Journal of Privacy and Confidentiality
Vol. 13 (1) 2023

Submitted Dec 2020
Published Aug 2023

SYNTHESIZING FAMILIAL LINKAGES FOR PRIVACY IN MICRODATA

GARY BENEDETTO AND EVAN TOTTY

Center for Enterprise Dissemination, Disclosure Avoidance; U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233
e-mail address: gary.linus.benedetto@census.gov

Center for Enterprise Dissemination, Disclosure Avoidance; U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233
e-mail address: evan.scott.totty@census.gov

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.

1. Introduction

Data providers face increasing demand from researchers to provide access to the actual microdata they collect, while also facing an increasingly 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 (Abowd et al., 2020). Methods for synthesizing attributes of people and firms have improved dramatically over the past couple of decades [see, e.g., Drechsler (2011), Raab et al. (2017), Raghunathan (2021), Raghunathan et al. (2003), Reiter (2002), Reiter (2005)]. 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. For example, researchers often want to know about income mobility across generations [e.g., Chetty et al. (2014b), Chetty et al. (2020),

Key words and phrases: Synthetic Linkages, Synthetic Relationships, Synthetic Edges, Synthetic Microdata.

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, CBDRB-FY22-068, CBDRB-FY22-CED001-0008.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.767

Ferrie et al. (2021)], firm effects and co-worker effects in employee wages [e.g., Abowd et al. (2003), Cornelissen et al. (2017)], how people's behavior or income impacts their spouse [e.g., Bertrand et al. (2015), Cesarni et al. (2017)], and the effect of schools, teachers, and classmates on student outcomes [e.g., Chetty et al. (2014a), Sacerdote (2011), Totty (2020)]. These linkages can be difficult to model, but still can present significant disclosure risk due to, for example, very large families, the pattern of how many jobs an employee holds over multiple years, large numbers of divorces and marriages, and the set of schools in which a student is enrolled over time.

In this paper, we describe a non-formally private algorithm we developed to model such relationships that operates smoothly within the standard approach to generating synthetic data. This algorithm was used to create synthetic linkages between records in version 7.0 of the Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) (Benedetto et al., 2018; U.S. Census Bureau, 2018).[1] The SIPP Synthetic Beta is a synthetic version of a U.S. Census Bureau data set known as the SIPP Gold Standard File (GSF), which links individuals in the Survey of Income and Program Participation to administrative records from the Internal Revenue Service (IRS) and Social Security Administration (SSA). The algorithm described in this paper was used to synthesize spousal links and mother-child links between records. These links were left unsynthesized in prior versions of the SSB; only record-level attributes such as age and income were synthesized. To our knowledge, SSB version 7.0 was the first publicly available microdata set to synthesize linkages between records.

The approach is similar in concept to predictive mean matching (Little and Rubin, 2002). Given internal data sets that contain unique identifiers and attributes for two groups that share a link (e.g., husbands and wives) and given an internal crosswalk that indicates the true links, we build a prediction model using attributes for which we would like to preserve the correlations between linked records when making a synthetic crosswalk. This approach is based on drawing candidate matches using the conditional multivariate Normal distribution (e.g., we draw a candidate synthetic husband based on the synthetic wife's attributes) and then searching through the set of synthetic records to find the nearest neighbor to the candidate using the Mahalanobis distance measure (e.g., we find the synthetic husband most similar to the candidate synthetic husband). We use data reduction on the variables whose correlation we hope to maintain by using principal component analysis (PCA). We describe the approach in greater detail in the text, including practical issues such as how to apply the approach to one-to-one matching versus one-to-many matching, the sort order for drawing candidate links, and sampling from the potential matches with versus without replacement.

After documenting the algorithm, we explore the utility of the synthetic links and their impact on the privacy loss incurred by releasing linkage information. As an empirical test, we synthesize familial links between individuals in the U.S. Census Bureau GSF that are from the 2008 SIPP. In order to explore the marginal contribution of synthesizing linkages on utility and privacy relative to only synthesizing the attributes, we compare results across four different data sets for many of the analyses: original attributes with original links, synthetic attributes with original links, original attributes with synthetic links, and synthetic attributes with synthetic links. We find many results that are encouraging. Specifically, the synthetic links replicate several important results such as the general pattern of husbands' earnings relative to that of their wife and the distribution of the age gap between mothers and

[1] The algorithm described in this paper uses a data reduction step that was not used in the creation of SIPP Synthetic Beta v7.0.

their children. Furthermore, using k-marginal similarity scores, we find that synthesizing the links generates similarity scores relative to the true links that are comparable to k-marginal similarity scores for two independent samples of the true data. On the other hand, we also identify several relationships that the synthetic links fail to replicate. For example, synthetic links fail to exactly replicate the sharp discontinuity in the distribution of wives' income as a share of household income that occurs at 0.5 in the true data. The synthetic links also fail to exactly replicate the tendency of a husband and wife to often be nearly the same age. When evaluating the household structure as a whole (e.g., household size, racial composition, nativity, educational attainment of the adults in the household, and work status of the adults in the household) we find that the statistic based on the internal, confidential data falls within the 95% confidence interval from the synthetic data for 17 out of 22 statistics.

From a privacy loss perspective, the original links are re-created for only approximately 0.5% of links. To further assess the marginal impact of synthesizing links between records on disclosure risk, we perform a minimum distance re-identification experiment: for a given record with synthesized person attributes, we attempt to re-identify the record in the original data corresponding to that synthetic record.[2] We perform this re-identification on a data set with original links and another data set with synthetic links created using our algorithm. We find that synthesizing the links significantly reduces the likelihood of re-identifying the confidential record. Collectively, our results for accuracy, link re-creation, and re-identification suggest that our approach does a good job of maintaining correlations between linked records while significantly reducing the disclosure risk associated with releasing link information.

An important feature of our approach is that it divorces the approach for synthesizing links between records from the approach for synthesizing record attributes.[3] Our approach can therefore be applied to databases in which most of the attributes are not synthetic or are only partially synthetic. Another advantage is that our approach is relatively simple to apply compared to other options. Hu et al. (2018) suggest an approach for modeling links between records (or settings in which "the data comprise units nested within groups," in the language from their paper) using Dirichlet Process mixture models. Their approach simultaneously synthesizes both the unit and group attributes. The authors only evaluate their method for categorical variables, whereas our approach can easily be applied to settings in which the variables of interest are continuous. One important limitation of our approach is that the PCA step performs best when there are many continuous variables. We note another limitation which is that this approach is not formally private and therefore does not provide quantifiable privacy loss.

The remainder of the paper proceeds as follows. Section 2 provides some background information on the connection between record linkages and graph theory as well as background information on synthetic data in general. Section 3 details our method for generating synthetic links, including a discussion of the strengths and weaknesses of our approach. Section 4 describes the data we use to test our methodology. Section 5 discusses the results for the effect of our method on the utility and disclosure risk of releasing link information. Section 6 summarizes our findings and concludes with possible directions for future research.

[2] That is, our approach to record attribute synthesis replaces the original values with synthetic ones, such that there is a one-to-one mapping from each record in the confidential data to one record in the synthetic data.

[3] We synthesize record attributes using sequential regression multiple imputation. See the discussion in Section 4 for more details.

2. Background

2.1. Graphs in Household Data.

Social scientists often refer mathematically to the data sets they use in empirical analyses as simple matrices where the rows are individuals (or establishments) and the columns are the various characteristics of these entities that have been collected. However, often the most interesting analyses involve taking advantage of relationships between these rows such as identifying spouses or children, employees to employers or co-workers, and students to teachers or classmates. Mathematically, the easiest way to represent and discuss the networks formed by these relationships is with the concept of a graph. In graph theory, the entities in the data would be referred to as nodes (or vertices) and the relationships tying these entities together are called edges. Moreover, often these edges describe a connection that has a direction (such as a mother and her child) in which case we call it a directed graph.

Figure 1 shows an example of a graph one might observe in a typical household survey. In this case, there is a house linked to a householder, and then a spouse, one child, and one grandchild are linked to the householder. This graph would occupy one record in a household-level file and four records in a person-level file. This example of a household graph would likely be found many times with different individuals throughout a large file, but there are some graphs that would be very rare or even unique (such as when there are a large number of children). This problem is a bit different than the social network problem [e.g., Ali et al. (2014), Pérez-Rosés and Sebé (2015)] in that there is much less connectivity in the full graph. Each relatively small cluster (or household) is fully connected, but the clusters are all completely disjoint from each other.

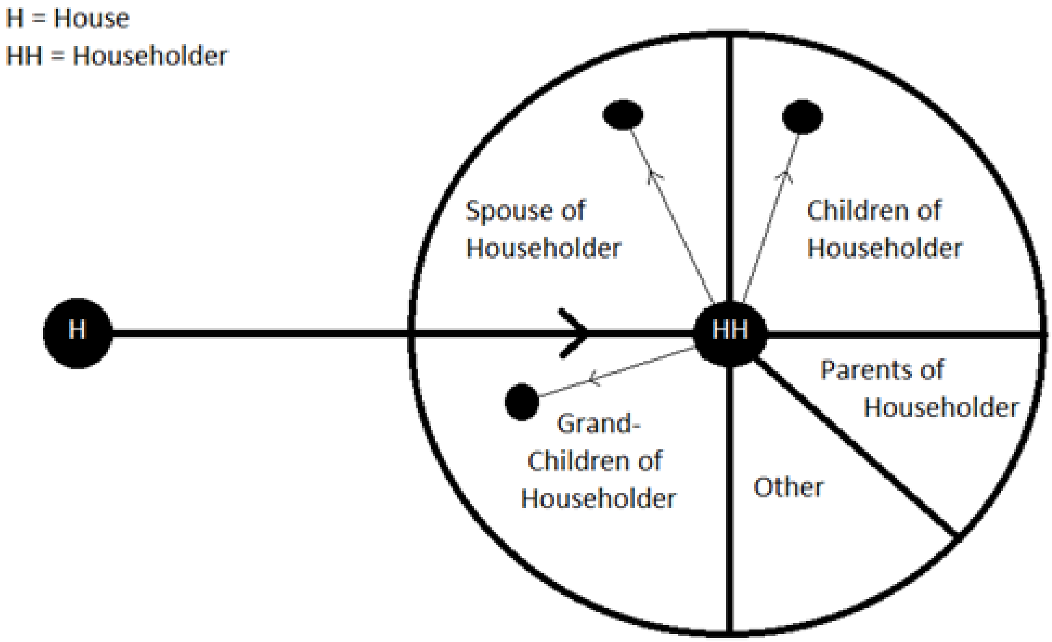
For the remainder of the paper, we will refer to the traditional variables describing characteristics of just the individuals by themselves as nodal attributes, and a file of nodal attributes as the flat file. We will refer to the linkages between nodes (rows of the flat file) as edges, and a file containing the set of edges as a crosswalk file.

2.2. 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. Reiter

Figure 1. Hypothetical Graph for Household Family Structure



(2003) presents the proper formulae for synthetic data, while Reiter (2004) also presents the proper formulae for data in which missing values are first multiply imputed and then multiple synthetic replicates are created for each completed data set.

Since then, a number of real-world applications of partially and fully synthetic data have been developed, but these do not include synthetic edges in a graph. Hu et al. (2018) suggested an approach to modeling these edges using Dirichlet Process mixture models.[4] 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 graph) 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 nodal attributes have already been synthesized, and all that remains is to synthesize the edges joining the nodes into graphs.

[4] They did not refer to these as edges in a graph, but the effect is the same.

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 nodal 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 (marital status is a nodal attribute that can be synthesized during the flat file synthesis).[5]

Consider a set of k_w vectors of nodal attributes for potential wives, $x_1, \ldots, x_{k_w}$, (these are $N_w \times 1$ vectors where N_w is the number of potential wives) and a set of k_h vectors of nodal attributes for potential husbands, $y_1, \ldots, y_{k_h}$, (these are $N_h \times 1$ vectors where N_h is the number of potential husbands) 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 sets of attributes. For example, if we were matching houses to householders, the two sets of attributes could be house characteristics and householder characteristics, respectively. In the case of spouses, these sets of attributes were the same, and, therefore, in this paper $k_w = k_h$. 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 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 $k_h^* \leq k_h$. In other words, we calculate:

(1) $X^* = XW$ 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 $k_w^* \times k_w^*$ diagonal matrix of eigenvalues of $X'X$ sorted from greatest to least (i.e., $\lambda_{1,1}^w \geq \lambda_{2,2}^w \geq \lambda_{3,3}^w \ldots$)

(2) $Y^* = YH$ 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 $k_h^* \times k_h^*$ diagonal matrix of eigenvalues of $Y'Y$ sorted from greatest to least (i.e., $\lambda_{1,1}^h \geq \lambda_{2,2}^h \geq \lambda_{3,3}^h \ldots$)

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

We use principal component analysis (PCA) for the dimension reduction. Other possible methods such as variable selection techniques can be computationally intensive, but more importantly they all select only a subset of variables. This limitation may be harmful to the utility of the data after synthesis if variables that are correlated with spousal characteristics are excluded. Another option is to keep all variables, but this approach becomes fraught as the number of variables increases. A large number of variables increases computational

[5] 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."

intensity and may lead to slow convergence, long run times, or estimation failures due to collinearity or model complexity. PCA allows us to summarize all the information into a smaller number of variables so that we reap the rewards of more powerful estimation (larger N/k) but also have every possible predictor contributing to the model even if only in a small way.

Next, we map each of these principal components independently to approximately standard normal distributions using a non-parametric transform developed by Woodcock and Benedetto (2009). 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), $\widehat{F}_x(x)$; (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, $\Phi^{-1}(\widehat{F}_x(x))$. We will denote as $T_{A^*}(A) = \widetilde{A}$ the mapping that performs this transformation to all the columns of matrix, A, independently, estimated on A^*. The transformed random variables, $\widetilde{x}^*$ and $\widetilde{y}^*$, represented by the columns of $\widetilde{X}^* = T_{X^*}(X^*)$ and $\widetilde{Y}^* = T_{Y^*}(Y^*)$ 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, $(\widetilde{x}^* \ \widetilde{y}^*) \backsim N(\mu = [\mu_w \ \mu_h], \Sigma = \begin{bmatrix} \Sigma_{ww} & \Sigma_{wh} \\ \Sigma_{hw} & \Sigma_{hh} \end{bmatrix})$. 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 $\widehat{\mu}$ and $\widehat{\Sigma}$. 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: $\widetilde{X}^*_s = T_{X^*}(X^*_s)$ and $\widetilde{Y}^*_s = T_{Y^*}(Y^*_s)$.

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: $N(\mu_h + \Sigma_{hw}\Sigma^{-1}_{ww}(\widetilde{y}^* - \mu_w), \ \ \Sigma_{hh} - \Sigma_{hw}\Sigma^{-1}_{ww}\Sigma_{wh})$. 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.[6]

3.2. **One-to-Many: Parent-Child Link.** The method we use to link one-to-many is 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[7]; however, this could also be

[6] 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 in our experience this does not improve performance.

[7] We followed the SSB in only linking children to mothers in this paper and not children to fathers. This is because of data quality issues: in early years of the SIPP the father person ID variable is often missing. Furthermore, the frequency of single fathers is quite small relative to single mothers.

Algorithm 1 One-to-one matching (husbands to wives)

```
(1) Calculate principal components of $X$ and $Y$, and keep first $k_w^*$ and $k_h^*$ respectively in $X^*$ and $Y^*$.
(2) Use eigenvectors and eigenvalues from step 1 to create corresponding $k_w^*$ and $k_h^*$ principal components in synthetic data, $X_s^*$ and $Y_s^*$.
(3) Use KDE transform to transform columns of $X^*$ and $Y^*$ into approximate standard Normals, $\widetilde{X}^*$ and $\widetilde{Y}^*$.
(4) Use same transformation to transform columns of $X_s^*$ and $Y_s^*$ into $\widetilde{X}_s^*$ and $\widetilde{Y}_s^*$.
(5) Estimate means and variance/covariance matrices of $X^*$ and $Y^*$, $\mu = [\mu_w \ \mu_h]$ and $\Sigma = \begin{bmatrix} \Sigma_{ww} & \Sigma_{wh} \\ \Sigma_{hw} & \Sigma_{hh} \end{bmatrix}$.
(6) Using $\mu$ and $\Sigma$ as the mean and variance parameters, draw predicted husbands for each synthetic wife from the assumed multivariate Normal distribution.
(7) Sequentially (after random sort) find distance (Mahalanobis) minimizing synthetic husband from each predicted synthetic husband from step 6, to generate a synthetic edge between synthetic wife and synthetic husband. Store each synthetic edge in synthetic crosswalk file.
```

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 synthesis of different types of edges should be done sequentially, conditioning on the results of prior synthetic edges. For instance, after women have been synthesized an edge connecting them to a spouse, potential edges connecting those women to children can be synthesized conditional on characteristics of these women updated with characteristics of their partners (or lack thereof).

The first few steps pertaining to getting the principal components and transforming them to standard Normal distributions remains the same. Rather than immediately generating all 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 the mother as the number of children the mother has (which should be a nodal 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.[8]

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

[8] We never ran out of candidate children because the original GSF data has children without a mother identified in the data. However, if we were to run out of candidate children, then we could simply generate extra candidate synthetic children as described in footnote 6 for the case of one-to-one matching.

an inappropriate assumption in the case of matching employers to employees, since different businesses will have different staffing needs. In such a case, we might want to synthesize, in advance, a variance group for the employer as a nodal attribute, and then perform the previous steps independently for each variance group.

Algorithm 2 One-to-many matching (mothers to children)

(1) Calculate principal components of original matrices of nodal attributes for mothers and children.
(2) Use eigenvectors and eigenvalues from step 1 to create corresponding principal components in synthetic matrices of nodal attributes for mothers and children.
(3) Use KDE transform to transform columns of original matrices into approximate standard Normals.
(4) Use same transformation to transform columns of synthetic matrices of nodal attributes.
(5) Estimate means and variance/covariance matrices of mothers' and children's nodal attributes.
(6) Using these estimates as the mean and variance parameters, draw one predicted child for each synthetic mother from the assumed multivariate normal distribution.
(7) Sort children of each mother by descending age in original data.
(8) Estimate covariance matrix of nodal attributes for a sibling from the next oldest sibling in original data.
(9) Using these estimates as the parameters of the assumed multivariate normal, draw one predicted, next oldest child for the most recently predicted child. Repeat until the number of predicted children matches the synthetic number of children from the nodal attributes for each mother.
(10) Sequentially (after random sort) find distance (Mahalanobis) minimizing synthetic child from each predicted synthetic child, to generate a synthetic edge between synthetic mother and synthetic children. Store these edges in crosswalk.

3.3. Discussion: Strengths and Weaknesses Relative to Alternative Methods.

The primary benefit of this method is the simplicity and computational low cost of the algorithm. There is no sophisticated optimization routine, for which the dimensionality of all possible graphs would present a major challenge, nor does it require any machine learning. We are simply matching together existing records in a way that mimics the existing matches. The search for the minimum distance match is the most time-consuming part of the process, and that could be sped up by breaking the potential pairs into smaller sets that can run in parallel.[9] As a result, we can lean on the strength of the synthesis of the flat file for which many methods exist and have proven to provide both sufficient privacy protection and sufficient accuracy.

Another important feature of this algorithm is that, since it simply matches existing records together into a graph, it can be used to create partially synthetic data, even to the point where the only synthetic feature is the graph. Alternate methods that currently exist simultaneously synthesize the characteristics of the individuals and the graph joining

[9] The algorithm, programmed in SAS, created four implicates of synthetic spouses for about 27,000 potential partners in 13:38.55 of real time using a scheduler limiting the job to at most 4 processors and 1 GB RAM from a shared compute node with 64 CPUs available and 754 GB RAM available.

individuals. Many real-world microdata products are protected using partial synthesis of just a few particularly sensitive variables and/or records.[10] If the data provider wants to pursue the route of using partial synthesis to protect the microdata, and if the data provider is primarily worried about the sparsity of certain types of graphs, then this method can be used to create partially synthetic data while keeping the characteristics in the existing records otherwise unperturbed.

One important limitation of our approach is that the PCA step performs better when there are many continuous variables available to condition upon in the nodal attributes. This can be partially addressed by including higher order terms and interactions between continuous variables. Furthermore, the advantage of our approach in terms of statistical simplicity and low computational cost may become a weakness as the complexity of the graphs in the confidential data grows. As a result, this method is more appropriate for use cases where the first priority is the accuracy of the synthetic flat file, and the edges are a feature of secondary importance. Alternative methods that make use of synthetic network generation techniques may be more appropriate for data sets with many more edges (such as social networks) and possible graph types (Ali et al., 2014; Pérez-Rosés and Sebé, 2015). Finally, this approach is not formally private.

4. Data

We used the SIPP 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 IRS and 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 (version 7.0) includes synthetic familial links using the methodology described in this paper.[11] 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 more thoroughly assess our methodology. The sample is based on individuals from the 2008 SIPP panel with non-missing information for race, Hispanic status, foreign born status and time of arrival in the USA, education level, homeowner 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 nodal attributes used for synthesizing the links are shown in Table 1.[12] The years 2009 and 2010 were used because they are the first two full reference years in the 2008 SIPP panel.[13] Because of data quality concerns and to limit the scope of this paper, we restrict ourselves to directed graphs linking husbands to wives and mothers to children.[14] Figure 2 illustrates

[10] For example, the American Community Survey public use microdata uses partial synthesis for group quarters (Hawala, 2008).

[11] The only difference is that in this paper we use PCA for data reduction of the variables used to create the links, whereas SSB v7.0 had no data reduction.

[12] Categorical variables were turned into dummy variables for use in the PCA data reduction step.

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

[14] Mothers in the SIPP data include biological, step, adopted, and foster mothers.

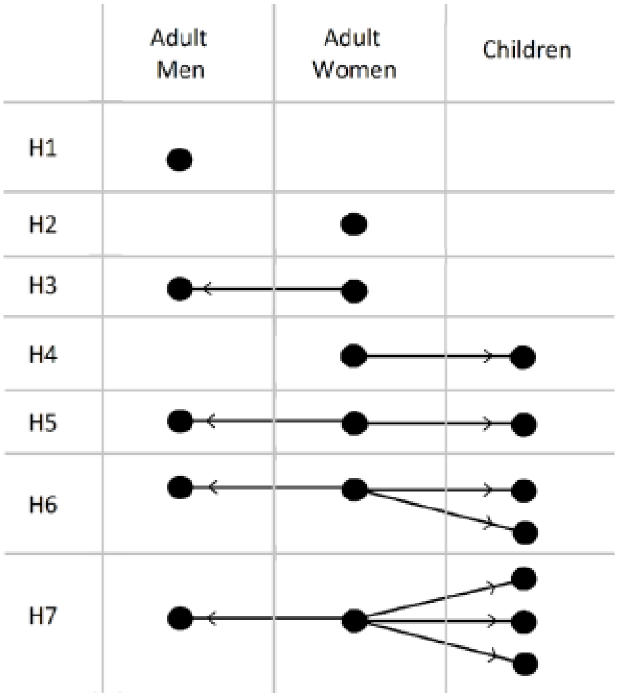
some examples of possible household graphs in this setting (where a household identifier is represented in the first column).

Table 1. Nodal Attributes Used in SIPP Synthetic Familial Linkages

Variable	Description	Range
Married	Recode - Indicator for married	Binary (0/1)
Kids	Recode – Count of kids linked to adult woman	Count
Hispanic	Hispanic ethnicity	Binary (0/1)
foreign_born	Foreign born	Binary (0/1)
time_arrive_usa (& squared)	Date of arrival in USA (days since Jan. 1, 1960)	Continuous
total_der_fica_2009 (& squared)	SSA Detailed Earnings Record – 2009 earnings	Continuous
total_der_fica_2010	SSA Detailed Earnings Record – 2010 earnings	Continuous
sipp_birthdate (& squared & cubed)	Birthdate in the SIPP (days since Jan. 1, 1960)	Continuous
Nonwhite	Recode – non-White race	Binary (0/1)
Black	Recode – Black race	Binary (0/1)
pos_der_fica_2009	Indicator for positive 2009 DER earnings	Binary (0/1)
pos_der_fica_2010	Indicator for positive 2010 DER earnings	Binary (0/1)
educ_d1-educ_d5	Indicators five-category highest education level	1-5
sipp_birthdate X total_der_fica_2009	Interaction term	Continuous
sipp_birthdate X educ_5cat	Interaction term	Continuous
total_der_fica_2009 X educ_5cat	Interaction term	Continuous

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

Figure 2. Example Family Graphs in Our Data



We synthesized both the links (edges in the graph) and the nodal attributes themselves separately. The nodal attributes are synthesized using a sequential regression technique.

More specifically, we build up the synthetic data as a series of conditional marginals, using only previously synthesized variables as explanatory variables. After estimating the model, we impute a value for each variable based upon the most up-to-date synthetic data. Hence while the synthetic variables are not used in the model estimation, they are used to impute other synthetic values in order to keep the synthetic data internally consistent. In implementing this sequential regression approach, we made four decisions for each variable that was synthesized. First, we chose what type of model to use (ordinary least squares, logistic, Bayesian bootstrap); second, we designated parent-child relationships among variables; third, we defined restrictions to be placed on the values of variables when necessary; fourth, we chose a set of grouping and conditioning variables to use in modeling. More information on the attribute synthesis can be found in Benedetto et al. (2018).

For the sake of differentiating between the effects of synthesizing the nodal attributes and synthesizing the edges, we use four different versions of the microdata in the analysis below: (1) original edges with original attributes, (2) synthetic attributes with synthetic edges, (3) synthetic attributes with original edges[15], and (4) original attributes with synthetic edges (just for spousal links). We generated four replicates for each synthesis.[16]

5. Results

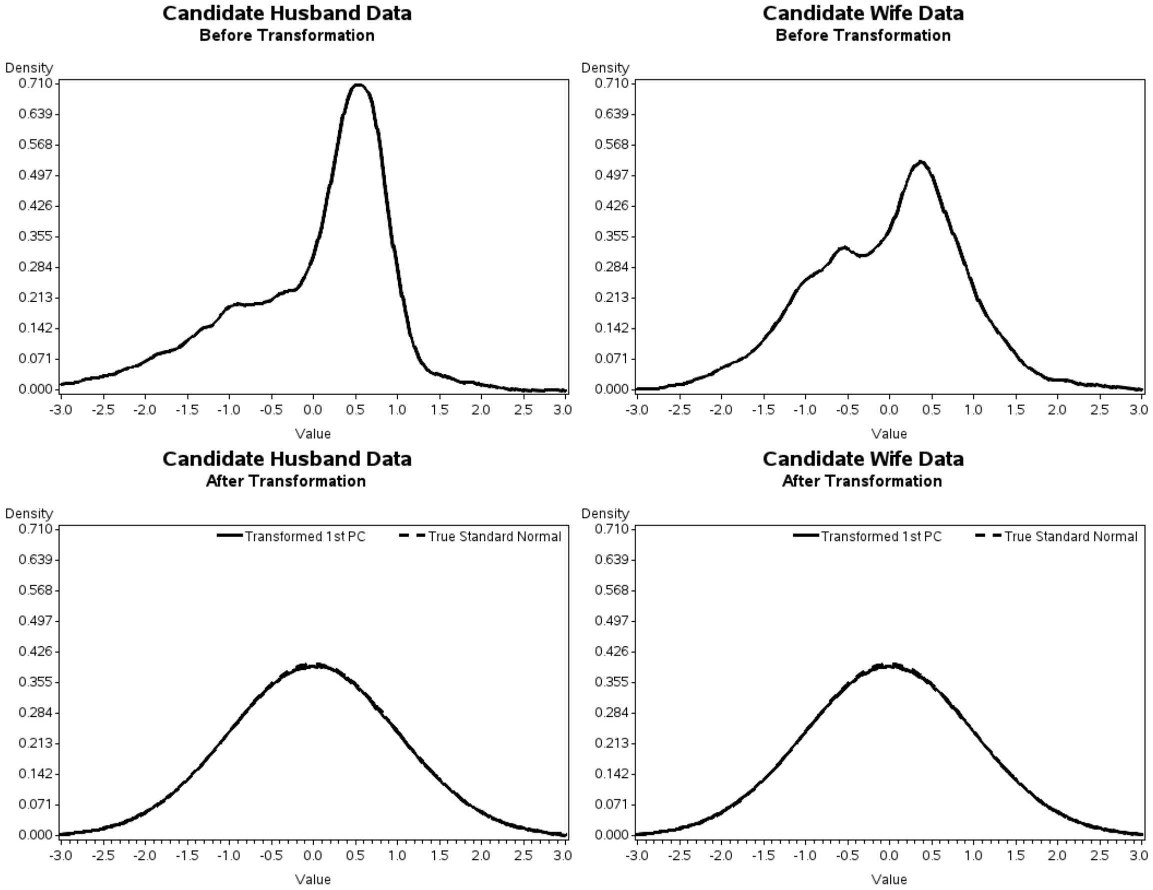
5.1. **Test for Multivariate Normality.** We first analyze the quality of the assumption of normality for the transformed principal components before turning to results for utility and privacy. Figure 3 shows the univariate density distribution for the first principal component from the potential husbands and potential wives, before and after transformation. The panels of Figure 3 illustrate that the principal components do not resemble a standard Normal distribution before transformation but do afterwards.

Figure 4 shows the bivariate density of the first principal component for husbands and wives, before and after transformation. For reference, the lower panel of Figure 4 displays a bivariate Normal density, with the covariance matrix set to the sample covariance matrix of the transformed principal components. The distribution after transformation does not perfectly resemble a bivariate Normal density (there is a ridge in the surface creating a bit

[15] The original edges are able to be maintained despite synthesizing the attributes because we retain the original person ID, spousal ID, and mother/child ID while replacing the original attributes with synthetic ones. I.e., we effectively perform partial synthesis in which the familial crosswalk is left unsynthesized. Note that the first two variables in Table 1 – marital status and number of kids – are left unsynthesized for this method. Furthermore, when we synthesize the attributes, we use all available information, including link characteristics. I.e.., for the data with original edges and synthesized attributes, we condition on presence of spouse/kids and any attributes already synthesized for that spouse/kid. This allows the method with original edges and synthesized attributes to still perform well despite maintaining original links.

[16] We used four total replicates (as did the production SSB) for two reasons. First, while more replicates are better for accuracy of both the point estimate and the actual components of variance, more synthetic replicates also represent an increased privacy risk since accuracy and privacy form a necessary trade-off. Second, we wanted to evaluate the properties of the proposed approach in realistic production settings for the number of replicates. In practice, the computational cost to the researcher can become intractable if the number of replicates grows too large. For the SSB, the synthetic data simulates the missing data pattern of the underlying data in addition to the values of the non-missing data, and researchers are encouraged to use multiple imputation to properly address the uncertainty introduced by missing data. If the researcher chooses to create M=4 completed implicates per synthetic replicate (and there are R synthetic replicates), then the researcher is performing the analysis on M*R total implicates.

Figure 3. Univariate Distribution of First Principal Component



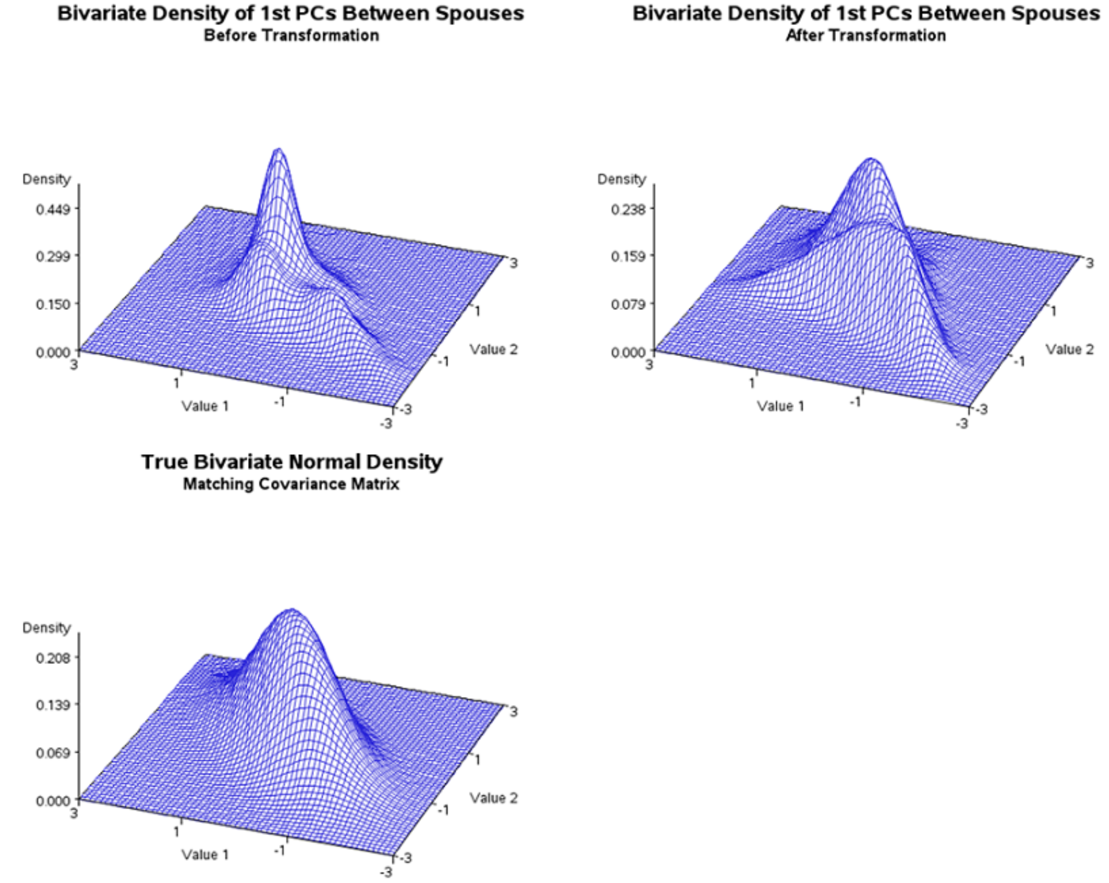
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

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.

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. Royston (1982) suggests that the test only be used for samples up to 2,000. After estimating the principal components, transforming them to standard normal, and linking couples

Figure 4. Bivariate Distribution of Husband's and Wife's First Principal Component



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

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.[17]

5.2. **One-to-one: Spousal link.** We begin our analysis of the utility of synthetic edges by analyzing the one-to-one spousal links. We evaluate utility in a few different ways. First, we check the distance between predicted spouses and linked spouses for the real and synthetic

[17] 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 when the multivariate normality assumption is reasonable and when it is not.

TABLE 2. Multivariate Normality Test for Husband and Wife Principal Components

	(1) Husbands	(2) Wives	(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).

links. Second, we check the distribution of key variables of interest, such as the distribution of spousal age difference and relative earnings of spouses using the real and synthetic spouses. 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.

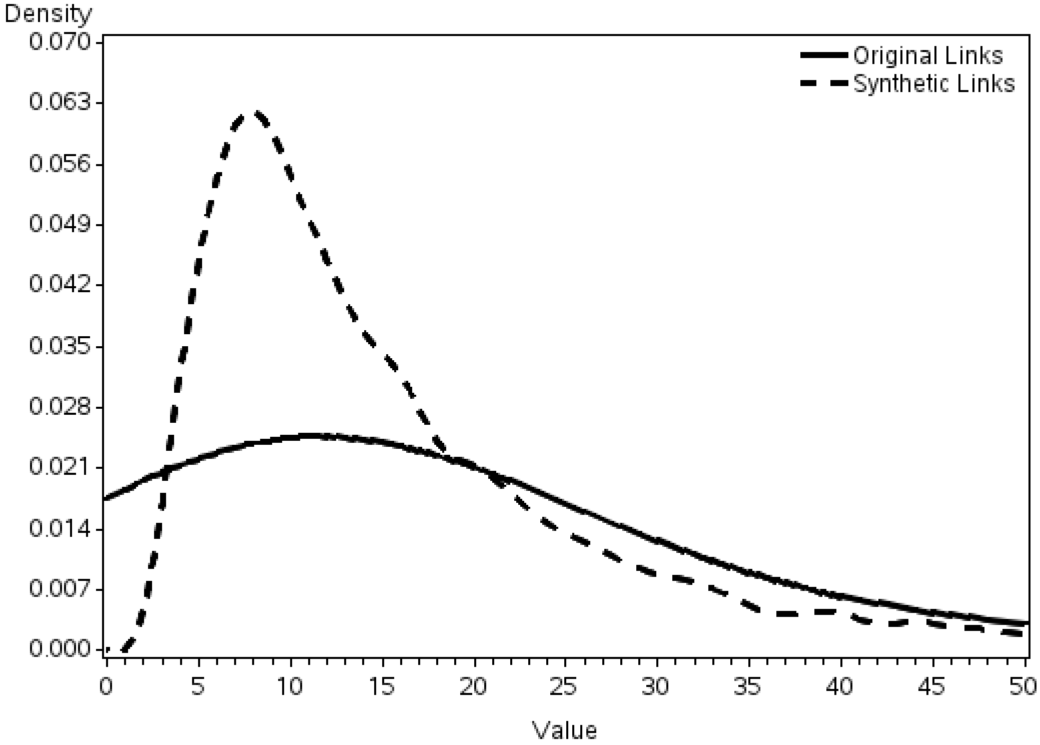
As described above, the synthesis process for spousal links involves randomly sorting the wives, drawing a candidate husband for each wife using the conditional multivariate normal distribution, and then finding the nearest neighbor to the candidate husband using the Mahalanobis distance measure. Figure 5 shows the univariate distribution of Mahalanobis distance between the predicted and linked spouses for both the real and synthetic linkages.[18] 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 20-50. 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 6 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 replicate the tendency of spouses to be of similar age and for the husband to often be slightly older than the wife. The synthetic densities are somewhat flatter. As is apparent in the figure, this appears to be mostly due to a relative reduction in the frequency of spouses with an age gap of -2 to 5 and a relative increase in the frequency of spouses with an age gap of -10 to -3 and 5 to 10. 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.

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 et al., 2015). Other work has shown that this pattern may be partially due to survey misreporting related to gender norms (Murray-Close and Heggeness, 2019).

[18] That is, for both the real and synthetic links, we compute the Mahalanobis distance between a spouse's linked partner and their predicted partner. Assessing the difference between predicted spouses and actual spouses for both types of links is useful for assessing the extent to which real spousal links may be driven by characteristics that are unobserved to our model.

Figure 5. Distance of Predicted Mean Spouse to Linked Spouse



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

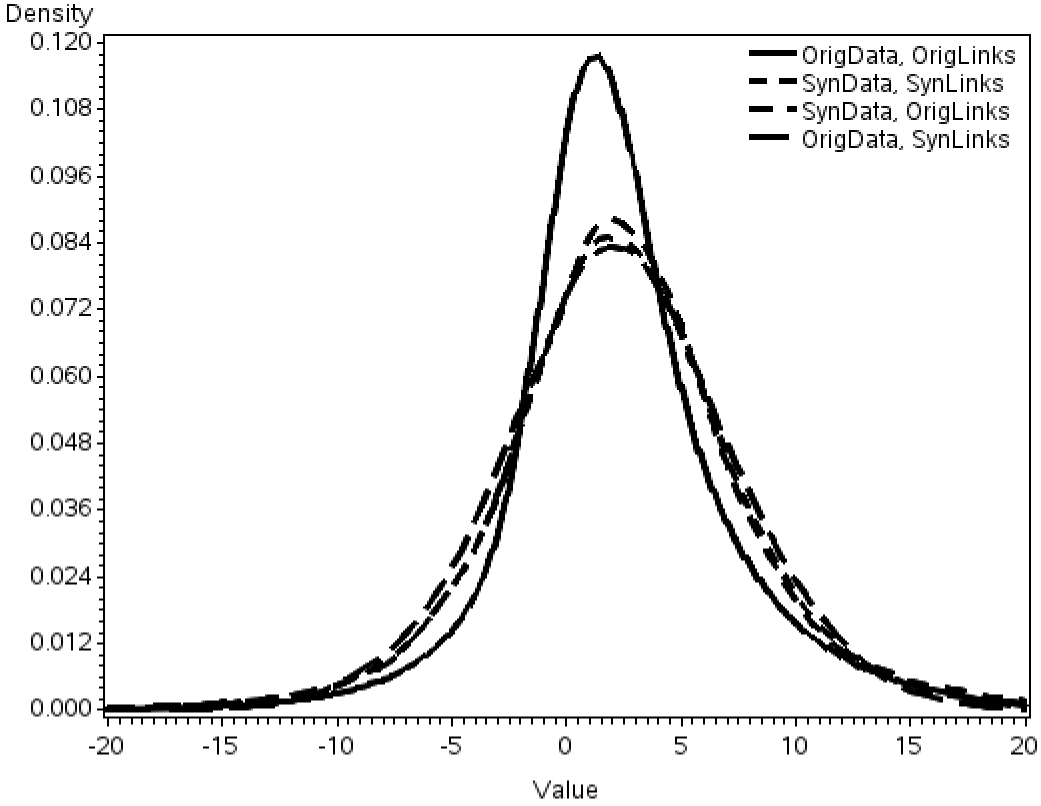
Analogous to Figure 1 in Bertrand et al. (2015), Figure 7 shows the univariate distribution of the share of spousal earnings that are earned by the wife for the same four sets of spousal links used for age differences: (1) the original variables with original links, (2) synthesized variables with synthesized links, (3) synthesized variables with original links, and (4) original variables with synthesized links.[19]

The original-original version shows the stark discontinuous drop in density around the 50% threshold. Synthesizing only the nodal attributes with existing links or only the links for existing nodal attributes fails to replicate this discontinuous drop. The figure based on synthesized attributes and synthesized links does show a drop in density across the 50% threshold, although it is a much smaller drop than on the original data. It is perhaps not surprising to see that the synthetic edges struggle to perfectly replicate this difference in density around an otherwise arbitrary threshold, particularly when the modeling does not attempt to account for it directly.

Figure 8 illustrates that our method shows some improvement over other methods used in prior versions of the SSB. Specifically, it performs better in the tails and also shows evidence of a small drop in density across the 0.5 threshold rather than the opposite as seen in version 7. Version 6 of the SSB left the first spousal link unsynthesized. Version 7

[19] Bertrand et al. (2015), Figure 1, shows the results as estimated on the GSF, after the authors built their code on the SSB. Figure 7, top left panel, shows the analog to their published figure. The bottom right panel shows the figure they would have obtained while using our synthesis method.

FIGURE 6. Univariate Distribution of Husband's Age minus Wife's Age



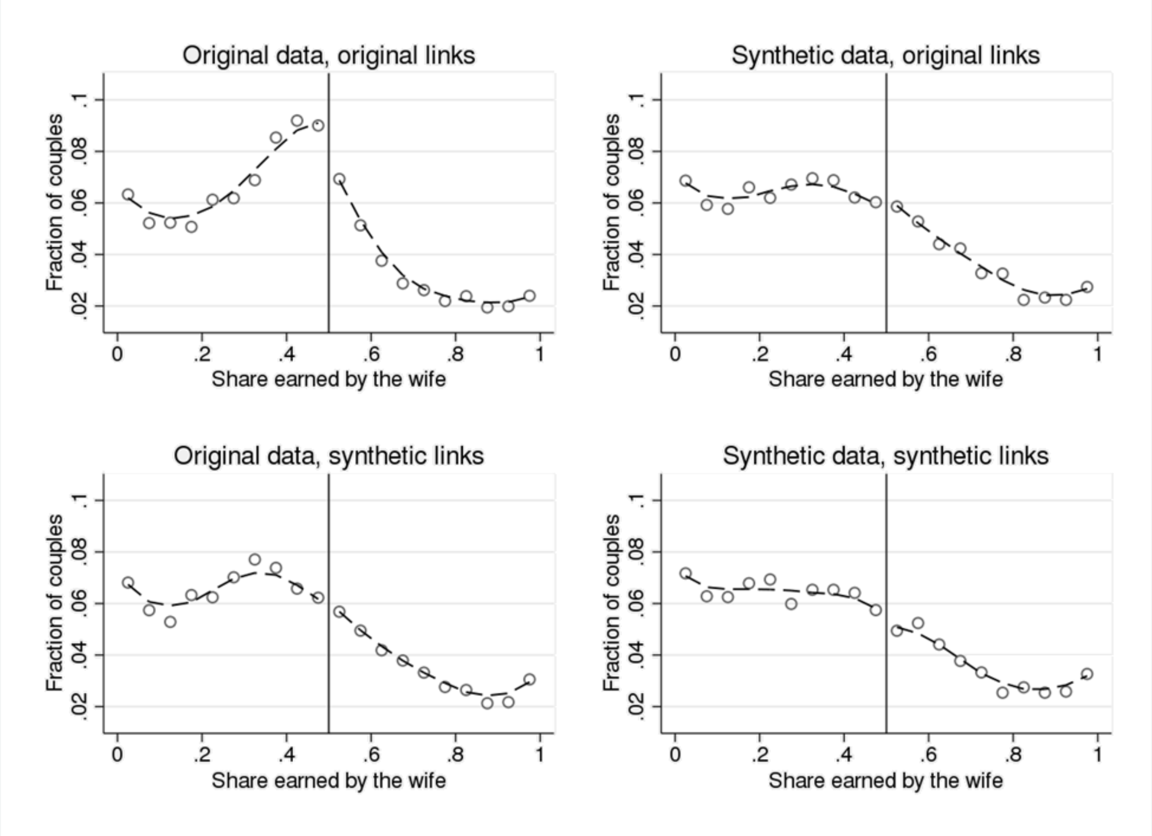
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

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 may provide meaningful improvements.

Next, we use two-way k-marginals to evaluate similarity in the bivariate distribution of eight 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.[20] This is done on both the real set of couples and the synthetic couples, and then the absolute value of the difference in the density between the real and synthetic data

[20] 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. Share of Spousal Earnings Earned by the Wife



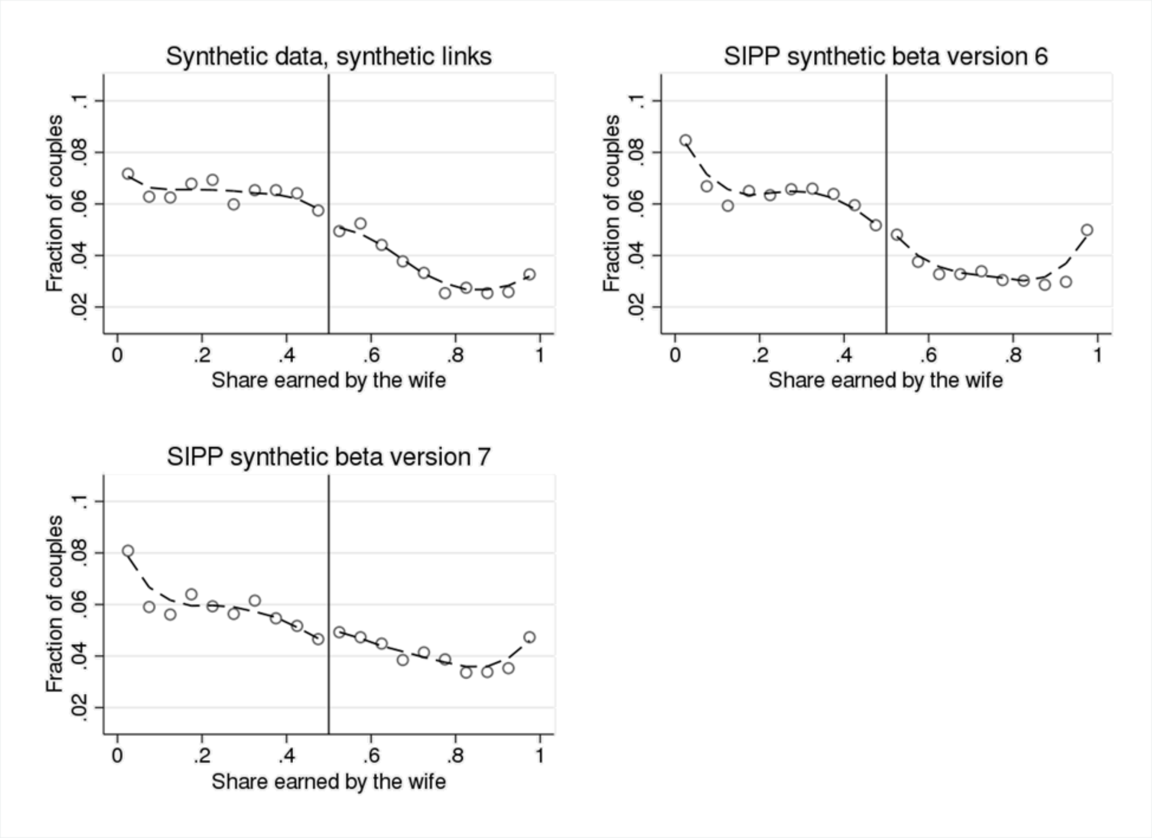
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

in each cell is summed across all cells to create the k-marginal score. 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 small. The lowest possible score is 0, which indicates identical distributions for the given variables and categories. The highest possible score is 2, which indicates two distributions with zero overlap.

The score was constructed for 64 (8 by 8) 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. We first highlight some specific two-way k-marginal results before reporting the overall two-way k-marginal score by averaging across all combinations.

Figure 9 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

Figure 8. Share of Spousal Earnings Earned by the Wife in SIPP Synthetic Beta



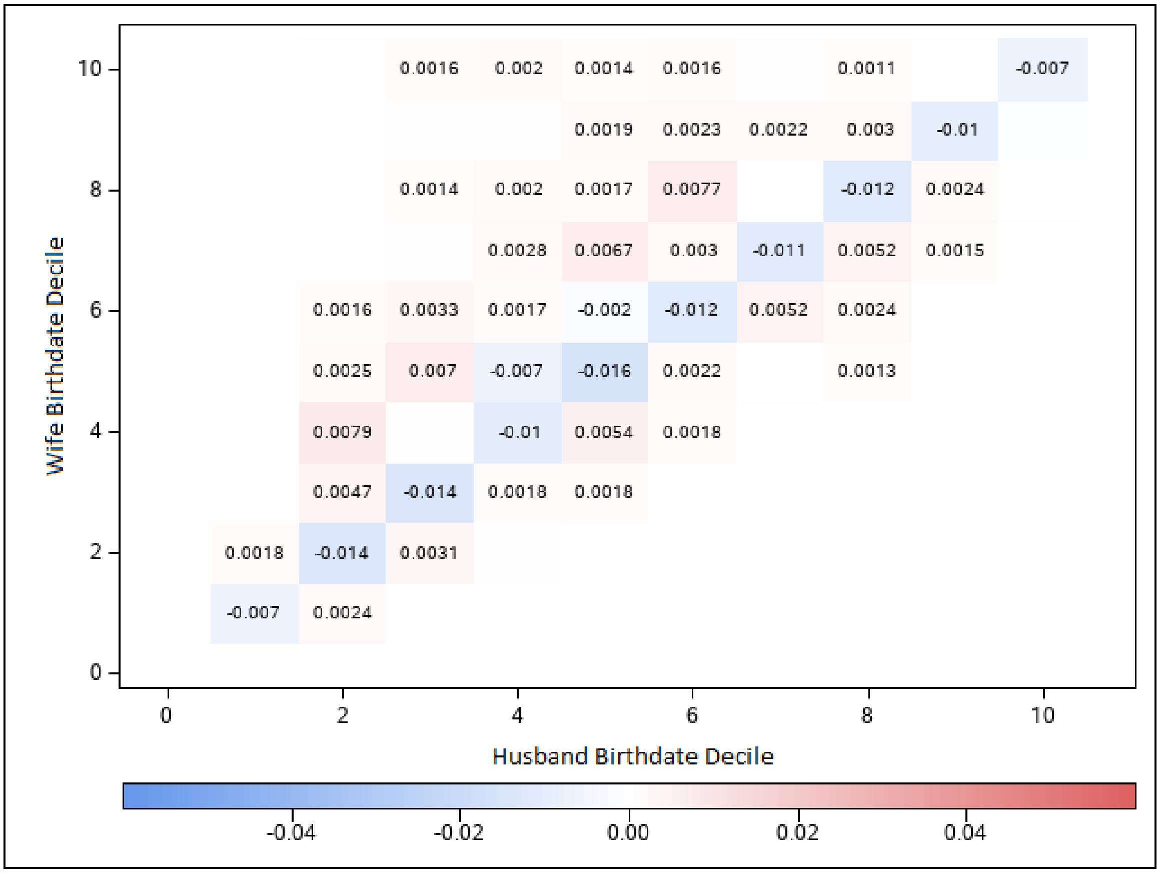
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

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 6, where the synthetic data shows less density around the 0 age difference region and more density around the (-10,-2) and (5,13) regions.

Figure 10 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 7: 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.[21] 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.

Figure 9 and Figure 10 are interesting because they show the same interpretable patterns as Figure 6 and Figure 7, respectively. This illustrates the flexibility and usefulness of k-marginal heatmaps. While Figure 6 and Figure 7 were created using code generated from scratch and tailored for that specific comparison, Figure 9 and Figure 10 were created using generic heatmap creation code that can be run for any two variables. Thus, heatmaps offer a fast and easy way to evaluate correlations between original and synthetic data.

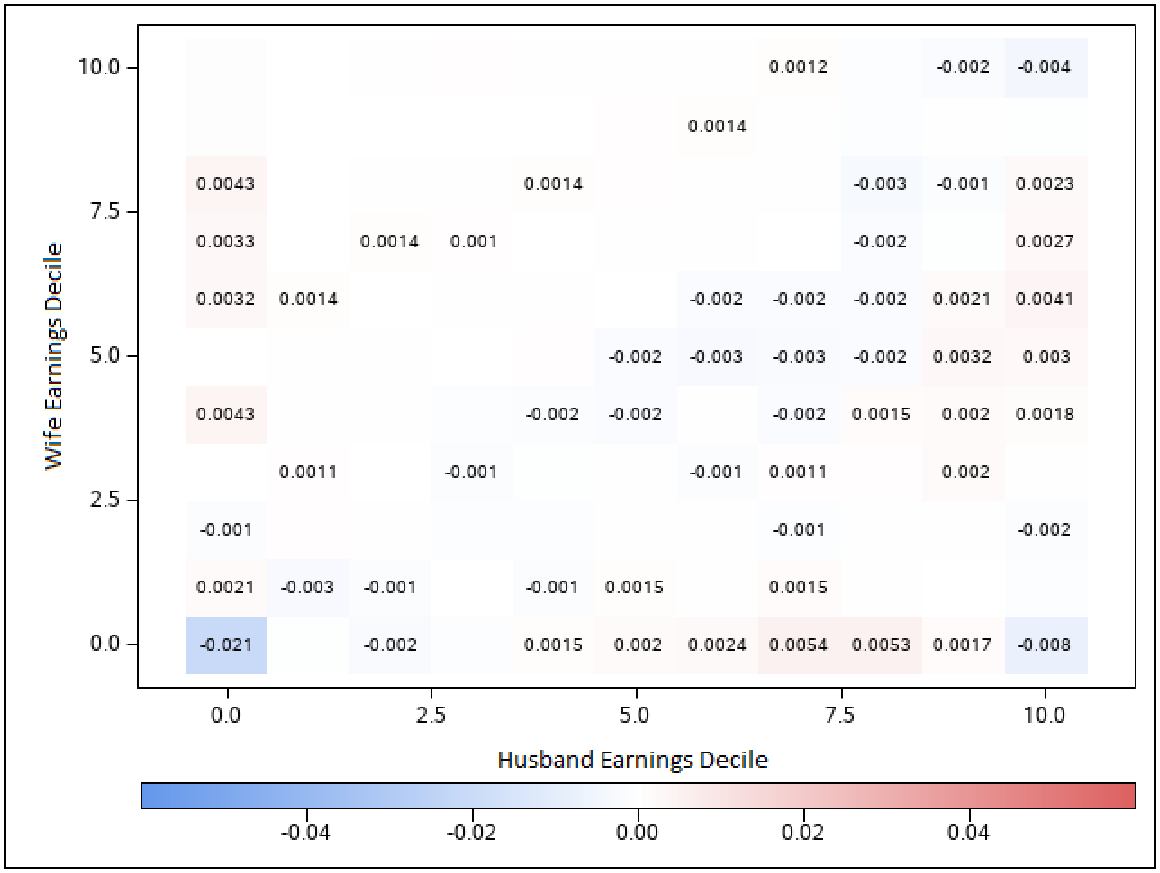
Figure 9. Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Spousal Birthdate



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).
Note: Empty cells have data, but the table only reports cells with an absolute density difference of at least 0.001.

[21] 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 10. Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Spousal Earnings

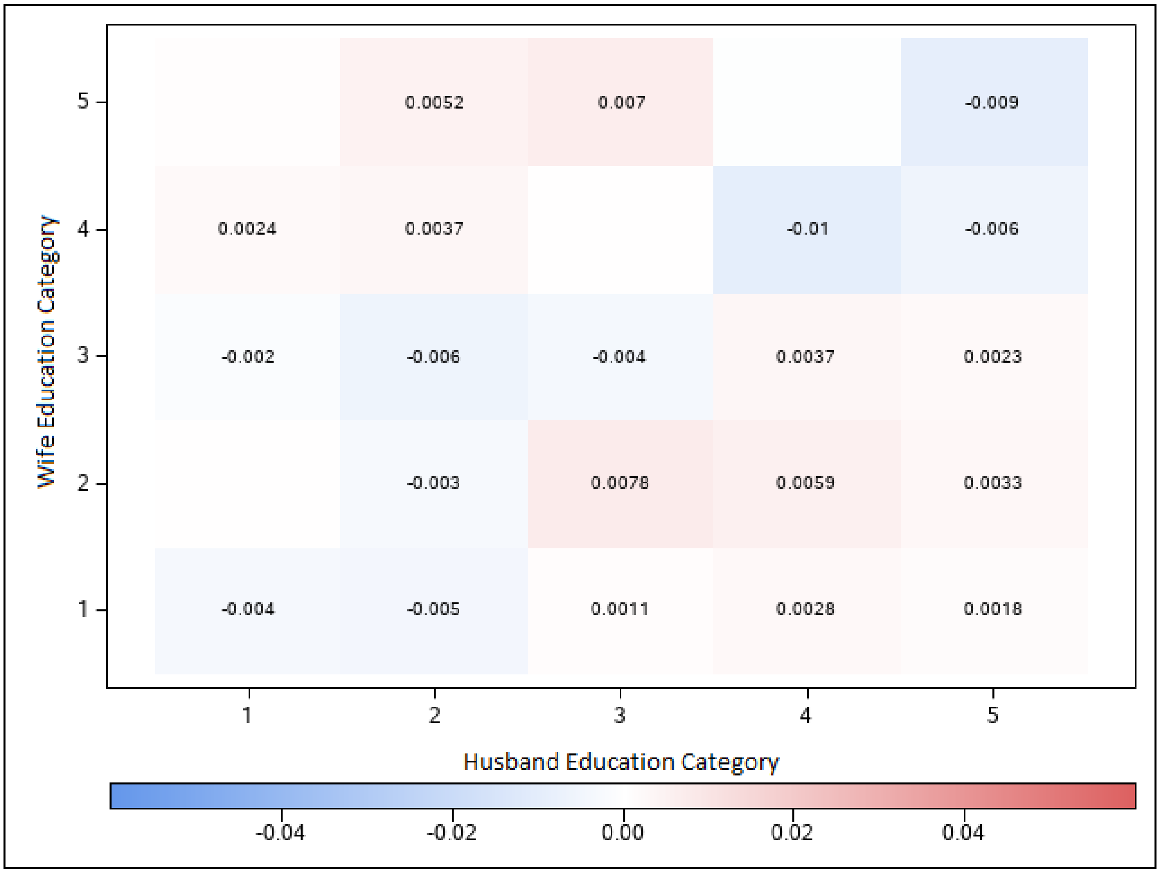


Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).
Note: Empty cells have data, but the table only reports cells with an absolute density difference of at least 0.001.

Figure 11 shows a heatmap for spousal education levels. Once again, we see less density in the synthetic data along the diagonal, which corresponds to spouses with the same level of education. A clear pattern across the three heatmaps is that while the synthetic data appear to do a good job of replicating spousal characteristics overall, it fails to fully capture the tendency of couples to often be nearly identical in particular characteristics. This suggests to us that there are complex socio-economic aspects of spousal pairings that are difficult to capture. The tendency of spouses to often be nearly the same age and to have the same education level likely reflects the presence of real-life interactions through social networks, while the tendency of wives to have earnings just below that of their husband likely reflects remnants of traditional gender norms (Bertrand et al., 2015).

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

Figure 11. Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Spousal Education



Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).
Note: Empty cells have data, but the table only reports cells with an absolute density difference of at least 0.001.

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. 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 baseline comparison using two 50% samples of the data and approximately 100% larger than the baseline comparison of a 50% sample of the data to the full data.

5.3. **One-to-many: Parent-child link.** Figure 12 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

Table 3. Two-Way K-Marginal Scores for Spousal Linkages

Synthetic File	Comparison File	(1) Spousal Links Score
Synthetic both #1	Original data	0.066
Synthetic both #2	Original data	0.062
Synthetic both #3	Original data	0.064
Synthetic both #4	Original data	0.064
Original data 50% Sample	Original data	0.027
Original data 50% Sample	Original data 50% sample	0.041

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

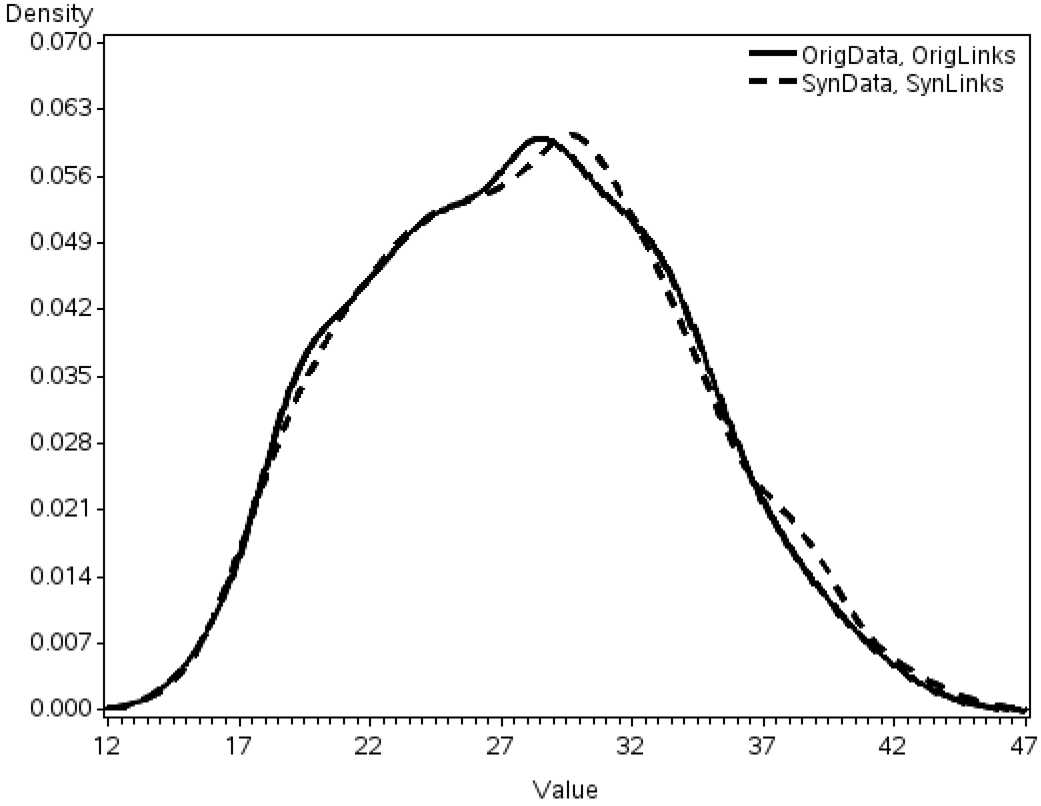
variables with synthesized links.[22] The synthetic version does a good job of matching the distribution of age differences between mothers and children. Figure 13 shows the two-way k-marginal heat map for SIPP birthdate for linked mothers and children. Birthdate is split into ten deciles based on the distribution of observed birthdates for all linked mothers and children in the original data. Greater deciles indicate later birthdates. 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. Because these results looked quite strong overall, we did not consider alternative sorting methods for the linking, such as allocating the first child for all mothers, then the second child for all mothers, and so forth. Alternative sorting could prove useful in other applications.

Table 4 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 approximately 50% larger than the baseline comparison using two 50% samples of the data and approximately 100% larger than the baseline comparison of a 50% sample of the data to the full data. The sampling-based comparisons 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 illustrates 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 14 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

[22] 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.

Figure 12. Univariate Distribution of Mother's Age minus Child's Age



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

Table 4. Two-Way K-Marginal Scores for Mother-Child Linkages

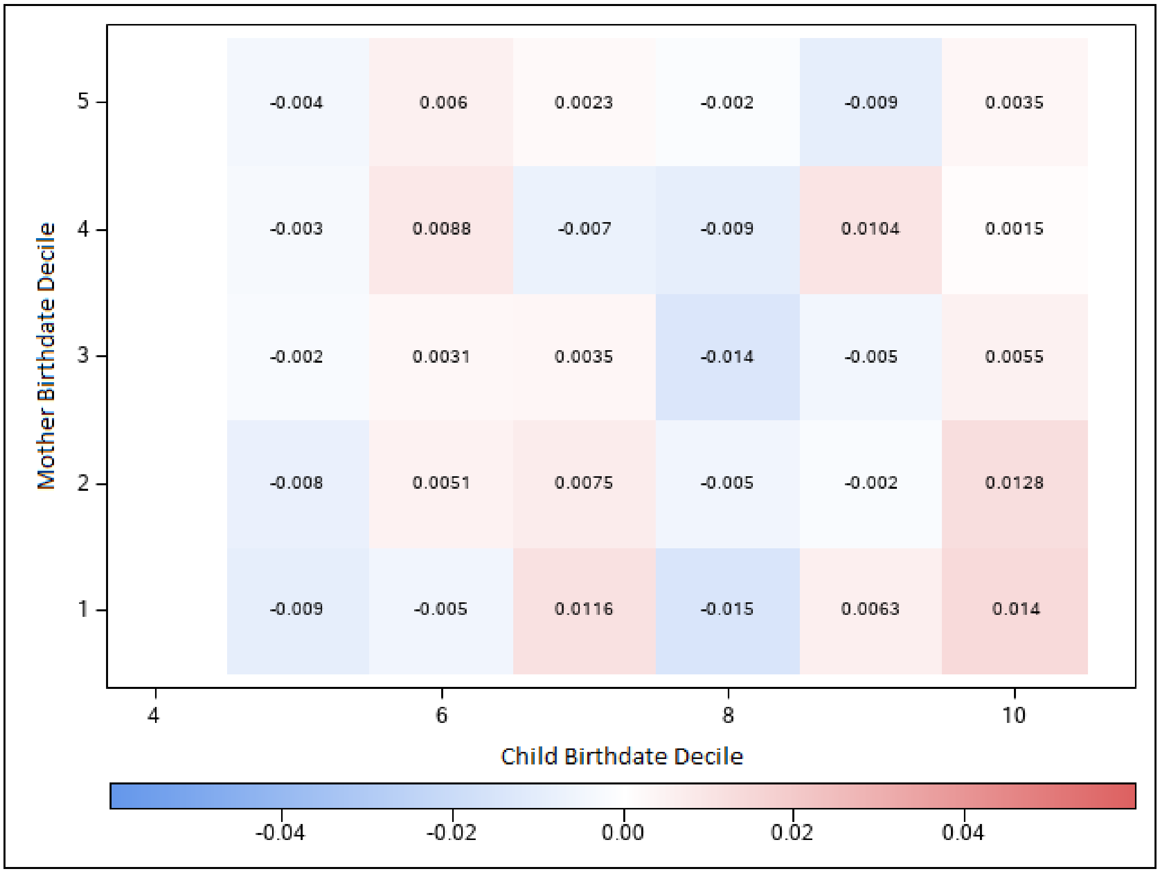
		(1)
Synthetic File	Comparison File	Parent-Child Links Score
Synthetic both #1	Original data	0.146
Synthetic both #2	Original data	0.131
Synthetic both #3	Original data	0.129
Synthetic both #4	Original data	0.132
Original data 50% Sample	Original data	0.067
Original data 50% Sample	Original data 50% Sample	0.092

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

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 15 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. The figure shows that younger siblings (i.e., those on

Figure 13. Difference between Original Data Density and Synthetic Data Density (Synthetic minus Original) – Mother-Child Birthdate



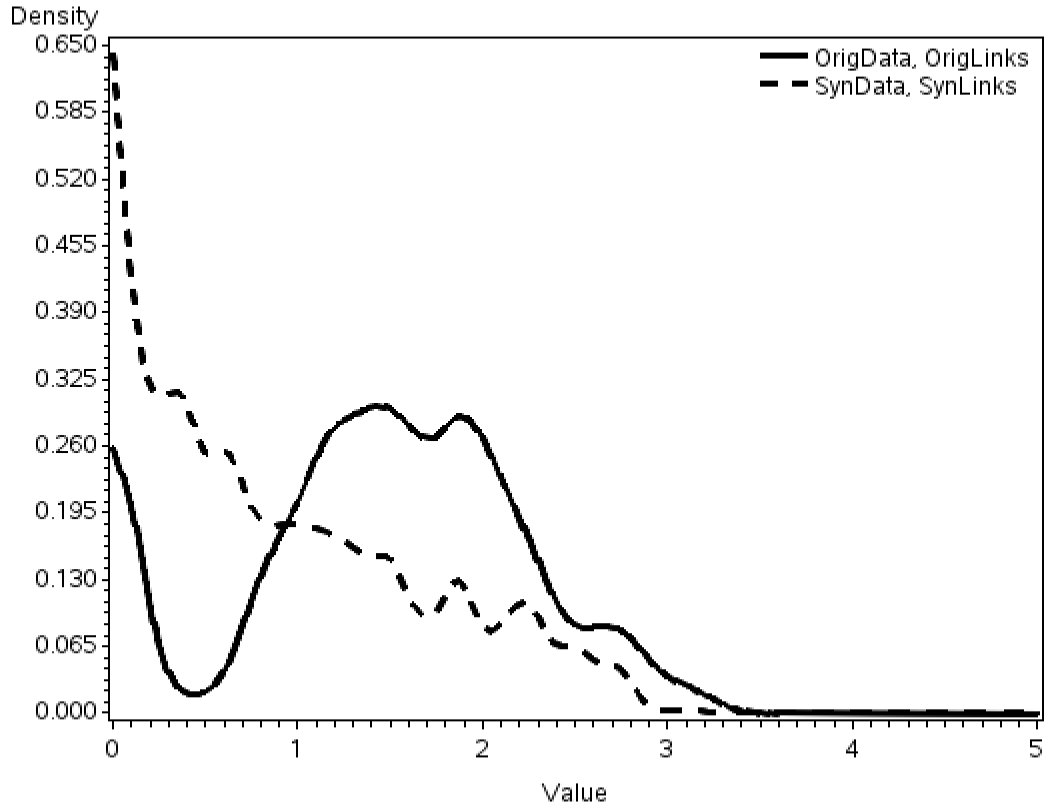
Source: U.S. Census Bureau Gold Standard File (linked SIPP-IRS-SSA).

the y-axis) born in later deciles are more likely to be matched with siblings also born in later deciles. That is, there are more siblings who are close together age, which is consistent with Figure 14.

Finally, Table 5 shows the overall two-way k-marginal scores for linked siblings. The synthetic sibling links have larger scores than the spousal and mother-child links, suggesting that sibling links do a worse job of preserving two-way correlations. However, the synthetic sibling scores still have about the same ratio to the sampling comparison scores as the prior links had, illustrating again the importance of considering the size of synthesis error relative to sampling error.

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 6 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).

Figure 14. Univariate Distribution of Older Sibling Age minus Younger Sibling Age



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

Table 5. Two-Way K-Marginal Scores for Sibling Linkages

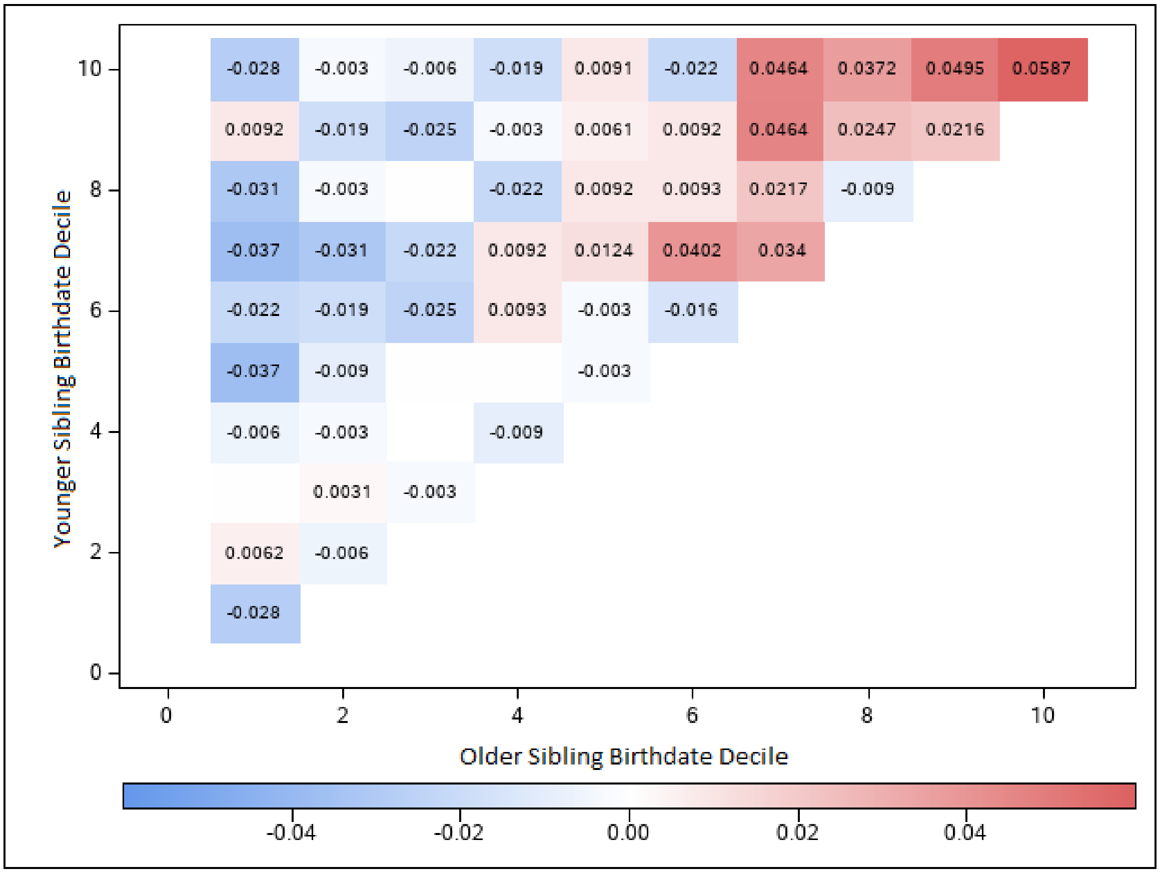
Synthetic File	Comparison File	(1) Sibling Links Score
Synthetic both #1	Original data	0.349
Synthetic both #2	Original data	0.323
Synthetic both #3	Original data	0.334
Synthetic both #4	Original data	0.310
Original data 50% Sample	Original data	0.152
Original data 50% Sample	Original data 50% Sample	0.205

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

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. Synthesis of these attributes performed quite well, with an average confidence interval overlap of 90.5%.

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

FIGURE 15. Univariate Distribution of Older Sibling Age minus Younger Sibling Age 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).
Note: Note: Empty cells above the diagonal have data, but the table only reports cells with an absolute density difference of at least 0.001. Empty cells below the diagonal have zero density because the younger sibling cannot have an earlier birthdate than the older sibling.

the adults in the household, and the work status of the adults in the household. Many of the statistics are quite similar based on the degree of interval overlap in their 95% confidence intervals [a standard often used in evaluating the quality of synthetic data: Karr et al. (2006)]. Educational attainment of the household stands out as the group of characteristics that our method struggles with in terms of confidence interval overlap. Our best guess for why this characteristic performs the worst is the same as our interpretation of the spousal educational attainment heatmap in Figure 11: spousal pairings likely occur in part through real-life interactions in social networks and these complex socio-economic features can be difficult to fully replicate. This includes interactions with peer groups in high school and college. Furthermore, categorical variables, such as educational attainment, always pose a strong challenge for synthesis and imputation models.

Table 6. Household Characteristics

	(1)	(2)	(3)	(4)	(5)
	Q-Original	95% CI-Original	Q-Synth	95% CI-Synth	Overlap
Household size					
=2: married couple	0.8483	(0.8425,0.8542)	0.8466	(0.8393,0.8538)	96.58
=3: married couple, 1 kid	0.0877	(0.0831,0.0923)	0.0857	(0.0758,0.0956)	100
$\geq$ 4: married couple, 2+ kids	0.0122	(0.0104,0.0140)	0.0113	(0.0089,0.0137)	91.67
=2: single mother, 1 kid	0.0457	(0.0423,0.0491)	0.0484	(0.0432,0.0536)	86.76
=3: single mother, 2 kids	0.0057	(0.0045,0.0070)	0.0073	(0.0053,0.0094)	68.00
$\geq$ 4: single mother, 3+ kids	0.0004	(0.0001,0.0008)	0.0007	(0.0001,0.0013)	100
Number of different races in household (white, black, or other)					
=1	0.9527	(0.9493,0.9562)	0.9469	(0.9387,0.9550)	82.61
=2	0.0468	(0.0433,0.0502)	0.0526	(0.0443,0.0608)	83.88
=3	0.0005	(0.0001,0.0008)	0.0006	(0.0001,0.0011)	100
Spousal/mother college education					
married couple, no degrees	0.5367	(0.5285,0.5448)	0.519	(0.5099,0.5281)	0
married couple, one degree	0.204	(0.1975,0.2106)	0.2413	(0.2335,0.2491)	0
married couple, two degrees	0.2075	(0.2009,0.2141)	0.1832	(0.1759,0.1904)	0
single mother, no degree	0.0436	(0.0403,0.0469)	0.0413	(0.0362,0.0465)	93.69
single mother, degree	0.0082	(0.0068,0.0097)	0.0151	(0.0121,0.0181)	0
Spousal/mother work status					
married couple, neither working	0.1732	(0.1671,0.1794)	0.1602	(0.1461,0.1743)	58.54
married couple, one working	0.2982	(0.2907,0.3056)	0.3203	(0.3074,0.3330)	0
married couple, both working	0.4768	(0.4686,0.4849)	0.4631	(0.4437,0.4826)	85.89
single mother, not working	0.0128	(0.0110,0.0146)	0.0147	(0.0123,0.0171)	64.48
single mother, working	0.0390	(0.0359,0.0422)	0.0418	(0.0362,0.0473)	94.30
Foreign born in household					
none	0.9631	(0.9601,0.9662)	0.9646	(0.9597,0.9695)	100
some, but not all	0.0210	(0.0186,0.0233)	0.0210	(0.0162,0.0257)	100
all	0.0159	(0.0139,0.0180)	0.0144	(0.0121,0.0168)	71.81
Average Interval Overlap					*67.19*

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

5.6. **Disclosure Risk Assessment.** Finally, we evaluate the level of disclosure limitation 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, we perform two tests to assess the disclosure risk of creating synthetic links. One intuitive test of the protection provided is to measure how often the actual links are recreated when synthesizing edges in the original data. In other words, if one does not synthesize any of the characteristics, but throws out the known edges in the original data, would this method introduce sufficient error into the graph? As we mentioned earlier, this method could be used to create partially synthetic data where the only synthesized feature is the graph. In such a case, the data provider would want to know to what extent that original graph is recreated. Table 7 shows the percent of links re-created for each of four different replicates of the synthetic links with original nodal attributes. Even though no other characteristic is perturbed, the link re-creation rate is small across all four replicates, suggesting that the privacy loss from these synthetic edges is low. For disclosure risk assessments, people often consider identity protection (protecting the presence of an individual in the data) as well as attribute protection (protecting values of characteristics even when an individual is known to be in the data). We focused on the stronger of the two, identity protection, because if identity is protected then certainly attributes are protected. It is worth noting, though, that this test of link re-creation does indicate strong attribute protection in the sense that even if we know the record of a certain married individual in the original data, the synthetic links will generally not link to the correct spouse, thus protecting the spouse's attributes.

Table 7. Percentage of Original Links Re-Created with Original Data and Synthetic Links

	(1) Implicate #1	(2) Implicate #2	(3) Implicate #3	(4) Implicate #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).

Next, we use a minimum distance, re-identification experiment to assess the marginal impact on disclosure risk from synthesizing the edges between records on a data set that already contains synthetic nodal attributes. The concept of this exercise is to imagine a powerful intruder armed with the original confidential data, and to see if the intruder can match (partially) synthetic records back to the correct original records with a reasonable minimum distance strategy. The re-identification process blocks on un-synthesized variables (for the file with synthetic nodal attributes and synthetic edges, there are no blocking variables, and for the file with synthetic nodal attributes but unsynthesized edges, the intruder can block on the type of family graph). Then, within those blocks, for each record in file A (the original file), we search for the record in file B (the synthetic file) that is closest based on some distance metric. If the closest record from file B corresponds to the same individual prior to the synthesis that it is matched to in file A, then it was a true match. Otherwise, it was a false match. We used four different distance metrics: Euclidean, Euclidean with each variable standardized by dividing by its standard error, Mahalanobis where the intruder only knows the covariance matrix from the synthetic data, and Mahalanobis where the intruder knows the covariance matrix in both the synthetic and original data (Abowd et al., 2006).

Because we are synthesizing every value of every characteristic in the files, one would expect that, when synthesizing the i-th record, it would very rarely become the closest match to the original i-th record. In fact, one would expect that to happen roughly $1/N$ times where N is the sample size. However, when we do not synthesize the graph, and the hypothetical intruder can block on the type of graph to which the i-th record belongs, then one would expect a successful re-identification rate of $1/N_j$ where j is the type of graph to which i belongs and N_j is the number of records in the whole sample that belong to graphs of type j. For very common graphs, this will still be small, so the overall re-identification rate would still be expected to be small. But for rare graphs, the re-identification rate should become quite large.

Because the successful reidentification rate is so low, for disclosure avoidance purposes we can only report that, in the file with synthetic edges, there were fewer than 15 successful reidentifications with every distance metric.[23] Moreover, with each distance metric, there

[23] As a reminder, our data used 54,000 individuals, including 13,500 linked spouse pairs, 2,200 linked moms, and 2,500 linked kids.

were approximately 20 times as many successful reidentifications in the file without synthetic edges even though every other characteristic was synthesized.

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. 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 a wife's share of spousal earnings at the 50% threshold and the pregnancy-specific age gap in siblings, that would be challenging for any model to re-create without explicitly controlling for them. However, the method appears to do a reasonable job of replicating many other characteristics of within-household links that we studied. This is noteworthy because while the method in Hu et al. (2018) provides a more theoretically coherent joint model for nodal attributes and edges, it is not easily applicable to all types of data. Our method is easily applied to many different variable types, or partially synthetic data, and still shows satisfactory results for synthetic attributes and edges. We also find that the method significantly reduces privacy loss risk in the form of re-identification risk when releasing link information.

Future work in the area of creating synthetic links between records in household survey data should focus on settings with many more edges and possible graph types, for which our method may or may not be feasible given its relative statistical simplicity. A method for creating synthetic links that satisfies formal privacy would also be desirable but is beyond the scope of this paper.

7.

Acknowledgment

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

References

J. M. Abowd, F. Kramarz, and D. N. Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 2003. https://doi.org/10.1111/1468-0262.00020.

J. M. Abowd, M. Stinson, and G. Benedetto. Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Working paper, U.S. Census Bureau, 2006. https://hdl.handle.net/1813/43929.

J. M. Abowd, G. Benedetto, S. L. Garfinkel, S. A. Dahl, A. N. Dajani, M. Graham, M. B. Hawes, V. Karwa, D. Kifer, H. Kim, P. Leclerc, A. Machanavajjhala, J. P. Reiter, R. Rodriguez, I. M. Schmutte, W. N. Sexton, S. Singer, and L. Vilhuber. The modernization of statistical disclosure limitation at the U.S. Census Bureau. Working paper, U.S. Census Bureau, 2020. https://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/ThemodernizationofstatisticaldisclosurelimitationattheU.S.CensusBureau.pdf.

A. M. Ali, H. Alvari, A. Hajibagheri, K. Lakkaraju, and G. Sukthankar. Synthetic generators for cloning social network data. Conference paper, ASE BigData/SocialInformatics/PASSAT/BioMedCom, 2014. http://eecs.ucf.edu/~halvari/10.pdf.

G. Benedetto, J. C. Stanley, and E. Totty. The creation and use of the SIPP Synthetic Beta v7.0. CES Technical Notes Series 18-03, Center for Economic Studies, U.S. Census Bureau, 2018. https://ideas.repec.org/p/cen/tnotes/18-03.html.

M. Bertrand, E. Kamenica, and J. Pan. Gender identity and relative income within households. *The Quarterly Journal of Economics*, 130(2):571–614, 2015. https://doi.org/10.1093/qje/qjv001.

D. Cesarni, E. Lindqvist, M. Notowidigdo, and R. Ostling. The effect of wealth on individual and household labor supply: Evidence from Swedish lotteries. *American Economic Review*, 107(12):3917–3946, 2017. https://doi.org/10.1257/aer.20151589.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632, 2014a. https://doi.org/10.1257/aer.104.9.2593.

R. Chetty, N. Hendren, P. Kline, and E. Saez. Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014b. https://doi.org/10.1093/qje/qju022.

R. Chetty, J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter. The opportunity atlas: Mapping the childhood roots of social mobility. Working paper no. 25147, National Bureau of Economic Research, 2020.

T. Cornelissen, C. Dustmann, and U. Schonberg. Peer effects in the workplace. *American Economic Review*, 107(2):425–456, 2017. https://doi.org/10.1257/aer.20141300.

J. Drechsler. *Synthetic datasets for statistical disclosure control: Theory and implementation.* Springer Science & Business Media, 2011.

J. Ferrie, C. Massey, and J. Rothbaum. Do grandparents matter? Multigenerational mobility in the United States, 1940–2015. *Journal of Labor Economics*, 39(3):597–637, 2021. https://doi.org/10.1086/711038.

S. Hawala. Producing partially synthetic data to avoid disclosure. Conference paper, Proceedings of the Joint Statistical Meetings, American Statistical Association, 2008. http://www.asasrms.org/Proceedings/y2008/Files/301018.pdf.

J. Hu, J. P. Reiter, and Q. Wang. Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis*, 13(1):183–200, 2018. https://doi.org/10.1214/16-BA1047.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):1–9, 2006. https://doi.org/10.1198/000313006X124640.

R. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426, 1993. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley, 2002.

M. Murray-Close and M. L. Heggeness. Manning up and womaning down: How husbands and wives report their earnings when she earns more. Working paper 28, Federal Reserve Bank of Minneapolis, Opportunity & Inclusive Growth Institute, 2019.

H. Pérez-Rosés and F. Sebé. Synthetic generation of social network data with endorsements. *Journal of Simulation*, 9(4):279–286, 2015. https://doi.org/10.1057/jos.2014.29.

G. M. Raab, B. Nowok, and C. Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97, 2017. https://doi.org/10.29012/jpc.v7i3.407.

R. Raghunathan. Synthetic data. *Annual Review of Statistics and Its Applications*, 8: 129–140, 2021. https://doi.org/10.1146/annurev-statistics-040720-031848.

T. E. Raghunathan, J. P. Reiter, and D. B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–17, 2003. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/multiple-imputation-for-statistical-disclosure-limitation.pdf.

J. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544, 2002. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/satisfying-disclosure-restrictions-with-synthetic-data-sets.pdf.

J. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20030026785.

J. Reiter. Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(2):235–242, 2004. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20040027755.

J. Reiter. Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3):441–462, 2005. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf.

J. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society*, 31(2):115–124, 1982. https://doi.org/10.2307/2347973.

J. Royston. Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society*, 32(2):121–133, 1983. https://doi.org/10.2307/2347291.

D. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):462–468, 1993. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/

discussion-statistical-disclosure-limitation2.pdf.

D. Sacerdote. Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. A. Hanushek, S. Machin, and L. Woessmann, editors, *Handbook of the Economics of Education*, chapter 4, pages 249–277. Elsevier, 2011.

S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965. https://doi.org/10.1093/biomet/52.3-4.591.

E. Totty. High school value-added and college outcomes. *Education Economics*, 28(1):67–95, 2020. https://doi.org/10.1080/09645292.2019.1676880.

U.S. Census Bureau. SIPP Gold Standard Files v7.0. Data set, D.C., Washington: U.S. Department of Commerce, 2018.

S. Woodcock and G. Benedetto. Distribution-preserving statistical disclosure limitation. *Computational Statistics & Data Analysis*, 53(12):4228–4242, 2009. https://doi.org/10.1016/j.csda.2009.05.020.