Final report

Octopize-MESR-LNDS

2024-12-10

Table of contents

List of Figures

List of Tables

1 Overview

Data linkage aims to associate individuals from a data source A are to individuals from a data source B in a way that global statistics on $A \cup B$ can be discovered and exploited.

Although the library can be used on any data, the original need for data linkage functionalities comes from a context where data at each source cannot be shared for legal or competition reasons. Data linkage should therefore be compatible with anonymized data.

1.1 Example

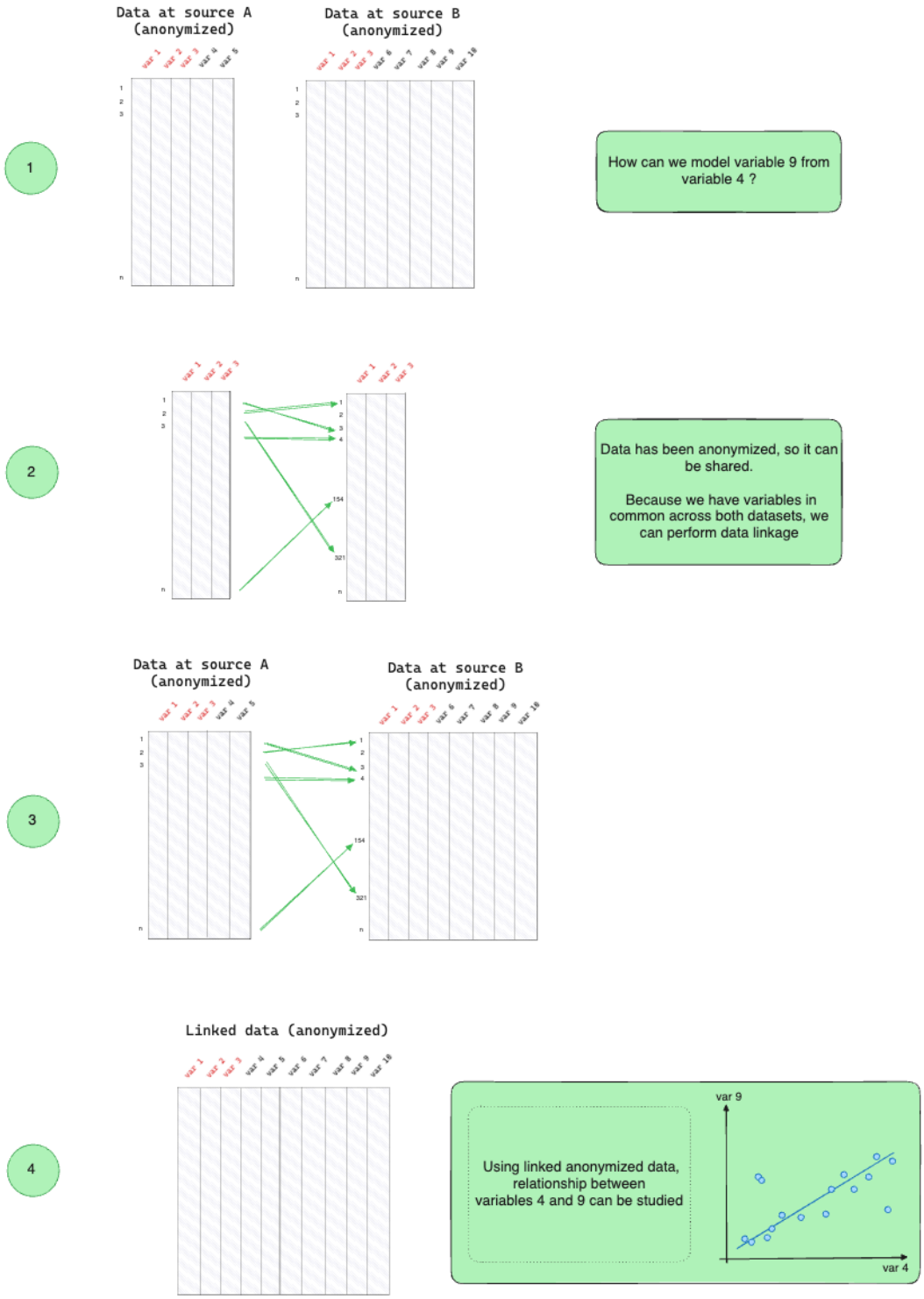

Figure 1.1: Linkage example

1.2 Scope

The library currently focuses on linkage in the following contexts:

- **There must be some variables in common**. The solutions made available use a notion of distance between individuals in both sources. This distance is computed using those common variables.
- **Both sources contain data on the same individuals**. Linkage algorithms can be adapted to handle linkage of different populations but this is not currently done and evaluation on such contexts has not been carried out to date.
- Evaluation of the proposed linkage solutions has been carried out on contexts with **only 2 data sources**. While handling more data sources by sequentially applying linkage is possible, there is no evidence on the quality of the resulting linked data.
- Data to be linked is **contained in a single file at each source** where one row represents one individual to link. Relational databases are not handled.

1.3 Linkage steps

Data linkage should follow some key steps. First, in most contexts data will need to be anonymized, so that it can be shared before being linked. It is then necessary to evaluate the potential for linkage. If the variables common to both datasets are too few or not representative enough of the datasets, then success of linkage cannot be guaranteed and it is recommended to look for additional or alternative common variables before proceeding with linkage. Pre-linkage metrics are available to measure the chances for a linkage to be successful. Following computation of pre-linkage metrics, linkage can be performed, resulting in a single linked data file. When possible (i.e. in experimental or development contexts), post-linkage metrics can be computed to compare a reference dataset to the linked data.

Those steps and insights about linkage performance are detailed in dedicated pages: - Anonymization - Pre-linkage metrics - Linkage - Post-linkage metrics - Experiments

Future work ideas are listed in: - Future work

2 Anonymization and privacy

This page describes how data can be anonymized with the avatar solution and what linkage means in terms of privacy.

2.1 Anonymization with the avatar solution

Data linkage can be done in the context of original data but the project and the library emerged from a need in a context of anonymized data where original data cannot be shared for legal or competition reasons.

In this latter context, the typical pipeline would go as follows:

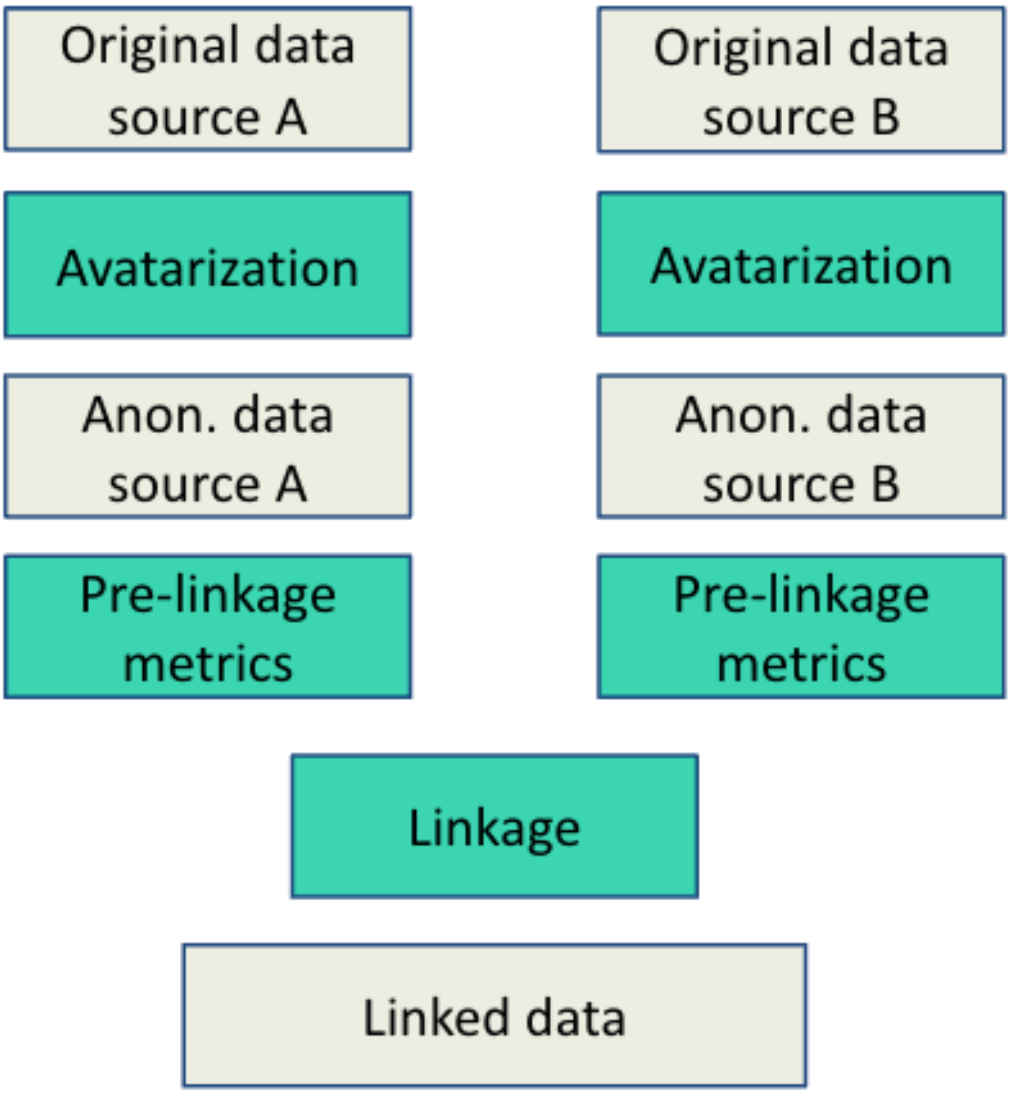

Figure 2.1: Base pipeline

Avatar is an anonymization solution developed by Octopize that produces data respecting GDPR criteria: singling-out, linkability and inference.

The solution is available on SaaS and on-Premise. Scripts demonstrating the full pipeline make use of the SaaS version.

For more information, we recommend going through the avatar public documentation and reading the paper describing the avatar method:

Guillaudeux, M., Rousseau, O., Petot, J. et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. Nature Digital Medicine. 6, 37 (2023)

2.2 Privacy

The avatar solution comes with privacy metrics and a privacy report can be automatically generated. A dataset can only be considered anonymous if this is confirmed by privacy metrics. The metrics covers the 3 GDPR criteria as described in details here.

Privacy metrics confirm that it is impossible to re-identify an individual in the data (singling-out), to link individual from the dataset with data from another source and to infer sensitive information from the data.

2.2.1 Linkage and GDPR linkability criteria

While it may seem impossible to link data for which the linkability criteria is respected, this is not the case as the GDPR linkability criteria prevents the linkage of data from one individual to be linked with his data coming from a different source. **The linkage as performed in this library is not impacted by the respect of this criteria because linkage does not aim at recreating exactly the original data but instead aims at associating individuals in a way to preserve global statistics of a dataset (for example correlations).**

2.2.2 Linkage and increased privacy

Through the experiments carried out, it has been observed that data linkage does not add any privacy risk. The experiments where original data was split then linked showed that applying avatar at each source yielded anonymized data at each source and that their linkage resulted in data with some statistical properties shared with the original data and **an increased level of protection**. This can be explained by the fact that the linkage is approximate and can be interpreted as an additional source of noise.

3 Pre-linkage metrics

Pre-linkage metrics are necessary to evaluate the potential for linkage. If variables common to both datasets are too few or not representative enough of the datasets, then success of linkage cannot be guaranteed and it is recommended to look for additional or alternative common variables before proceeding with linkage. Pre-linkage metrics are available to measure the chances for a linkage to be successful.

3.1 Unicity

Given a dataset A with variables V_A, the unicity score for a set of shared variables V_S is given by the ratio between number of unique value combinations of V_S (its cardinality *card(V_S)*) and the number of records in the dataset:

$$U(V_S, A) = \frac{card(V_S)}{|A|}$$

The unicity score represents how unique records are when only defined by the common variables V_S. The best possible linkage variable is a direct identifier as one value represents exactly one individual. Its unicity score is 1. On the other hand, a variable such as *gender* is an example of a variable with poor unicity score. If V_S only contains such a variable, then it should not be expected that linkage will be successful. Variables with low cardinality should still be considered for inclusion in V_S because it can give high unicity score if combined with other variables with low cardinality.

Note that the unicity score needs to be computed for both data sources.

3.2 Contribution score

Common variables V_S can also lead to good linkage if those are representative of the dataset as a whole. To model this representativeness, a model can be learnt at a data and contribution of the variables in V_S be computed.

Although many models can be considered, we use factor analysis to project the data. This type of model is also used when computing some distances in the linkage phase. Such model represents the data using a set of dimensions, each explaining for a certain proportion of the variance in the data. For example, a PCA can represent data with 10 dimensions where the first 3 dimensions will explain respectively 55%, 25% and 10% of the variance.

For each dimension, it is possible to retrieve the individual contribution of each variable. For example, a specific variable (*age*) may contribute to 80% of the first dimension of a PCA.

The pre-linkage contribution score makes use of the concepts of proportion of variance explained and contribution. For a model M of size (number of dimensions $|M|$), the contribution of a variable i from V_S to the dimension j is $\alpha_{i,j}$. For the same model, the variance explained by each dimension j is σ_j^2. The contribution score is defined as:

$$C(V_S, A) = \frac{\sum_{i \in V_S} \sum_{j \in |M|} \alpha_{i,j} * \sigma_i^2}{\sum_{i \in V_A} \sum_{j \in |M|} \alpha_{i,j} * \sigma_i^2}$$

$C(V_S, A)$ is defined between 0 and 1. If $V_S = V_A$, then $C(V_S, A) = 1$.

Note that the contribution score needs to be computed for both data sources.

3.3 Relationship between pre-linkage metrics

Both metrics can be considered prior to performing data linkage. They measure different concepts and are not correlated with each other.

4 Linkage methods

This page describes the different approaches available to perform linkage between two data sources.

4.1 Concept

Linkage of datasets requires two main components: a notion of distance between two records and an algorithm to associate records from A to records from B.

4.2 Distances

4.2.1 Gower distance

The gower distance can be used on data containing both numeric and non-numeric variables. This distance is a natural option to consider when computing distances between records in datasets as most real-life datasets contain mixed data types. Gower can be interpreted as a combination of Euclidean and Hamming distance.

For a dataset with a set of categorical variable C and a set of numerical variables N, the Gower distance between two records (x) and (y) is given by:

$$D_{gower}(x, y) = \sum_{i \in C} (\{x_i\} - \{y_i\}) + \sum_{i \in N} 1 - \frac{|x_i - y_i|}{R_i}$$

$R_i, i \in N$ refers to the value range of variable i in the whole dataset

This library uses the *gower* library available on pypi.

4.2.2 Euclidean distance

Although, Euclidean distance cannot be used directly on non-numeric data, a dataset can be projected into a multidimensional numeric space in which all records have numeric only coordinates. Factor Analysis can be used for this purpose and this is the solution used in this library by means of Principal Component Analysis (PCA), Factor Analysis of Mixed Data (FAMD) and Multiple Correspondence Analysis (MCA).

Following a projection P (representing records in |P| dimensions), The Euclidean distance between two records (x) and (y) is calculated between their coordinates as:

$$D_{eucl}(x, y) = \sqrt{\sum_{i \in |P|} (P_i(x) - P_i(y))^2}$$

$P_i(x)$ refers to the projection of x on the i-th dimension (i.e. its i-th coordinate)

Before projection, a model needs to be fitted. This can be done either on data from source A or from source B or from both sources (this only really has an influence if anonymization has been performed). The 3 options are available but no significance difference has been observed. For most experiments, a model fitted on both data is used.

This library uses *saiph*, available on pypi for projection using factor analysis.

4.3 Linkage algorithms

4.3.1 Linear Sum Assignment (LSA) linkage

LSA is a classic combinatorial optimization problem which aims at assigning a set of objects to another set of objects in a way that the overall assignment cost is minimized.

A solution to a LSA problem is required to be bijective, i.e. any object in A must be assigned to exactly one object in B.

In the context of data linkage between two data sources A and B, the objective is to find a mapping M such that the sum of distances over all associated pairs of individuals x and y is minimized:

$$min \sum_{x,y \in |A|} D(x, y) * M_{x,y}$$

LSA is not bound to any distance and so all distances described in the earlier section can be used.

There is active research on solving LSA problems. In this library, we rely on the solver provided by scipy.

For further details about LSA solvers (including the one used by scipy), we suggest the following paper:

Dell'Amico, M. and Toth, P. *Algorithms and codes for dense assignment problems: the state of the art.* 2000. Discrete Applied Mathematics.

It is important to note that the complexity of the best LSA solvers is $\mathcal{O}(n^3)$. This may impact linkage of datasets containing many records.

Although not covered in this library, this could be handled by dividing the data in smaller subsets on which LSA can be run in reasonable time. This would however results in an approximation of what a solver would obtain on the whole dataset.

4.3.2 Greedy linkage

The library also includes some simple greedy algorithms. Those algorithms are simple and easily understandable. However, their performance is limited in comparison with LSA and their use is not recommended on real use cases.

`Min order`

The first greedy algorithms, currently named `min_order`, orders all records in A by $min(D(x))$, its closest distance to another record in B. Ordered records are then allocated in this order to the closest non-assigned records in B.

`Min re-order`

The second greedy algorithms, currently named `min_reorder` only differs from `min_order` by the fact that records in A are re-ordered by $min(D(x))$ *after each allocation*. This algorithm provides better results that `min_reorder` but requires significantly more resources.

4.4 Baseline algorithms

To understand the performance of the different linkage options, two baseline linkage methods are available.

4.4.1 Row order linkage

Row order linkage between two sources A and B will allocate the i-th individual from A to the i-th individual from B.

$$M = \begin{bmatrix} 1 & & & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1 \end{bmatrix}$$

In experimental contexts where the real order is known and the data not anonymized, *row_order* linkage yield perfect matching. On similar contexts where the data has been anonymized (but not shuffled), *row_order* will not produce the best linkage but a good linkage. We can consider *row_order* as a good objective. Linkage algorithms approaching the quality of *row_order* can be considered as good.

4.4.2 Random linkage

To compare linkage solutions, a random linkage approach is made available. Its results will represent a lower bound.

5 Post-linkage metrics

Because post-linkage metrics rely on accessing the original un-split data, they are only suitable to specific contexts: - experimental context where the data is purposely split under different scenarios to study the performance of linkage solutions or to develop new ones. This is the context in which results presented in *Experiments* where generated. - Development/configuration phase for a real-life use case. Prior to deploying a functionality using linkage in production, it may be possible to assess its future performance on samples or on synthetic data. In this context, data with primary key (or identifiers) can be shared and the un-split data (i.e. ground truth) reconstituted and compared with the linked data.

We present here two generic metrics.

5.1 Correlation retention

The main objective of data linkage is to preserve or reconstitute relationship between variables that are in different datasets or sources. Naturally, pairwise correlations can be calculated on the linked data and compared to those of original data. For this purpose, we use Pearson correlations coefficients whose absolute value represent how strong a correlation is between two variables.

Correlation retention is the absolute difference between correlations of original and linked data for each pair of variables in A and B respectively. Several statistics can then be used to measure success of linkage: *mean correlation retention*, *max correlation retention*, *sum of correlation retention*.

Correlation retention can also be plotted. Below are two examples of good (using a *row order* linkage) and poor (using a *random* linkage) correlation retention.

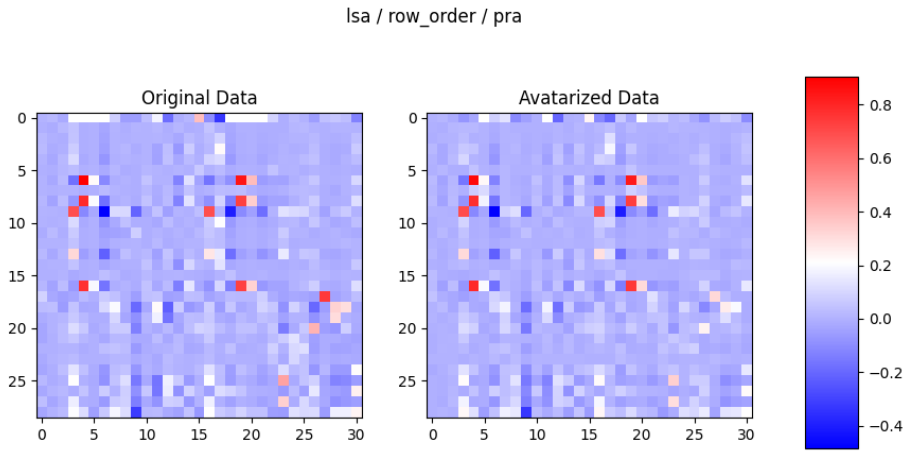

Figure 5.1: Example of good correlation retention

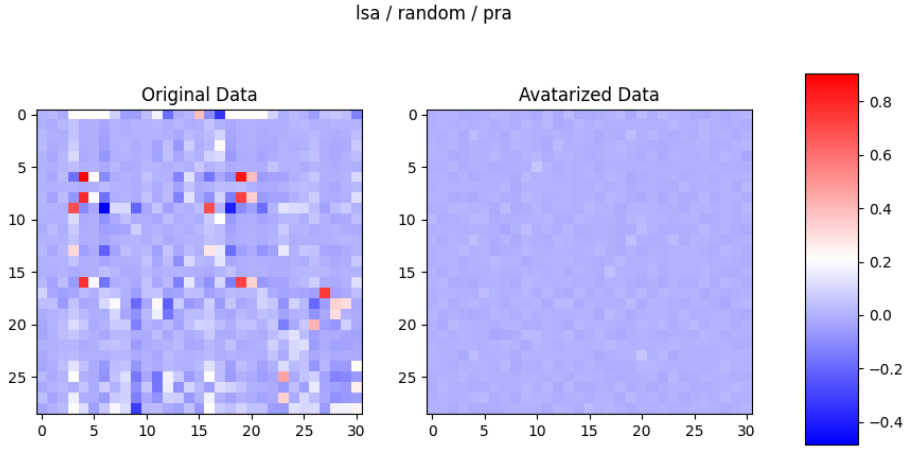

Figure 5.2: Example of poor correlation retention

5.2 Reconstruction score

Correlation retention focuses on pair-wise correlation and so is a bi-variate metric. However, a dataset can have more complex relationships between more than 2 variables. The reconstruction score relies on a factorial analysis model (e.g. PCA) in order to capture all relationships that may exist in the data and this without limitations in term of number of variables in those relationships.

The reconstruction score is the difference between the reconstruction error of the original data and the reconstruction error of the linked data. Reconstruction error is computed by first fitting a model on the original data and by keeping the first

dimensions (the ones explaining most variance). This model represents the original data. By projecting any data on this model and reconstructing the data (i.e. inverse transform), we can compute the reconstruction error.

If the data has similar statistical properties to the original data, the reconstruction error will be close to the reconstruction error of the original data. On the other hand, if the data has different statistical properties, the reconstruction error will be higher. Measuring this difference is a way to capture how globally similar the linked data is to the original data. This metric is multivariate since each model dimension potentially combines information from all variables.

Because reconstructed data needs to be in the original space, it may be a mix of numeric and categorical types, so the reconstruction error is measured by means of the Gower distance.

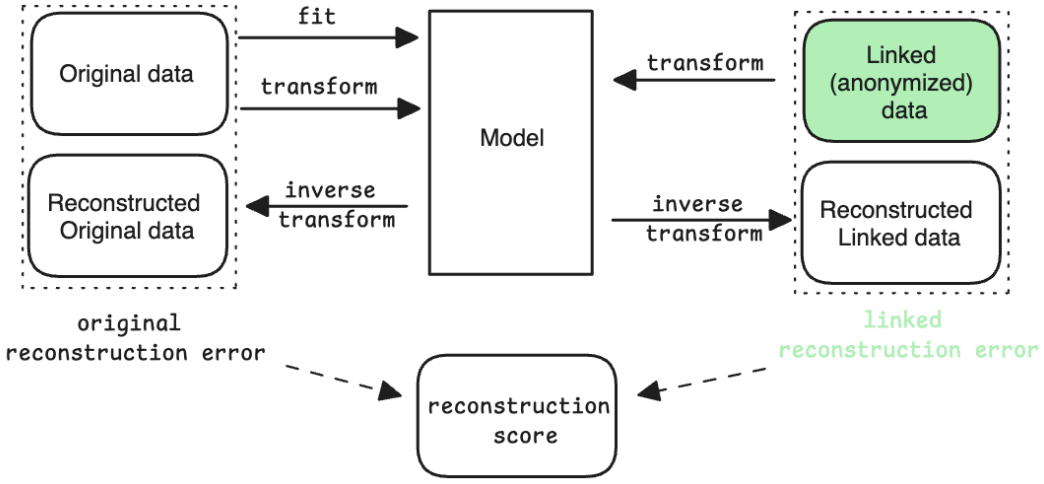

Figure 5.3: Reconstruction score

6 Experimental results

6.1 Differences between the selected datasets

Dataset sizes

The number of rows in a dataset is a property which may be important when interpreting results because it probably has an impact on linkage. This is especially true because the proposed linkage solutions are approximative: the aim of the linkage is not to link individual 1 from source A with individual 1 from source B but instead to link individual 1 from source A with another individual from source B in such way that the global statistical properties of the datasets are preserved.

We can expect that - linkage on a few individuals is less tolerant to approximation: for linkage of a given individual at source A, there are less candidate at source B than on datasets with a lot of individuals. - Also, a proper anonymization method (such as avatar) will modify to a greater extent individuals from a small dataset than from a large dataset (i.e. many individuals). - **It is expected that linkage of larger datasets (in terms of number of individuals) should preserve better the global statistical properties.**

Dataset	Number of individuals
student performance	395
student dropout	4424
adult	10000 (out of 48842)
pra	12430

Unicity scores

Based on the variables present in each datasets and in particular their cardinality, the unicity scores differ across datasets.

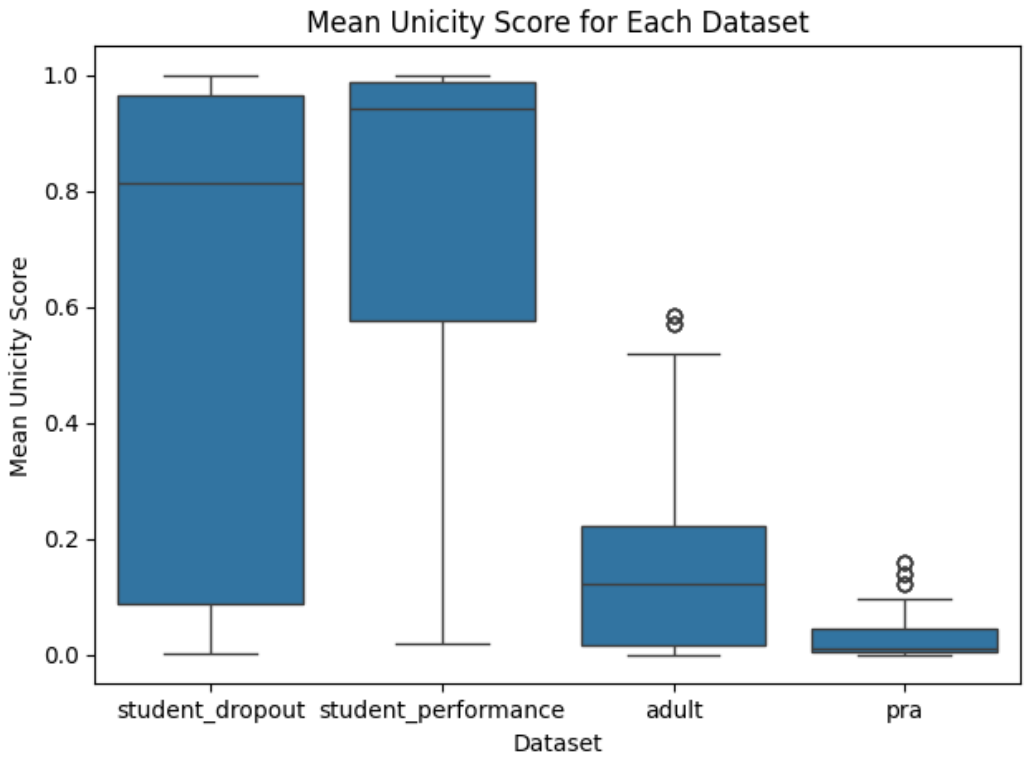

Figure 6.1: image info

Interpretation: - Unicity score can be very hiogh on student_dropout and student_performance datasets. This is caused by the smaller size of those datasets in comparison with the other datasets. Unicity score should be considered alongside number of records - Because they are the datasets with the fewest individuals, the anonymization should also yield more differences to the original data (this should be seen on assessment of linkage of avatars).

6.2 Results with k=10

Correlation differences

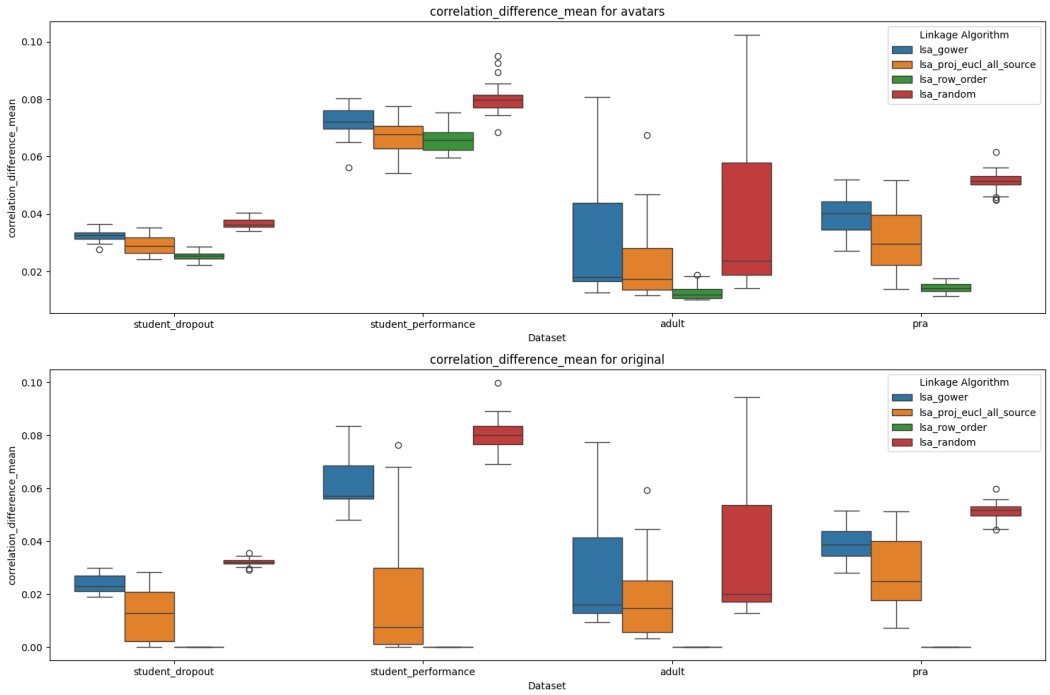

Figure 6.2: image info

Interpretation: - Across all datasets, *lsa + euclidean distance in a projected space* is the method giving the best results (i.e. correlation difference is the lowest)

Correlation differences (for high unicity scores only)

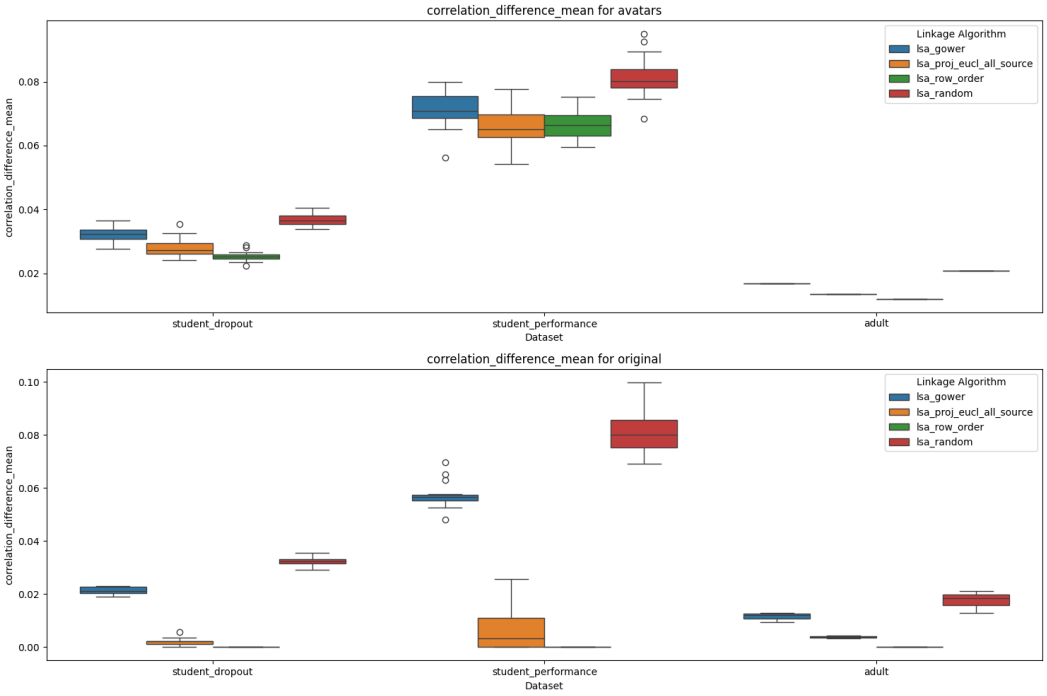

Figure 6.3: image info

Interpretation: - When the unicity score is high (e.g. > 0.5), the linkage with the best method is of similar or comparable quality to *row_order*. It can even outperform it

Mean Correlation differences for differerent level of unicity score

An acceptable level of correlation difference between original and sythetic linked data is difficult to define as it depends on the context and how the data will be used.

As a generic threshold, it is often considered that a difference lower than 0.1 is acceptable. For illustration purpose, we use 0.1 as a threshold when showing maximum correlation differences and we use 0.05 for mean correlation differences.

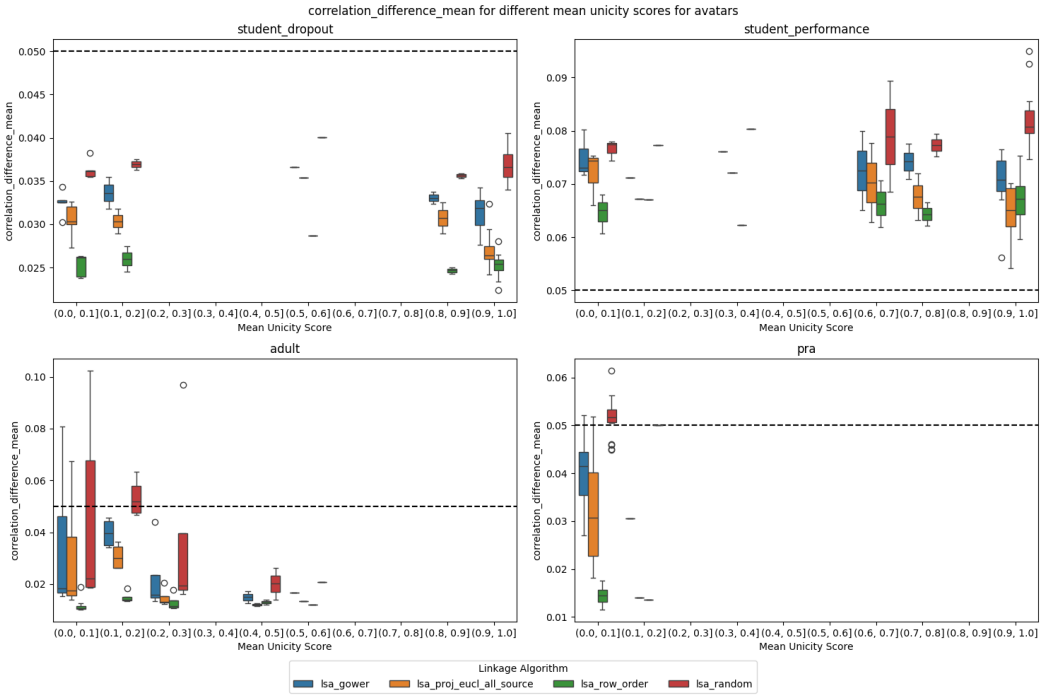

Figure 6.4: image info

Interpretation: - Whether results are above or beyond the indicative acceptable threshold strongly depends on the dataset. - Linked avatars yields acceptable mean correlation difference

Max Correlation differences for differerent level of unicity score

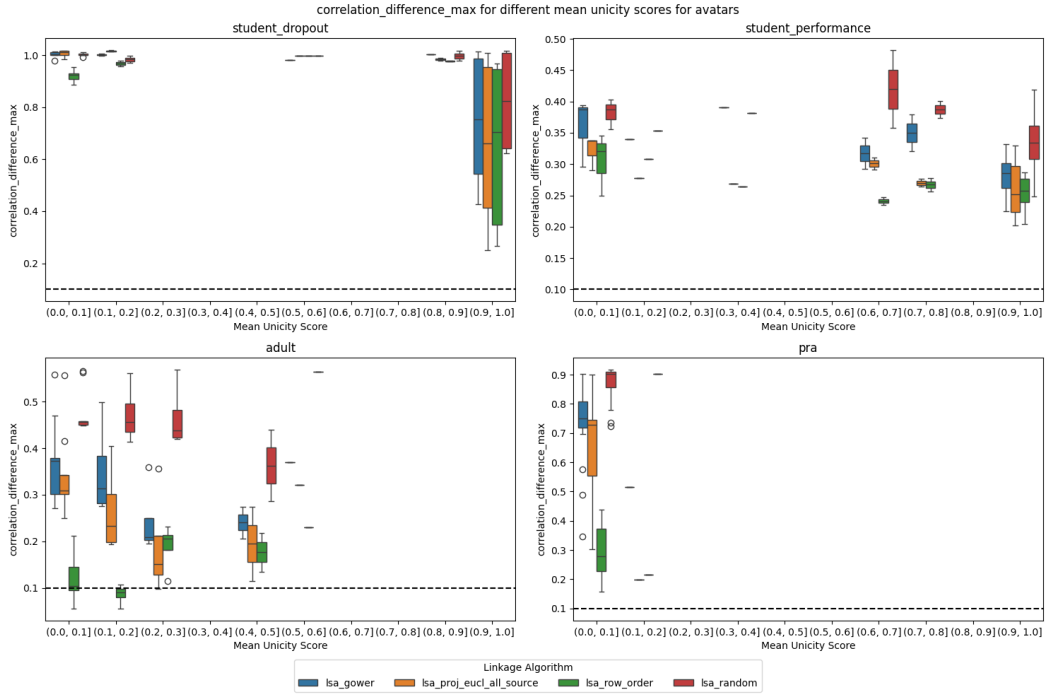

Figure 6.5: image info

Interpretation: - Maximum difference in correlation are not below the threshold and it is expected that some pairwise correlations will be altered. - However, looking at one run of the best method on pra, we see that only a few correlations are displaying large differences. We also observe that this correlation difference is high with the reference *row_order* linkage. - We observe that a global correlation may not be kept at linkage but we also see that no non-existent correlation is created. This is important wrt. the contexts in which such linked data can be used. - The likely reason behind this is that with some data splits, a variable globally correlated with another becomes completely uncorrelated and independent in its own split. Anonymization of the split alters this variable independently and global correlation is lost. There is no way such correlation can be restored at linkage. Note that this does not happen when a variable is globally correlated with several other variables because such variables would not become independent when the data is split. - Changing anonymization parameters does not help (not shown here but similar results in terms of maximum correlation differences are obtained when using k=3 instead of k=10 in avatar)

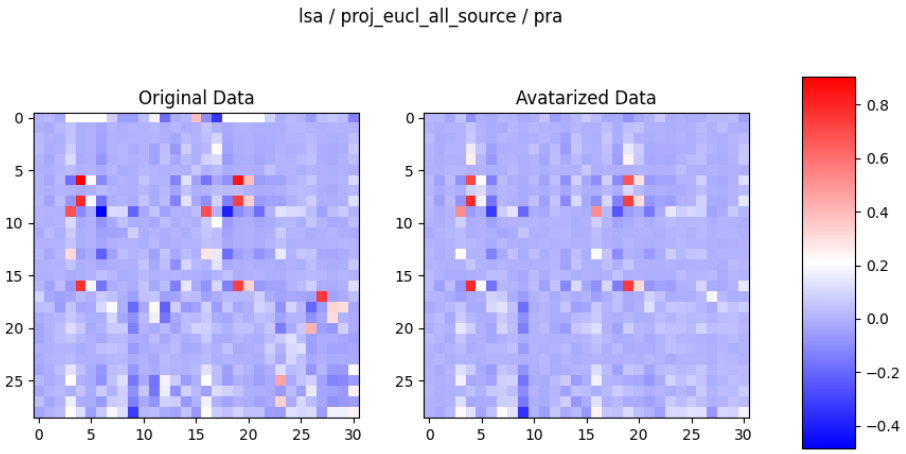

Figure 6.6: image info

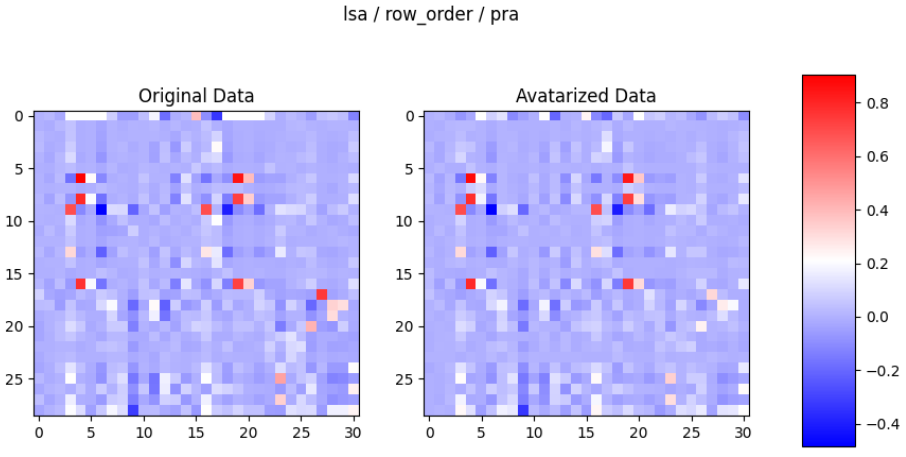

Figure 6.7: image info

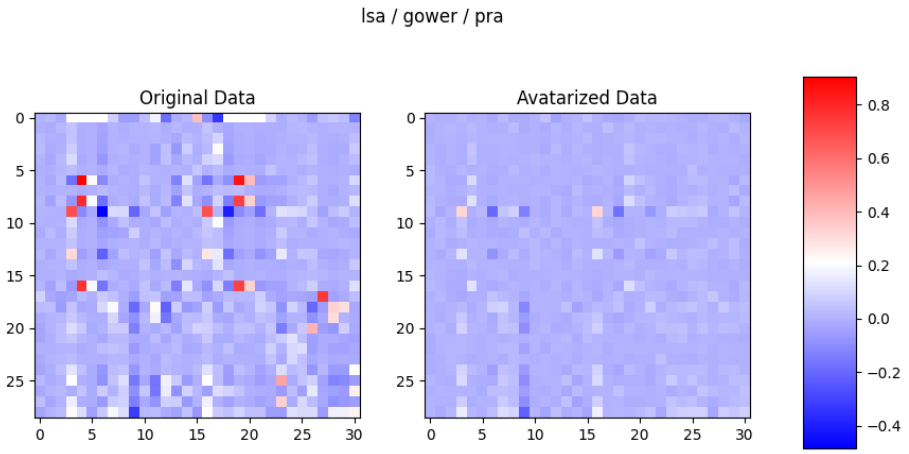

Figure 6.8: image info

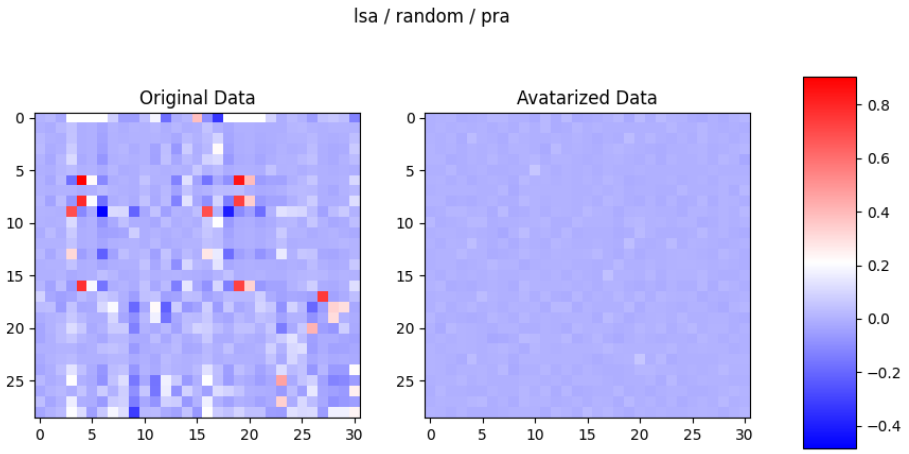

Figure 6.9: image info

6.3 Can we predict post-linkage results from pre-linkage metrics ?

Correlation between Unicity score (pre-metric) and correlation difference (post-metric)

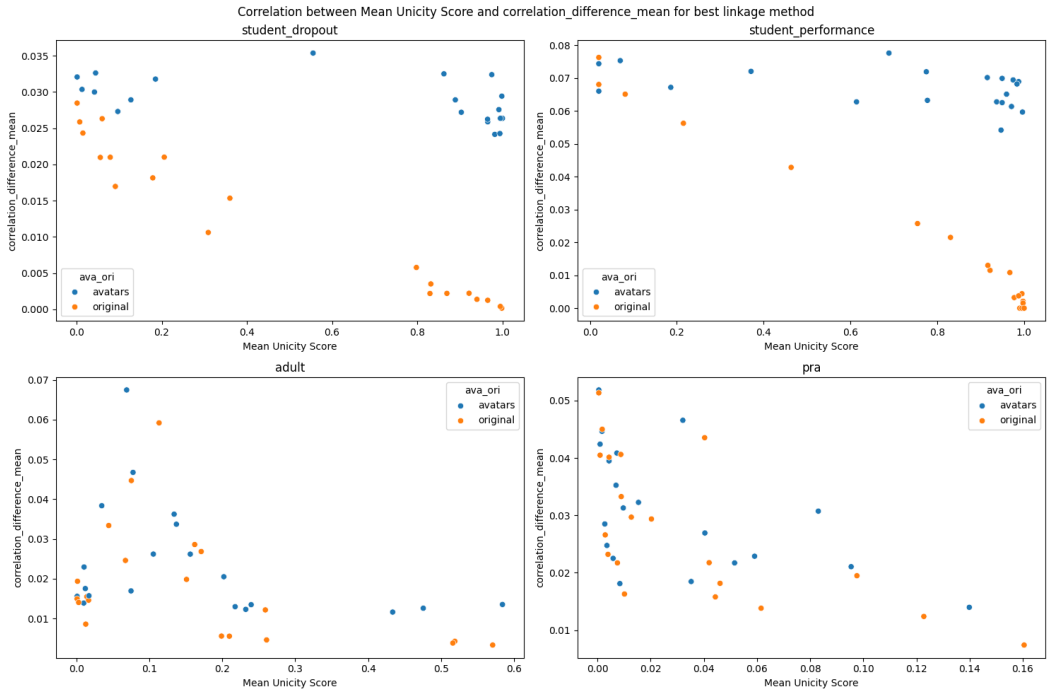

Figure 6.10: image info

Interpretation: - The trend (correlation) between pre and post linkage metrics is clear when linking subsets of the original data - The correlation can be observed on avatar linkage on adult and pra.

Correlation between Contribution score (pre-metric) and correlation difference (post-metric)

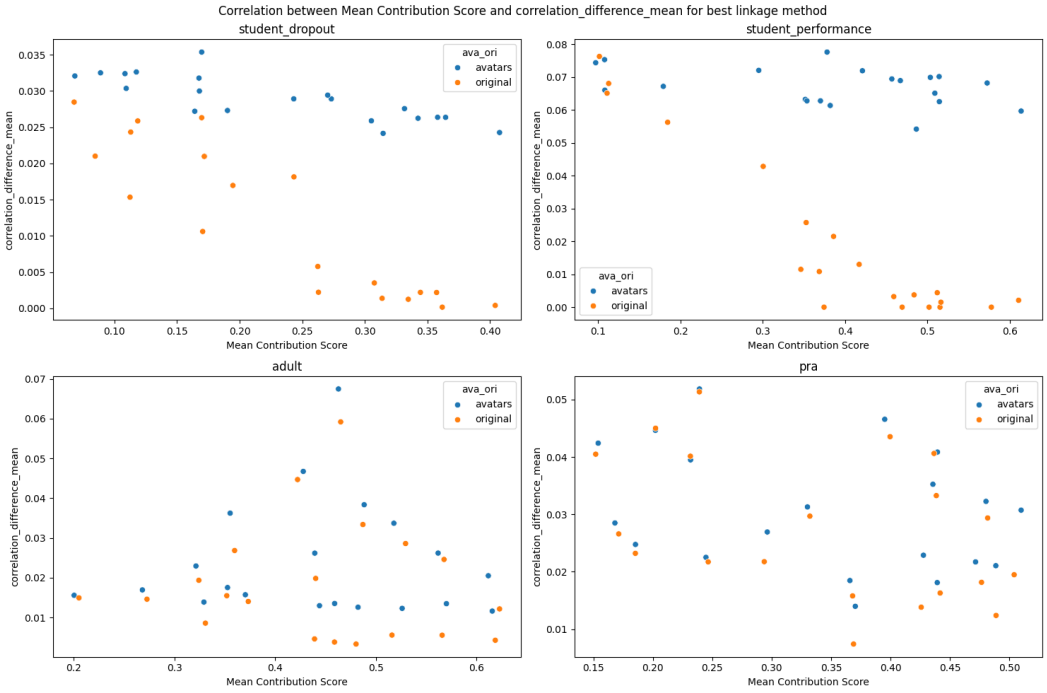

Figure 6.11: image info

Interpretation: - No systematic clear correlation with post-linkage metric when using this contribution score. Focusing on pra and adult where a correlation is observed on unicity score, we see that this is not the case here.

Impact of dataset size on linkage of avatars

To study the impact of the number of records, we use the two largest datasets and sample them (with seeds) to obtain datasets of respective size 10000, 5000, 1000 and 500 on which avatarization and linkage by means of *lsa + euclidean distance in a projected space*.

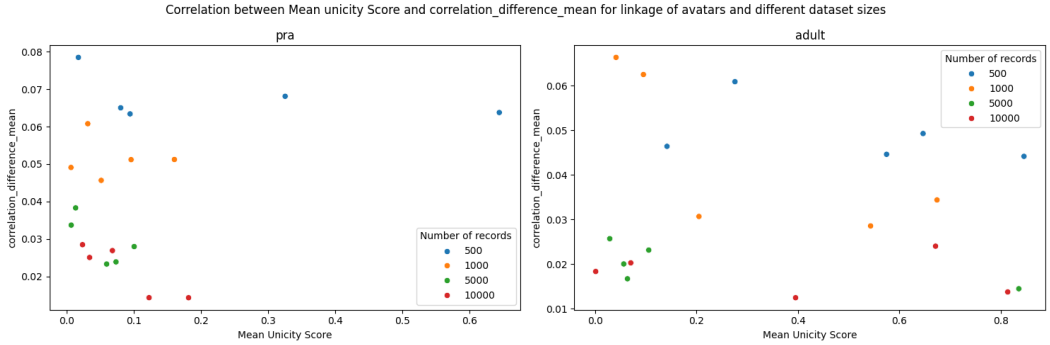

Figure 6.12: image info

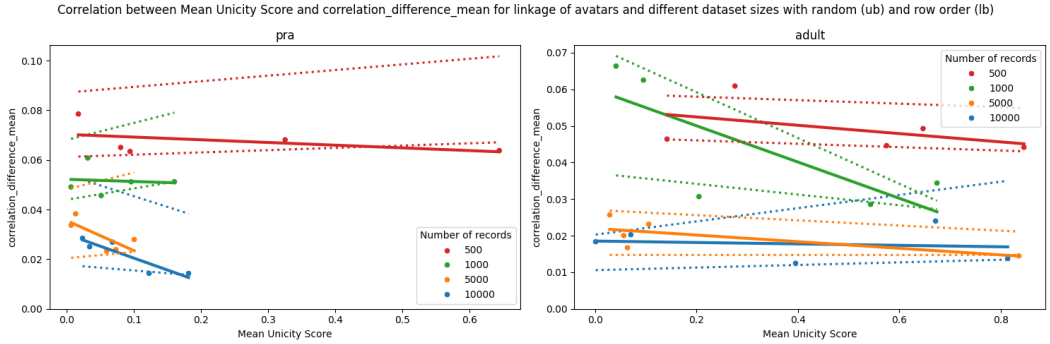

Figure 6.13: image info

Interpretation: - The more records, the closer to original. This is an expected findings: more individuals in a dataset means that there are more potential good candidates for association during the linkage step.

6.4 Take-home messages

- Unicity score should be combined with number of records to decide whether linkage could be performed
- Mean correlation difference is a post-linkage that can be used.
- When having sufficient number of records (i.e. > 5000), unicity score is correlated to post-linkage metric and so should be maximised.
- Some global correlations may not be "preserved" but non-existent correlations will not be created at linkage

  - :point_right: **Correlation in linked data can be considered as real correlations.**
  - :point_right: **But a lack of correlation in linked data may not necessarily mean that there is no correlation.**

- Based on 2 datasets large enough in this study, we suggest that unicity score should be greater than ~0.2 before attempting linkage. Additional runs on more datasets should be executed to determine a threshold.
- This library should serve as a basis to assess additional options to measure and perform linkage of synthetic data. Additional distances, algorithms and metrics can easily be added to it and evaluated.

7 Future work

7.1 Pre-linkage metrics

- Measure contribution of individual variables to dataset (i.e. contribution to projection). Potential correlations involving variables that are almost random (low contributions) are at risk to be lost after anonymization (and linkage cannot fix this). Identifying those variables would be good to know what use case / explorations can be considered on the linked anonymized data.
- Include dataset size (number of individuals) to existing metrics (because it is an important factor of linkage success).

7.2 Linkage methods

- The current solutions do not bias linkage towards specific variables. This could be done to ensure some correlations are kept at linkage.
  - Use LDA as a projection method on which euclidean distances are computed
  - Use weights on specific variables in saiph projections
- Work on greedy linkage algorithms

7.3 Post-linkage metrics

7.4 Experiments

- Run pipeline on many more datasets