begin by visually analyzing the quality of the standard normal assumption for the transformed principal components. Figure 1 shows the univariate density distribution for the first principal component from the potential husbands and potential wives, before and after transformation. The figures illustrate that the principal components do not resemble a standard normal distribution before transformation but do afterwards.

Figure 1: Univariate Distribution of First Principal Component

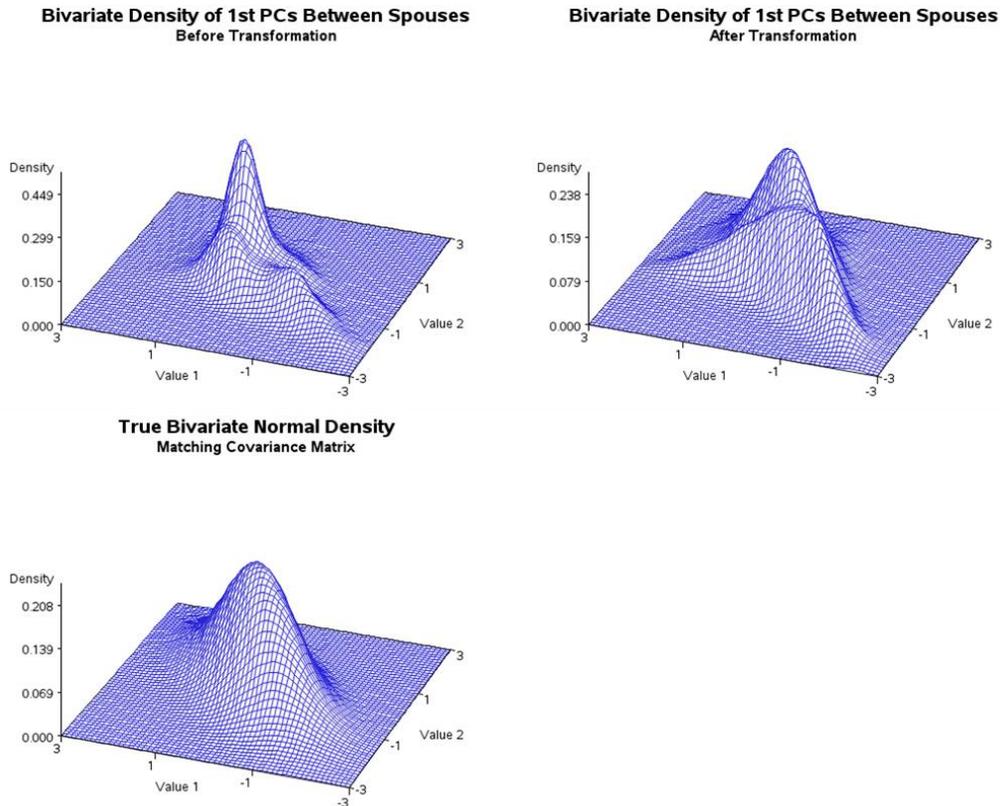


Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Figure 2 shows the bivariate density of the first principal component for husbands and wives, before and after transformation. It also shows the true bivariate normal density for comparison, with a covariance matrix equal to the sample covariance matrix of the transformed

principal components. The distribution after the transformation does not perfectly resemble a bivariate normal density (there is a ridge in surface creating a bit of bimodality), most likely due to imperfect model assumptions and perhaps somewhat to sample uncertainty. However, the density displays improved bivariate normality relative to the distribution prior to transformation.

Figure 2: Bivariate Distribution of Husband's and Wife's First Principal Component



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Next, we formally test our assumption of multivariate normality using the Royston (1983) test. This method first tests each of the principal component transformations for univariate normality using the Royston (1982) extension of the Shapiro-Wilk test (Shapiro and Wilk, 1965) to larger samples. The univariate test results are then combined into one test statistic for multivariate normality by transforming the Shapiro-Wilk statistics into an approximately Chi-squared random variable. The degrees of freedom are estimated by taking into account correlation between the univariate test statistics.

We performed the Royston (1983) test using the first ten principal components constructed from the potential wives' variables and the first ten principal components constructed from the potential husbands' variables. The variables used were those shown in Table 1. The Royston (1982) extension of the Shapiro-Wilk test suggests the test only be used for samples up to 2,000. After estimating the principal components, transforming them to standard normal, and linking couples together, we took the principal components from a random sample of 2,000 couples to use in the test.

Results for the test are shown in Table 2. We tested multivariate normality for the ten principal components from husbands and wives separately and together. The table reports the Royston test statistic, the estimated degrees of freedom, and the p-value associated with the null hypothesis of multivariate normality. All three tests fail to reject normality, suggesting that the model assumption of multivariate normality for the transformed principal components is not unreasonable.⁹

Table 2: Multivariate Normality Test for Husband and Wife Principal Components

	(1) Husbands Principal Components	(2) Wives Principal Components	(3) Both
Royston test statistic	14.99	10.12	5.01
Equivalent degrees of freedom	19.38	9.76	9.81
P-value	0.74	0.41	0.88
Number of principal components	10	10	20

Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

5.2 One-to-one: Spousal link

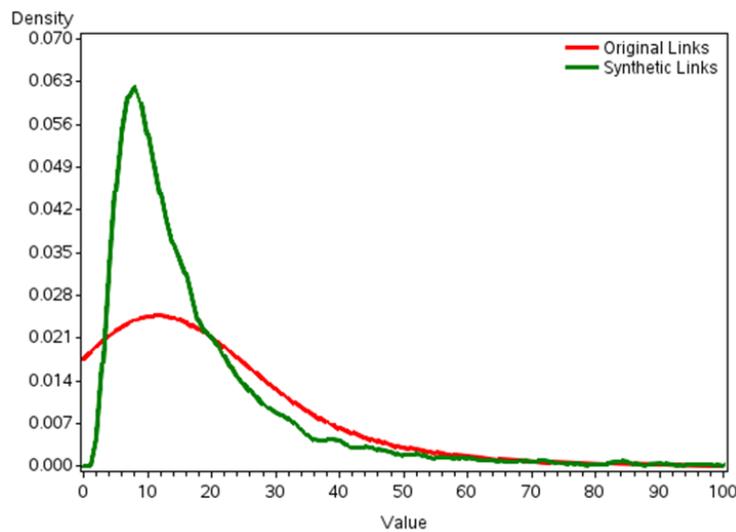
We begin analysis of the quality of synthetic edges by analyzing the one-to-one spousal links. We evaluate the quality in a few different ways. First, we check the distance between predicted spouses and linked spouses for the real and synthetic links. Second, we check the

⁹ We found that the Royston test would reject multivariate normality when we tried this with too few continuous variables or too much smoothing in the KDE step of the transformation. It warrants further study to understand the where the multivariate normality assumption is reasonable and where it is not.

distribution of key variables of interest, such as the distribution of spousal age difference and relative earnings in the real vs synthetic couples. Third, we use two-way k-marginals to assess the overall similarity in the bivariate distribution of many pairs of characteristics between real and synthetic linked spouses. Finally, we check the percent of original links that are re-created in our synthesis when we only synthesize the edges (not the nodes) in order to assess its impact on privacy.

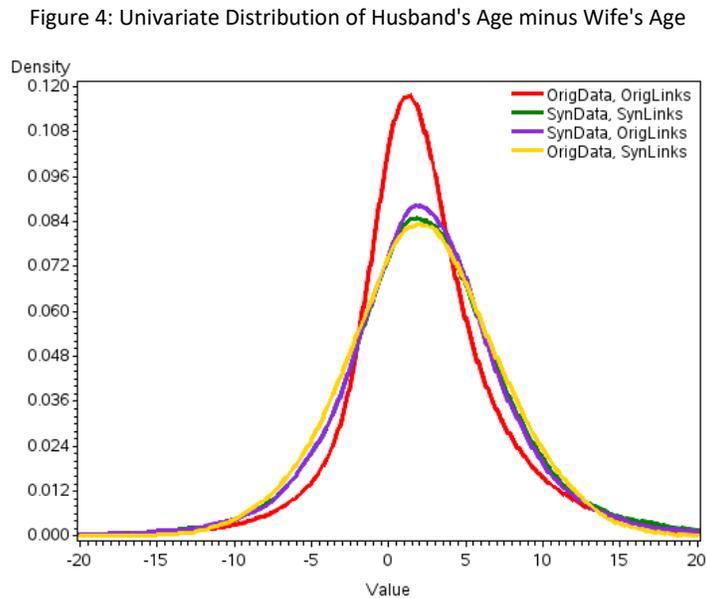
As described above, the synthesis process for spousal links involves randomly sorting the wives, drawing candidate husbands for each wife using the conditional multivariate normal distribution, and then finding the nearest neighbor to the candidate husband using the Mahalanobis distance measure. Figure 3 shows the univariate distribution of Mahalanobis distance between the predicted and linked spouses for both the real and synthetic linkages. The synthetic version generates more links with a relatively small distance of 5-20 at the expense of fewer linkages with a distance less than 5 or 20-60. Both linkages have long tails. The difference in these distributions suggest that there may be room for improvement in further exploring the sampling strategy (rather than simply randomly sorting the wives prior to matching).

Figure 3: Distance of Predicted Mean Spouse to Linked Spouse



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Figure 4 shows the univariate distribution of the age difference between spouses (husband age minus wife age) for four different spousal links: (1) original variables with original links, (2) synthesized variables with synthesized links, (3) synthesized variables with original links, and (4) original variables with synthesized links. All three of the synthesized versions have less density around the median and appear to be slightly skewed toward a larger positive age gap. Overall, the synthetic edges appear to do a reasonable job of maintaining a similar shape for the distribution of age differences, and perform about the same as when we leave the edge unsynthesized and synthesize each spouse’s attributes conditional on their partner’s attributes.



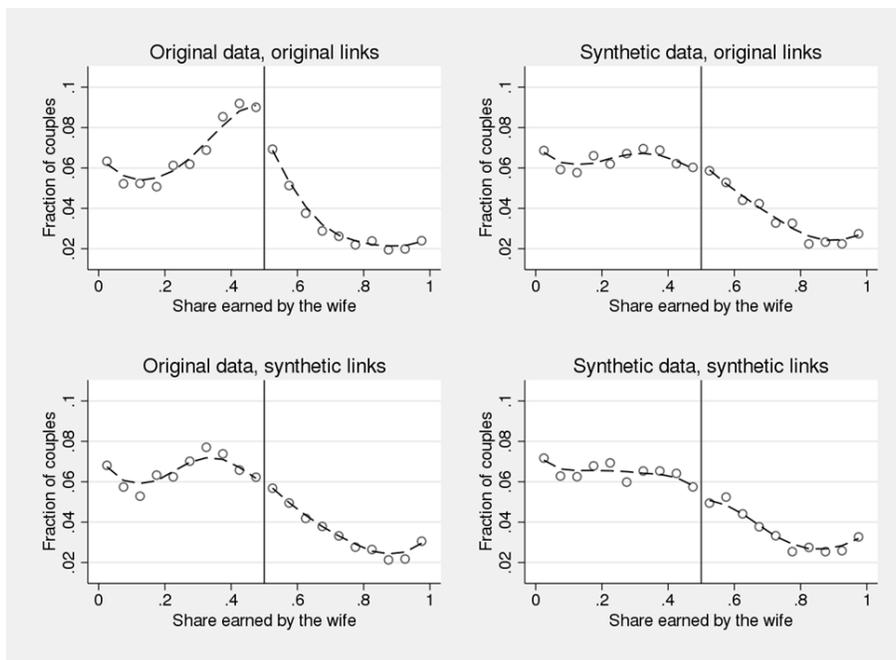
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

We turn next to relative earnings by analyzing the distribution of the wife’s share of the couple’s combined earnings. Previous research has shown that there is a spike in the density just below the 50% threshold and a large drop in density just above that threshold (Bertrand, Kamenica, and Pan; 2015). Other work has shown that this pattern may be partially due to survey misreporting related to gender norms (Murray-Close and Heggeness, 2018). It is reasonable to assume then that synthetic data could struggle to replicate this difference in density around an

otherwise arbitrary threshold, particularly when the modeling does not attempt to account for it directly.

Figure 5 shows the univariate distribution of the share of married couple’s earnings that are earned by the wife for the same four sets of spousal links used for age differences: (1) the original variables with original links, (2) synthesized variables with synthesized links, (3) synthesized variables with original links, and (4) original variables with synthesized links. Figure 6 also shows the distribution for two previous iterations of the SSB.¹⁰

Figure 5: Share of Spousal Earnings Earned by the Wife



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

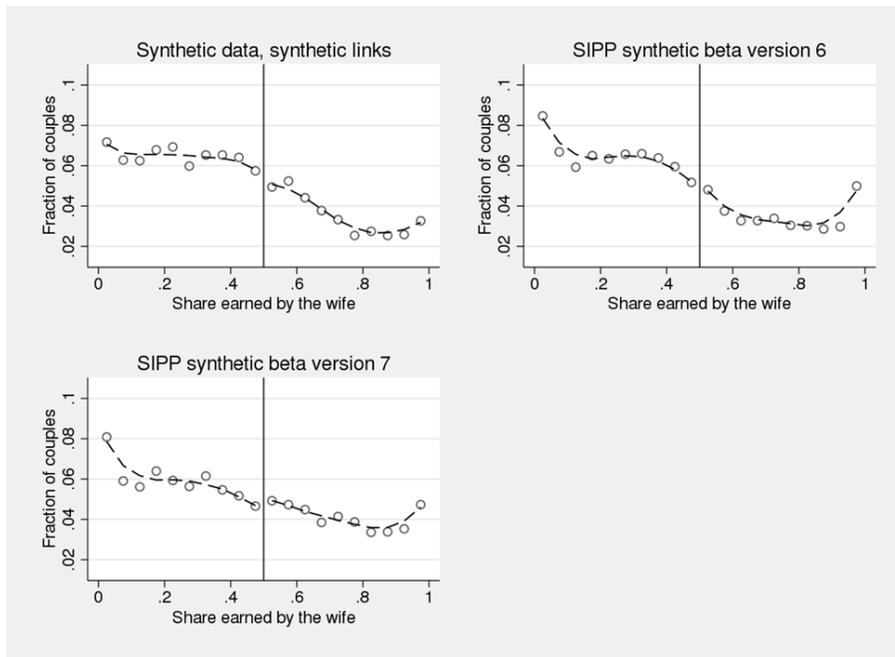
The original-original version shows the stark discontinuous drop in density around the 50% threshold. Synthesizing only the variables with existing links or only the links for existing variables fails to replicate this discontinuous drop. The figure based on synthesized variables and

¹⁰ Bertrand, Kamenica, and Pan (2015) used the SSB for their earnings share figure. The figure in the paper shows the results from the validation on the internal data after the authors built their code on the synthetic data.

synthesized links does show a drop in density across the 50% threshold, although it is a much smaller drop than on the original data.

This is a noticeable improvement over the two versions of previous SSB data shown in Figure 6. Version 6 of the SSB left the first spousal link unsynthesized. Version 7 synthesized the linkages using the same procedure described in this paper, except without reducing the set of variables used for husbands and wives to their principal components. Using principal components allows us to reduce the dimension of the data and therefore include more variables without issues of computational complexity. It also allows more variables to contribute in a meaningful way by summarizing many highly correlated variables into a smaller number of principal components and allowing other variables with important features to contribute strongly to other principal components. The results suggest that using the principal components rather than the full list of variables or variable selection techniques for drawing candidate spouses may provide meaningful improvements.

Figure 6: Share of Spousal Earnings Earned by the Wife in SIPP Synthetic Beta



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

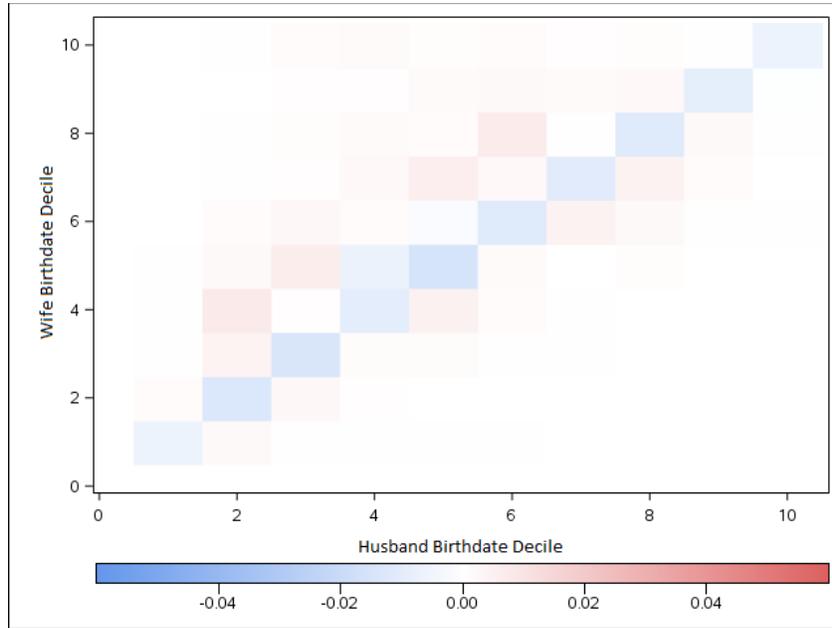
Next, we use two-way k-marginals to evaluate similarity in the bivariate distribution of many variables between the real and synthetic couples. The two-way k-marginal takes any pair of variables, makes discrete categories out of the full range of values for each variable, and then constructs the distribution density that falls within each cell of the two-way combination.¹¹ This is done on both the real set of couples and the synthetic, and then the absolute value of the difference in the density between the real and synthetic data is summed across all cells. If the real and synthetic couples have similar distributions for the two-way combination of variables, then the density difference in each cell will be very small. The smallest possible summed score is 0, which indicates identical distributions for the given variables and categories. The worst possible score is 2, which indicates two distributions with zero overlap.

The score was constructed for many different two-way marginals and then averaged to get an overall similarity score for the real and synthetic couples. In the context of spousal links, one of the two variables we used in the k-marginal was from the husband and the other was from the wife. The variables we used were race, Hispanic status, foreign born status, highest education level, time of arrival in USA, 2009 total FICA earnings, 2010 total FICA earnings, and SIPP birthdate.

Figure 7 shows a heat map of the two-way k-marginal score for spousal birthdates. Birthdate is split into ten deciles based on the distribution of observed birthdates among all spouses in the original data. Larger decile numbers indicate later birthdates. The density for each two-way cell is constructed on the original and synthetic data. Red-shaded areas indicate cells where the synthetic data had greater density than the original data and blue-shaded areas indicate cells where the synthetic data had less density than the original data. The figure indicates that the synthetic data has less density along the diagonal, which corresponds to spousal pairs whose birthdates are in the same decile, and more density just to the sides of the diagonal, which corresponds to spousal pairs where one spouse is slightly older than the other. The results are similar to Figure 4, where the synthetic data shows less density around the 0 age difference region and more density around the (-10,-2) and (5,13) regions.

¹¹ Categorical variables such as highest education level already have discrete categories. For continuous variables such as earnings, we created eleven categories which indicate the ten deciles of the full range of values plus a category for missing values.

Figure 7: Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Spousal Birthdate

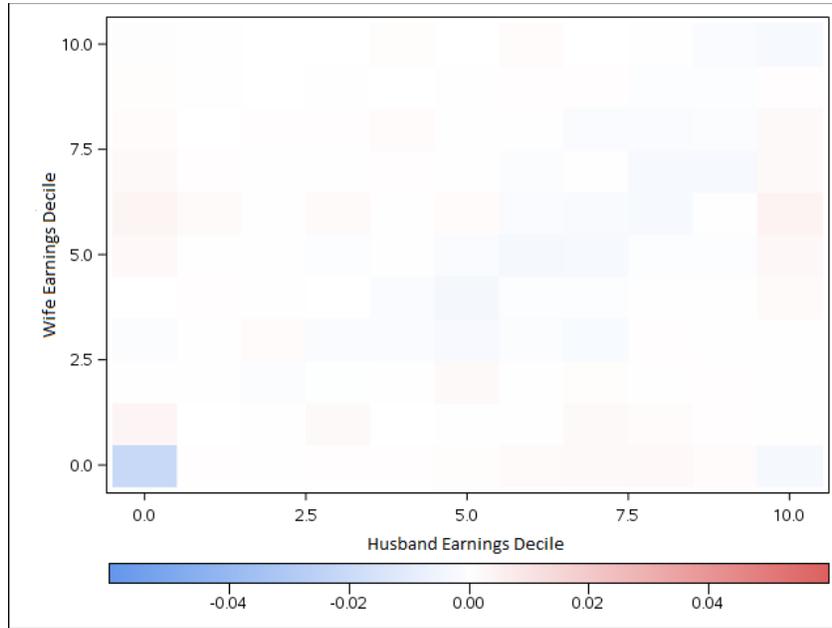


Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Figure 8 shows the heat map of k-marginal scores for 2009 total FICA earnings. Earnings are split into ten deciles based on the distribution of observed earnings among all spouses in the original data. Most of the cells are a lighter shade of red and blue than the previous figure, indicating overall greater similarity in distribution for synthetic spousal earnings than synthetic spousal birthdates. The figure shows less density in the synthetic data along the diagonal and also just below the diagonal. This corresponds to areas where the wife’s earnings are in the same decile as their husband’s or just below that of the husband. This is consistent with the wife’s spousal earning share distribution in Figure 5: the synthetic data does not fully replicate the bunching of the wife’s earnings just below that of the husband’s. There is also a relatively dark-blue cell at the (0,0) decile. Zero indicates that the earnings value was missing.¹² Thus, the synthetic data includes fewer spousal links where both the husband and the wife did not report any earnings from FICA-covered jobs with W-2 or Schedule C (self-employment) filings to the IRS.

¹² The heat map for birthdate did not include a zero decile because there was no missing birthdate information in the sample used for this analysis.

Figure 8: Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Spousal Earnings



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

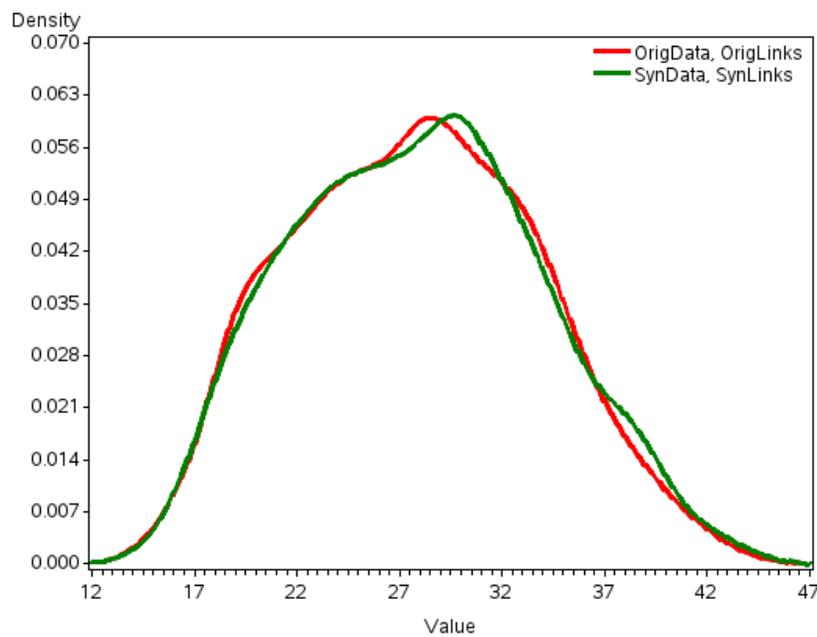
Column (1) of Table 3 shows the overall two-way k-marginal score after summing the absolute value of the difference in each cell for a given combination of variables, repeating this for all combinations of variables, and then averaging the resulting scores. The synthetic links were created four times in order to evaluate stability in scores across replications. The table also shows results for two sets of baseline comparison two-way k-marginal scores. One is based on taking two 50% random samples of the real couples and comparing them for similarity in distributions. The other is based on taking one 50% random sample of the real couples and comparing it to the full set of real couples. These scores are meant to indicate what type of similarity scores are expected when two sets of linked couples differ only due to sampling error.

The results show that the spousal link scores are consistent across the four replications of linkages. The synthetic links have scores that are approximately 50% larger than the two-sample baseline comparison, which is the most similar in spirit of the two comparisons.

Finally, we evaluate the level of disclosure limitation being provided by the synthesis of the edges in the graph. Since this is not a formally private algorithm, there is no way to clearly

quantify the level of privacy loss from the resulting synthetic edges. However, an intuitive test of the protection provided is to measure how often the actual links are recreated when synthesizing edges in the original data (as described above during the age difference checks). Table 4 shows the percent of links re-created for each of four different replicates of the synthetic links. The link re-creation rate is very small across all four replicates, suggesting that the privacy loss from these synthetic edges is low.

Figure 9: Univariate Distribution of Mother’s Age minus Child’s Age



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

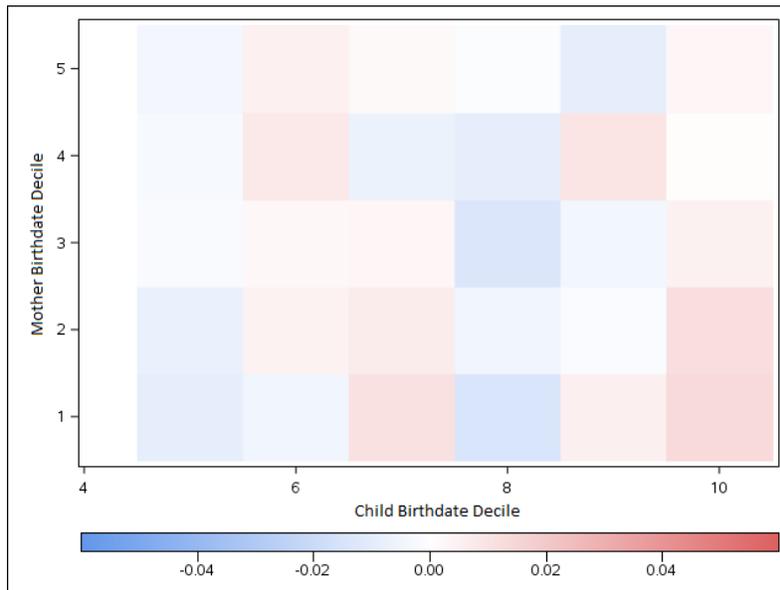
5.3 One-to-many: Parent-child link

Figure 9 shows the univariate distribution of the age difference between linked mothers and their children (mother age minus child age) for two different links: (1) the original variables with original links and (2) the synthesized variables with synthesized links.¹³ The synthetic version does a very good job of matching the distribution of age differences between mothers and

¹³ We did not create files with synthetic attributes and original edges because this would involve synthesizing all the attributes for each of a large number of possible children conditional on all the previously synthesized children. This becomes computationally burdensome, and severely reduces the sample size to estimate a model for the n-th sibling as n increases. This is one of the reasons why a separate method for synthesizing edges is necessary.

children. Figure 10 shows the two-way k-marginal heat map for SIPP birthdate for linked mothers and children. Birthdate is split into ten deciles based on the distribution of observed birthdates for all linked mothers and children in the original data. The figure does not show any clear pattern of systematic bias between the real and synthetic density distribution. If anything, there may be a small shift in the distribution toward relatively older mothers matched with younger kids.

Figure 10: Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Mother-Child Birthdate



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Table 3: Two-Way K-Marginal Scores for Familial Linkages

Synthetic File	Comparison File	(1)	(2)	(3)
		Spousal Links Score	Parent-Child Links Score	Sibling Links Score
Synthetic links file #1	Real links	0.066	0.146	0.349
Synthetic links file #2	Real links	0.062	0.131	0.323
Synthetic links file #3	Real links	0.064	0.129	0.334
Synthetic links file #4	Real links	0.064	0.132	0.31
Real Links 50% Sample	Real links	0.027	0.067	0.152
Real Links 50% Sample	Real Links 50% Sample	0.041	0.092	0.205

Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Table 4: Percentage of Original Links Re-Created With Original Data and Synthetic Links

	(1) Synthetic links #1	(2) Synthetic links #2	(3) Synthetic links #3	(4) Synthetic links #4
Percentage of links re-created	0.61%	0.66%	0.47%	0.58%

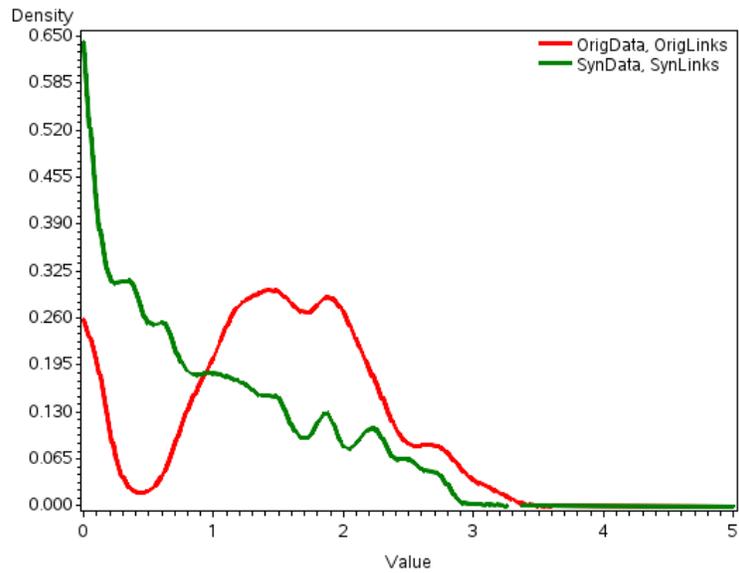
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Column (2) of Table 3 shows the overall two-way k-marginal scores for mother-child links. Similar to the results for spousal links, the four sets of mother-child links have scores that are about 50% larger than the two-sample baseline comparison. The sampling error and synthesis error scores for mother-child links are about twice as large as the scores for the spousal links. This suggests that, at least for the variables chosen, the distribution of linked mother-child characteristics has greater variation due to random chance. This is why it is important to include the baseline random sample comparison: so that we can consider the ratio of synthesis error to sampling error rather than just the level of synthesis error.

5.4 One-to-many: Siblings links

Next we turn to sibling links. Siblings are determined based on the parent-child links discussed previously: kids who link to the same mother become siblings. Figure 11 shows the univariate distribution of the age difference between linked siblings (sibling age minus next youngest sibling age) for two different links: (1) the original variables with original links and (2) the synthesized variables with synthesized links. The synthetic data shows greater density for age differences between 0 and 1 years and less density for age differences between 1 and 3 years. The density in the original data has a severe drop between 0 and 1 due to the length of time it takes to complete a new pregnancy after giving birth. The synthetic data fails to replicate this biological relationship and could likely be improved by hard-coding a penalty or restriction on linking siblings between 0 and 9 months apart.

Figure 11: Univariate Distribution of Older Sibling Age minus Younger Sibling Age

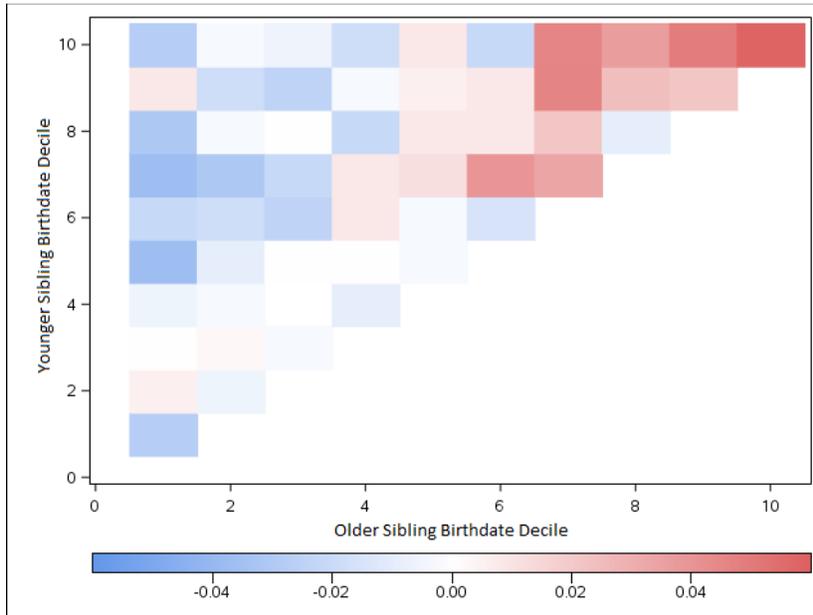


Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

Figure 12 shows the two-way k-marginal heat map for SIPP birthdate for linked siblings. Birthdate is split into ten deciles based on the distribution of observed birthdates for all linked siblings in the original data. The heat map is based on all sibling pairs and shows the older sibling on the horizontal axis. Along both dimensions, the synthetic density appears to be shifted toward higher deciles of birth dates; that is, younger children. Thus, the figure suggests that younger children in general are being linked together as siblings more often in the synthetic data than the original.

Finally, column (3) of Table 3 shows the overall two-way k-marginal scores for linked siblings. The synthetic sibling links have larger scores than the other two linkages, suggesting that sibling links do a worse job of preserving two-way correlations. However, the sibling synthesis error scores are still only about 65% larger than the two-sample baseline comparison scores, illustrating again the importance of considering the size of synthesis error relative to sampling error.

Figure 12: Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Sibling Birthdate



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

5.5 The synthetic household

Finally, we consider the entire household created by synthesizing both the spouse and mother-child links. In this analysis, since the GSF only contains the spouse link and the mother-child link, what we refer to as a household is every cluster where an adult woman is linked to anyone else (children and/or husband). Table 5 shows some summary statistics for these households in the original data (columns 1 and 2) and the synthetic data with synthetic links (columns 3 and 4).

The first characteristic, household size, is a direct result of attributes modeled at the nodal level governing whether to match a woman to a husband and how many children a woman has. All the other household summaries depend on the quality of the synthetic edges. We look at racial composition of the household, nativity of household, educational attainment of the adults in the household, and the work status of the adults in the household. Most of the statistics are very close both from the eyeball test, and from the degree of overlap in their 95% confidence intervals (a standard often used in evaluating the quality of synthetic data).

Table 5: Household Characteristics

	(1) Q-Original	(2) 95% CI-Original	(3) Q-Synth	(4) 95% CI-Synth
Household size				
=2	0.894	(0.8890,0.8990)	0.895	(0.8869,0.9031)
=3	0.0934	(0.08866,0.09814)	0.0903	(0.08454,0.1015)
=4	0.01211	(0.01032,0.01389)	0.01157	(0.00909,0.01406)
>=5	0.00048	(0.00013,0.00084)	0.0004	(-0.00002,0.00082)
Number of different races in household (white, black, or other)				
=1	0.9527	(0.9493,0.9562)	0.9469	(0.9387,0.9550)
=2	0.04677	(0.04333,0.05021)	0.05255	(0.04431,0.06078)
=3	0.00048	(0.00013,0.00084)	0.00058	(0.00007,0.00109)
Spousal/mother education				
married couple, no college degrees	0.5367	(0.5285,0.5448)	0.519	(0.5099,0.5281)
married couple, one college degree	0.204	(0.1975,0.2106)	0.2413	(0.2335,0.2491)
married couple, two college degrees	0.2075	(0.2009,0.2141)	0.1832	(0.1759,0.1904)
single mother, no college degree	0.04359	(0.04026,0.04692)	0.04134	(0.03619,0.0465)
single mother, college degree	0.00823	(0.00676,0.00971)	0.01511	(0.01214,0.01808)
Spousal/mother work status				
married couple, neither working	0.1732	(0.1671,0.1794)	0.1602	(0.1461,0.1743)
married couple, one working	0.2982	(0.2907,0.3056)	0.3203	(0.3074,0.3330)
married couple, both working	0.4768	(0.4686,0.4849)	0.4631	(0.4437,0.4826)
single mother, not working	0.0128	(0.01097,0.01463)	0.0147	(0.01227,0.01712)
single mother, working	0.03902	(0.03586,0.04218)	0.04176	(0.03622,0.04730)
Foreign born in household				
none	0.9631	(0.9601,0.9662)	0.9646	(0.9597,0.9695)
some, but not all	0.02096	(0.01863,0.02330)	0.02095	(0.01618,0.02573)
all	0.01591	(0.01387,0.01795)	0.01444	(0.01208,0.01680)

Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA). Census Bureau DRB approval: CBDRB-FY20-287.

6 Conclusion

Synthetic data has become a popular way for data providers to address the simultaneous increase in demand for microdata from researchers, and, from intruders, the increase in access to external data, computing power, and sophisticated techniques for re-identification and reconstruction. Methods for synthesizing attributes of people and firms have improved dramatically over time and many synthetic microdata products already exist. However, these methods typically do not take into account relationships between records in a data set, despite

the fact that such relationships are often of unique interest to researchers and can provide unique disclosure risks.

We provide a computationally low-cost method for synthesizing relationships between records that can be performed after synthesizing other record attributes, so it works easily with already existing methods of synthesizing nodal attributes. This method is similar to predictive mean matching. It involves drawing predicted partner attributes for a linked record based on the relationship between observed attributes of links in the internal data and the synthesized attributes in the synthetic data. We apply this method to household structures with spousal, mother-child, and sibling links. It could also be applied to a variety of other settings, including firm settings with employer-employee and co-worker links in matched employer-employee data or educational settings with school-student, teacher-student, and classmate links in matched school-teacher-student data sets.

This method struggles a little with certain deep-dives into the data, such as the discontinuity in the density of wife's share of couple's earnings at the 50% threshold and the pregnancy specific age-gap in siblings, that would be challenging for any model to recreate without explicitly controlling for them. However, the method appears to do a reasonable job of replicating most of the other characteristics of within-household links that we studied.

References

- Bertrand, Marianne, Kamenica, Emir and Pan, Jessica. (2015). Gender Identity and Relative Income within Households. *Quarterly Journal of Economics*, 130(2), 571-614.
- Benedetto, Gary, Stanley, Jordan and Totty, Evan (2018). The Creation and Use of SIPP Synthetic Beta v7.0. *CES Technical Notes Series 18-03*, Center for Economic Studies, U.S. Census Bureau.
- Hu, Jingchen, Reiter, Jerome P., and Wang, Quanli. (2018). Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data. *Bayesian Analysis*, 13(1), 183-200.
- D'Angelo, Gina M., D.C. Rao, and C. Charles Gu. (2009). Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proc.*, 3(7).
- Enders, C.K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research*, 37, 359-277.
- Enders, C.K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford.
- Howard, Waylon J., Mijke Rhemtulla, and Todd D. Little. (2015). Using Principal Components as Auxiliary Variables in Missing Data Estimation. *Multivariate Behavioral Research*, 50(3), pp 285-299.
- Little, R. J. A. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9, 407-426.
- Little, R. J. A., and D. B. Rubin. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.
- Murray-Close, Marta, and Heggeness, Misty L. (2018). Manning Up and Womaning Down: How Husbands and Wives Report their Earnings when She Earns More. *U.S. Census Bureau. SEHSD Working Paper #2018-20*.
- Reiter, Jerome P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, 29(2), 181-188.
- Reiter, Jerome P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology*, 30(2), 235-242.
- Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2), 115-124.

Royston, J. P. (1983). Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2), 121-133.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, 9, 462-468.

Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 65, 591-611.

Vogt, N.B. (1989). Polynomial Principal Component Regression: An approach to Analysis and Interpretation of Complex Mixture Relationships in Multivariate Environmental Data. *Chemometrics and Intelligent Laboratory Systems*, 7(1-2), pp. 119-130.

Appendix: Limitations to Existing Variable Selection Techniques and Potential Benefit of Using Principal Component Analysis

Common variable selection techniques include stepwise regression and best subset regression. Stepwise regression and best subset models are both automatic variable selection procedures that attempt to find the best set of predictors from a given list of potential predictors. Stepwise regression works by sequentially adding or removing variables based on some criterion. For example, a potential predictor may be added to or removed from a model based on whether its p-value is less than 0.05 when included in the model. This sequential process is repeated for every variable until no more variables can be added to or removed from the model. Best subset models estimate every possible combination of the potential predictors and then select that model that has the best fit according to some criterion, such as R-squared, the Akaike Information Criterion (AIC), or the Bayes Information Criterion (BIC).

Somewhat newer techniques include the least absolute shrinkage and selection operator (LASSO) and decision trees. LASSO performs both variable selection and regularization by forcing the sum of the absolute value of the regression coefficients on the predictor variables to be less than some threshold. This effectively forces coefficients on certain variables to zero such that the variables are no longer part of the model. Decision trees are classification techniques used to generate simple, interpretable predictions by splitting the population into many subsets based on whether their value of selected predictors falls above or below selected thresholds.

All of these approaches have their own limitations. Stepwise regression may miss important variables due to the one-at-a-time sequential nature of the testing and how this interacts with important correlations between possible predictors. Furthermore, criteria such as

p-values (or other possible criteria such as F-tests) may be misleading due to not accounting for multiple testing and due to the fact that a predictor that narrowly misses a p-value threshold may still provide important information for prediction. Best subset criterion can be more computationally intensive because it has to estimate all possible combinations of predictors. All of the approaches suffer from the fact that their objective (i.e., finding the best fitting model for the data at hand) is different than the final objective of producing high utility data after imputation or synthesis. This could lead to over-fitting on the data used to select the variables.

Another limitation of each of these approaches is that they all select only a subset of variables to use in the model and drop the rest of the variables. This potentially removes useful information if small but important variables are being dropped. Each approach ultimately uses some criterion that judges each possible predictor based on its relative contribution compared to the other variables or based on an absolute criterion. Variables that are excluded based on these criteria may still provide useful correlations for modeling certain relationships in the data. This problem is particularly relevant as the number of possible predictors increases.

This limitation may be harmful to the utility of data after imputation or synthesis. The utility of imputed data hinges on the missing at random (MAR) assumption (Little and Rubin, 2002), which requires that the incidence of missing data is uncorrelated with the missing values themselves after accounting for observed variables. That is, if observed variables predict missingness, then these variables need to be included in the imputation model. Similarly, the utility of synthetic data will depend on whether all observed variables that correlate with a given variable are accounted for in the synthesis model. If important predictors are excluded from the models, then parameter estimates on the resulting data will be biased. This makes

clear the issue related to variable selection routines keeping some variables and dropping others.

An alternative to using variable selection techniques is to keep all of the variables. While this seems a logical solution at first, this method becomes fraught as the number of observed variables increases. A large number of variables increases computational intensity and may lead to long run times. It may also lead slower convergence rates and estimation failures due to issues related to collinearity or model complexity (Enders, 2002, 2010).

We explore the feasibility of using principal component analysis to reduce the dimension of the possible predictors down to a smaller number of variables that can be used in synthesis or imputation regression models. This could be performed either on the remaining variables after variable selection, such that the factors are added to the selected predictors, or on all of the variables, such that the factors are the only predictors. This allows us to summarize all of the information into a smaller number of variables so that we reap the rewards of a more powerful regression (larger N/k) but also have every possible predictor contributing to the model even if only in a small way.

The most similar study to this one is Howard et al. (2015). The authors explore the use of PCA to reduce the dimension of possible predictors and then using the first principal component as a predictor when imputing missing data and find advantages compared to using just the original variables. We extend this approach to synthetic data models and the modeling of linkages between individuals. Other related studies include D'Angelo et al. (2009) who first use PCA for dimension reduction and then LASSO to detect gene-gene interactions in genome-

wide association studies and Vogt (1989) who uses principal components and their interaction terms from environmental chemical data to predict toxicological and ecological variables.