https://doi.org/10.1007/978-3-031-17146-8_2

Utility-Preserving Biometric Information Anonymization

Bill Moriarty[1], Chun-Fu (Richard) Chen[1], Shaohan Hu[1], Sean Moran[1], Marco Pistoia[1], Vincenzo Piuri[2], and Pierangela Samarati[2]

[1] JPMorgan Chase Bank, N.A., New York, NY, USA
{william.r.moriarty, richard.cf.chen, shaohan.hu, sean.j.moran, marco.pistoia}@jpmorgan.com
[2] Università degli Studi di Milano, Italy
{vincenzo.piuri, pierangela.samarati}@unimi.it

Abstract. The use of biometrics such as fingerprints, voices, and images are becoming increasingly more ubiquitous through people's daily lives, in applications ranging from authentication, identification, to much more sophisticated analytics, thanks to the recent rapid advances in both the sensing hardware technologies and machine learning techniques. While providing improved user experiences and better business insights, the use of biometrics has raised serious privacy concerns, due to their intrinsic sensitive nature and the accompanying high risk of leaking personally identifiable and private information.

In this paper, we propose a novel utility-preserving biometric anonymization framework, which provides a method to anonymize a biometric dataset without introducing artificial or external noise, with a process that retains features relevant for downstream machine learning-based analyses to extract interesting attributes that are valuable to relevant services, businesses, and research organizations. We carried out a thorough experimental evaluation using publicly available visual and vocal datasets. Results show that our proposed framework can achieve a high level of anonymization, while at the same time retain underlying data utility such that subsequent analyses on the anonymized biometric data could still be carried out to yield satisfactory accuracy.

1 Introduction

As sensing technologies get increasingly adopted into commodity electronic devices that people use in their daily lives, biometrics have become more accessible and appealing as an information source, for example to enable seamless authentication without manual password input [1]. What's more, the latest sensing technologies have gone way beyond just targeting more traditional biometrics such as fingerprints, whose sole usage is arguably authentication only. Today's sensing devices can collect rich biometrics such as facial imagery, voice, and even posture/gait, iris, and neural signal data. With the help of the recent rapid advances in machine learning techniques, a wide range of interesting analytics can then be performed on the rich biometric data [2], for example, to infer or extract

information such as age, gender, dialect, sentiment, emotion, focus level, medical condition, etc., which could then enable vast opportunities in various relevant services and business interests.

Despite the high potential value of biometric information, one major concern preventing its universal collection and utilization is its linkage to personal identity and potential privacy violation [3–5]. For example, a user might enjoy the convenience of Face Unlock on their personal electronic devices, but likely would not appreciate having their facial features and identity information collected and used for targeted advertisements. Similarly, businesses have deployed Voice ID authentication in their automated phone system to streamline their customer service call experience. It would be deeply problematic if a business extracts information such as age, gender, and race from the voice data and uses it to profile each of their individual customers for preferential treatments.

It is therefore our goal to devise a data transformation mechanism to resolve this conflict between the value of biometric data and the potential identity disclosure. The problem of de-identification has been studied for the past decades [6]. Ideally for our particular case of biometrics, a successful anonymization should transform the data such that no identity information could be recovered, but at the same time other interesting attributes are left intact. Such a biometric anonymization mechanism would be tremendously valuable across a multitude of use cases. For example, a marketing firm that has recruited a focus group to study people's preference towards different products by presenting to them series of images of new products and taking pictures of their facial reactions for analysis might want to anonymize their collected facial imagery data and transfer it to a technology company focusing on developing computer vision algorithms and software. Or, an international medical research institute that has collected detailed biometric records from a large population might have completed their study of a particular disease and would like to release an anonymized version of the dataset publicly so other medical researchers could carry out their own studies on the dataset and potentially make discoveries that are related, or even orthogonal, to the data's original purpose.

To make the data release and reuse possible, the key challenge lies in the high dimensionality of biometric data as well as in the intrinsic probabilistic nature of machine learning-based analytics performed on top of it. In comparison, for traditional tabular data where the useful information associated with each data record is simply the textual content itself (e.g., date of birth, zip code, etc.), a rich body of literature exists that provides promising anonymization results. For biometric data, on the other hand, each data record on itself (e.g., facial image, voice audio clip, etc.) is essentially just a blob of bits, and does not show its useful information without either manual labeling or automated machine learning-based analyses, which by nature is probabilistic. Even though from a philosophical point of view, our goal of preserving interesting attributes and removing identities might seem self-contradicting in that any features preserved could potentially be used for re-identification, we argue that our problem at hand around biometrics is far from being binary. On the contrary, the high dimension-

ality of the data itself and the probabilistic nature of machine learning-based analytics introduce a high degree of uncertainty that we can take advantage of to achieve retention of interesting attributes while performing anonymization.

In this paper we introduce a novel biometric data transformation framework that aims at accomplishing this exact goal, namely anonymize raw biometric data to prevent/minimize identity breaching in a manner that retains other data characteristics for successful subsequent analytics. Our contribution is three-fold:

- To the best of our knowledge, our proposed framework is the first one to introduce the concept of *utility preservation under the context of ML-based analytics with general biometric information anonymization.*
- We introduce a novel anonymization technique that uses a dynamically assembled random set and task-oriented machine learning models to help guide a selective weighted-mean based transformation to anonymize biometric records.
- We demonstrate the effectiveness of our method's identity protection and utility preservation via a thorough experimental evaluation using publicly available multi-modal datasets.

2 Basic Concepts & Problem Statement

Since our objective is to transform a private biometric dataset for public release such that personal identities cannot be recovered but data utility is preserved as much as possible, we would like to define a few terms we use as well as making a clear problem statement for our proposed utility-preserving data anonymization task, just so we are on level ground going forward with our discussion.

2.1 Basic Concepts

Regarding the utility of a biometric dataset, we define *attribute of interest* and *additional attributes*, as follows.

- *Attribute of Interest.* An individual's biometric data contains features that can be used to predict certain attributes about them. An *attribute of interest* is an attribute detectable from biometric data, whose value must be protected. For example, the sentiment states displayed in a set of facial images could be considered as an attribute of interest due to their potential uses in computer vision studies or business applications. Therefore, in anonymizing such a facial dataset, we want to preserve the discoverability of sentiment states of the images.
- *Additional Attribute.* Features detectable from the biometric data, in addition to the *attribute of interest*, are denoted *additional attributes*. For example, from a voice dataset, information such as age group and dialect can be extracted by analyzing each audio clip. If the age group information is the sole attribute of interest, the dialect information is considered an additional attribute. Preserving the dialect as well as the age group information while anonymizing the voice dataset could be desirable for the expanded potential usages of a public release.

2.2 Problem Statement

Due to the high dimensionality of biometric data and the high uncertainty of ML-based analytics, we argue it is impossible to formulate a provable security guarantee for our biometric anonymization problem at hand. Therefore, in this paper we propose a purely *data-driven* approach so that the level of utility preservation and the level of anonymization can both be quantified, experimentally through measurements.

For an original biometric dataset $\mathbb{D}$, suppose it has an attribute of interest p and a set of n additional attributes $\{q_n\}$, with their corresponding recognition models $\mathcal{P}(\cdot)$ and $\{\mathcal{Q}_n(\cdot)\}$ all trained from the original dataset $\mathbb{D}$. Suppose $\mathbb{D}$ has an identity classification model $\mathcal{I}(\cdot)$, also trained from the original dataset. Then, for any data transformation $\mathcal{T}(\cdot)$, we can represent the *Utility* $U(\cdot)$ of the transformed data as the collective attribute recognition accuracy

$$U(\mathcal{T}(\mathbb{D})) = \mathcal{P}(\mathcal{T}(\mathbb{D})) + \sum_{i=1}^{n} \alpha_i \mathcal{Q}_i(\mathcal{T}(\mathbb{D})),$$

and what we call *Identity Mixture* $M(\cdot)$ the degree to which the trained identity classification model is confused by the transformed data

$$M(\mathcal{T}(\mathbb{D})) = 1 - \mathcal{I}(\mathcal{T}(\mathbb{D})).$$

In the formulas, $\mathcal{T}(\mathbb{D})$ is the transformed biometric dataset, $\{\alpha_n\}$ are user input weights for the additional attributes. Each of the attribute recognition models $\mathcal{P}(\cdot)$ and $\{\mathcal{Q}_n(\cdot)\}$, as well as the identity classification model $\mathcal{I}(\cdot)$, takes as input an entire dataset and outputs its accuracy. Intuitively, to find the best anonymization for a biometric dataset $\mathbb{D}$ is to find the optimal $\mathcal{T}^*(\cdot)$ that maximizes both U and M (or achieves a good trade-off between them), which is to say that the corresponding transformed data thoroughly confuses the identity classification model but can still be used to reliably extract interesting attributes.

2.3 Attack Model

From our problem statement, we would like to make an important observation on the identity classification model $\mathcal{I}(\cdot)$: Only the data owner knows the ground-truth identity correspondence between the original data $\mathbb{D}$ and the transformed data $\mathcal{T}(\mathbb{D})$. Therefore, only the data owner can compute the accuracy $\mathcal{I}(\mathcal{T}(\mathbb{D}))$. Any attacker who tries to use an identity classifier $\mathcal{I}'(\cdot)$ would not be able to recover any identities because of the apparent lack of ground-truth identity correspondence between $\mathbb{D}$ and $\mathcal{T}(\mathbb{D})$. Therefore, even if the attacker's model $\mathcal{I}'(\cdot)$ correctly classified $x\%$ of the hidden identities in $\mathcal{T}(\mathbb{D})$, the attacker would not be able to tell which $x\%$ in $\mathcal{T}(\mathbb{D})$ the model $\mathcal{I}'(\cdot)$ got correctly. Thus, their attempted re-identification attack is reduced to random guess.

Additionally, we argue that it is reasonable to assume the data owner's model is always more powerful than the attacker's model, $\forall \mathbb{D} : \mathcal{I}(\mathbb{D}) \geq \mathcal{I}'(\mathbb{D})$, because the data owner and the attacker can both select the latest and most powerful identity classification algorithm, but the data owner has the advantage of having

access to the original unanonymized data, which the attacker does not have. Therefore, if we let the attacker's model be the same as its upper bound, $\mathcal{I}'(\cdot) = \mathcal{I}(\cdot)$, we can treat the data owner's measured identity mixture M, to be the lower bound of what the attacker can possibly experience. In other words, *the already hidden identity in the anonymized data would appear even more mixed to an attacker*. Therefore, in our discussion, we assume that *i)* the data owner only releases the final anonymized data, and nothing else, and *ii)* the identity classification model used by the attacker is effectively the same as the model used by the data owner.

Our attack model gives us a solid ground for our subsequent discussions We believe that, in practice, our data-driven approach can bring value to a wide range of application scenarios.

3 Rationale of Approach

To achieve our objective of utility-preserving anonymization for biometrics, the high dimensionality of the data and the uncertainty of ML-based analytics need to be accounted for. For each data record $\mathbf{d}$ we aim to anonymize, we dynamically assemble a random set containing $\mathbf{d}$ and perform a selective weighted-mean-based operation, where the weighting is only applied to the most important features, as guided by task-specific machine learning models. We intend to make our data transformation retain as much truthfulness as possible, hence our particular design follows the intuition of only utilizing information from the original biometric dataset, and purposefully avoiding external artificial noise. Therefore, the transformation step $\mathcal{T}(\cdot)$ randomly assembles a short-lived, parameter-driven (such parameters include desired set size, attribute purity, etc., which are discussed in detail in Sec. 4.1) set of feature vectors with which to calculate the weighted-mean for each of the target feature vectors being anonymized.

Under our proposal, each data record becomes different from its original form. Also, due to the high dimensional nature of biometrics, it is also unlikely for any anonymized data record to have an exact match in the original dataset, or vice versa. As will be demonstrated in Sec. 5.2, regardless of the particular attack method of choice—be it a direct distance measure between two data records or via a trained ML model to compute the probability of a match—the likelihood of an attacker being able to link any anonymized data record to its true corresponding original record is reduced to a random guess on the entire dataset. In other words, an anonymized record is equally likely to be the closest, or the farthest, or anywhere in between, to its true match, as far as re-identification is concerned. Hence, the attacker is unable to reliably recover any identities from the anonymized biometric dataset.

4 Methodology

In this section, we discuss our proposed framework for performing utility-preserving anonymization on biometric data. Our proposal is generally applicable to all types of biometrics, and not restricted to any particular data modalities or feature extraction methods. For example, as demonstrated in Sec. 5, our method is evaluated on both facial image-based and voice audio-based datasets, where

multiple different feature extraction methods are used, including no feature extraction at all (e.g., raw image pixels).

4.1 Dynamically Assembled Random Set

Regardless of the particular preprocessing and feature extraction, each biometric data record is essentially a feature vector, which we transform through a series of operations, starting with dynamically assembling a random set of other data records from the dataset to be anonymized. For each target data record $\mathbf{d}$ in the original dataset $\mathbb{D}$, we assemble a set F of g data records (g can be roughly interpreted as the size of the crowd that $\mathbf{d}$ is hiding in, and can be determined experimentally, as demonstrated in Sec. 5), where $\mathbf{d} \in F$, and the rest $g-1$ of the data records, $F \setminus \{\mathbf{d}\}$, are selected from $\mathbb{D}$ based on their attribute-of-interest values, according to a *purity* parameter

$$t = \frac{|\{\mathbf{f} \in F | p_{\mathbf{f}} = p_{\mathbf{d}}\}|}{g},$$

where $p_{\mathbf{f}}$ denotes the value of $\mathbf{f}$'s attribute of interest. For example, if $t = 1$, then all g elements in F share the same attribute-of-interest value as $\mathbf{d}$; if $t = \frac{|\{\mathbf{k} \in \mathbb{D} | p_{\mathbf{k}} = p_{\mathbf{d}}\}|}{|\mathbb{D}|}$, which is the proportion of $p_{\mathbf{d}}$ in the entire population $\mathbb{D}$, then all of F's elements are to be uniformly randomly selected from $\mathbb{D}$ regardless of their attribute-of-interest values; if $t = \frac{1}{g}$, then all other elements in F are selected to be of different attribute-of-interest values than $\mathbf{d}$. This way of assembling the set F is inspired by the k-anonymity, ℓ-diversity, and t-closeness methods, but differs in that our approach was designed with the main objective of preserving the attribute of interest, while also including mechanisms for trading off between attribute preservation and identity mixture, in the form of different set sizes $g \in \mathbb{Z}^+$ and purity levels $t \in \left[\frac{1}{g}, 1\right]$.

4.2 Selective Weighted Mean-based Transformation

After dynamically assembling a random set F, we transform the target biometric record $\mathbf{d}$ by computing its weighted mean with the rest of F's elements $F \setminus \{\mathbf{d}\}$. In order to preserve $\mathbf{d}$'s attributes, we want to protect its corresponding features by assigning them a higher *weight* such that they do not get completely buried when $\mathbf{d}$ is averaged with the rest of F. The higher weight, the more we anchor $\mathbf{d}$'s features in place during averaging.

One caveat of the weighting is that on one hand it protects the target's features, but on the other hand, it could potentially weaken the identity mixing effect of the averaging. For example, for $g \equiv |F| = 10$, a high weight $w = 1000$, would virtually completely anchor a target in place, rendering the transformation almost trivial. To mitigate this problem, we modify the weighting strategy such that the weights are only applied to selective "important" features of each biometric record as derived from the task-specific machine learning model for the attribute of interest. For example, a sentiment classifier on facial features might pay more attention to features around the mouth. Our weighted mean calculation would therefore apply nontrivial weights to the target $\mathbf{d}$'s mouth features,

Algorithm 1 Utility Preserving Biometric Anonymization

Inputs:

$\mathbb{D}$: the set of original biometric data feature vectors to be anonymized
$\mathcal{P}(\cdot)$: the classifier trained on $\mathbb{D}$ for the attribute of interest
$\{\mathcal{Q}_n(\cdot)\}$: the classifiers trained on $\mathbb{D}$ for the additional attributes
c_p : number of features to retain for attribute of interest
$\{c_{q_n}\}$: numbers of features to retain for each of the additional attributes
g : size of random set for anonymization
t : purity of the random set's attribute-of-interest value
w : weight parameter for computing weighted mean

Output:

$\mathbb{D}'$: the set of anonymized biometric data feature vectors

1: $I_p \leftarrow$ list of feature indices in descending order of importance from $\mathcal{P}(\cdot)$
2: $\{I_{q_n}\} \leftarrow$ lists of feature indices in descending order of importance from $\{\mathcal{Q}_n(\cdot)\}$
3: $I \leftarrow I_p[0:c_p] \cup \{\cup_{i\in\{n\}} I_{q_i}[0:c_{q_i}]\}$
4: $X_I \leftarrow$ indicator vector s.t. $X_I[j] = \begin{cases} 1, \text{ if } j \in I \\ 0, \quad \text{o.w.} \end{cases}$
5: $\mathbb{D}' \leftarrow \emptyset$
6: **for each** $\mathbf{d} \in \mathbb{D}$ **do**
7: Randomly select $F \subseteq \mathbb{D}$ s.t. $\mathbf{d} \in F$, $|F| = g$, and $\frac{|\{\mathbf{f}\in F | p_{\mathbf{f}} = p_{\mathbf{d}}\}|}{g} = t$
8: $\mathbf{d}' \leftarrow \frac{1}{w} \cdot \text{mean}(F) + \frac{(w-1)}{w} \cdot X_I \odot \mathbf{d}$
9: $\mathbb{D}' \leftarrow \mathbb{D}' \cup \{\mathbf{d}'\}$
10: **end for**
11: **return** $\mathbb{D}'$

and use $w = 1$ for other non-important features like the hair. This way, the weighting helps protect biometric attributes without running the risk of largely fixing target biometric records unchanged and causing low identity mixtures.

Note that not only can the attribute of interest be preserved by the weighting, so can any additional attributes. For example, in addition to the sentiment attribute of interest, the data owner of a facial image dataset might also want to preserve gaze directions as an additional attribute. In that case, they would query their gaze detector for a set of relevant features, which most likely are around the eyes. And these eye features would then be added to the list of features that nontrivial weights are applied to.

The algorithm pseudo-code is shown in Alg. 1. Line 1 through 4 collect the set of important features from the corresponding task-specific ML models. Line 7 prepares the dynamically assembled random set as discussed in Sec. 4.1. The selective weighted mean as discussed in Sec. 4.2 is computed on Line 8, where $\odot$ is the component-wise multiplication operator. Please note that even though we only use a single weight w here, the algorithm can be easily extended to incorporate multiple weights, one per attribute for example, by modifying the indicator-vector preparation on Line 4 and/or the averaging computation on Line 8. Lastly, each iteration of the for-loop in Line 6 through 10 is independent from the rest, leading to highly parallelizable and efficient computation in practice.

5 Experimental Evaluation

In this section, we experimentally evaluate our biometric anonymization technique using publicly available datasets. First, we describe the characteristics of the datasets we use, and the experimental settings. Next, we report the results of the various sets of experiments where we compare the effects of parameters in our proposed technique by examining its capabilities of identity mixture and biometric attribute preservation under various experimental settings.

5.1 Experimental Setup

Datasets. Our framework enables the preservation of multiple attributes of biometric data while performing anonymization. Thus, an ideal dataset for us to use to demonstrate this capability would be one that contains ground-truth label information for multiple interesting attributes. We curated two publicly available datasets that fitted our requirement for testing our method.

The first one is the facial image FER-2013 dataset [7], which contains grayscale images of human faces with associated ground-truth sentiment label information, and thus suits our purpose. A round of manual inspection was performed on the original dataset to remove problematic images that were duplicates, non-photographic, or of poor resolution, etc. We treat *sentiment* as an example biometric attribute of interest in our experiments. Moreover, we augment FER-2013 with the *mouth-slightly-open* attribute using a model pre-trained on the CelebFaces Attributes (CelebA) dataset [8] as an additional attribute. As a result, the final in-use FER-2013 dataset has 8,470 training images, 978 validation images and 1,060 testing images, and it has 4 classes for the sentiment attribute and two classes for the mouth-slightly-open attribute.

The other one is the voice AudioMNIST [9] dataset, which contains the waveform signal of different people speaking digits. We use *spoken digit* as the attribute of interest in our experiments. We sub-sampled the dataset to rebalance the difference classes since the original class distributions were highly skewed. We ended up with 7,200 training audios, 2,400 validation audios, and 2,400 testing audios. The dataset contains 24 speaker identities and has 10 classes for the spoken digit. For both datasets, the training and validation splits are used to train the classifiers and we use the testing split to evaluate our proposed method.

Data Preprocessing. For FER-2013, we experimented with multiple feature extraction methods as the representation for each data records: *i)* FaceGraph, which is the fully-connected graph built on facial landmarks extracted from each facial image (using the Swift Vision Library [10]); *ii)* Pixel, which simply uses the raw pixel values of an image as the feature vector; *iii)* Eigenface [11], which is the projection of an facial image onto the eigenspace computed from all facial images; and *iv)* Vggfeats, which is the feature of the final layer of the facenet [12] neural network. For AudioMNIST, on the other hand, we extract the embeddings by using HuBERT [13] on the voice signal and then average the embeddings of each token as our final data representation.

Evaluation Protocol. We use the *classification on attribute of interest* as a driving example for our experiments. Each classification task itself, however,

is *not* necessarily our focus—our goal is not to find the model that achieves absolutely the best accuracy for a classification test; rather, we are mostly interested in demonstrating that a model trained on the original biometric data can continue to successfully perform classification tasks even on the version of the biometric data transformed by our anonymization method. Therefore, we simply experimented with a few well-known classification algorithms and empirically picked the one that struck a balance between classification performance and training speed. We ended up picked the off-the-shelf Random-Forest [14] classifier from scikit-learn [15] for our experiments for attribute classification. It provided good accuracy on both the attribute of interest and the additional attribute on FER-2013 as well as AudioMNIST.

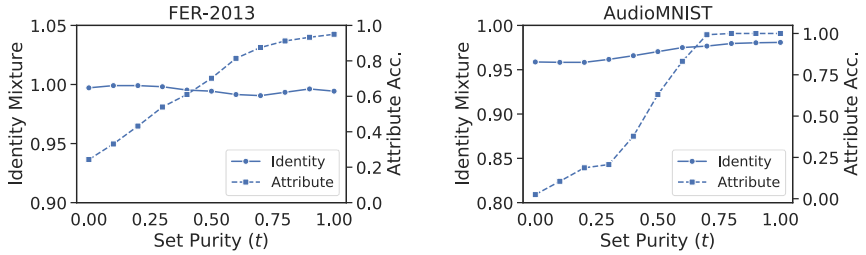
The evaluation protocol is setup as follows. First, to evaluate the preserved attribute of interest, we train and test a random-forest classifier on the original unanonymized data. We then apply this classifier on the anonymized data to check the level of preservation on the attribute of interest. We also evaluate the level of identity mixture on the anonymized data. Since the FER-2013 dataset does not come with identity information, when evaluating the level of identity mixture, we simply consider each image as a different identity and then measure the *cosine distance* between each anonymized data record to all originals to find the closest one as the potential match. AudioMNIST, on the other hand, does include the identity information. So, we employ an ML-based method to measure the level of identity mixture, where we train a multi-layer perceptron (MLP) over the identity labels using the original dataset and then evaluate its performance on the anonymized dataset.

Feature Ranking. In our proposed method, we need to rank all data features in order to decide which ones to retain. There are existing feature ranking methods that determine the importance of each feature [16, 17]. The random-forest classifier also ranks each feature upon building its decision trees, which we directly use as our metrics to gauge the importance of each feature.

Parameter Settings. We carried out a thorough scan through the parameter space in order to uncover all interesting trends and crucial regions in our experiments. For the sake of presentation brevity, we report in Sec. 5.2 only the representative results, under the following parameter settings:

- Set attribute purity t: ranges from 0.0 to 1.0 with step size 0.1;
- Set size: $g = 8, 32, 128$;
- Feature retention ratio $r_p = c_p/|\mathbf{d}|$ for the attribute of interest: $r_p = 0.1\%, 1\%, 10\%, 50\%, 100\%$;
- Feature retention ratio $r_q = c_q/|\mathbf{d}|$ for the additional attributes: $r_q = 0\%, 0.1\%, 1\%, 10\%, 50\%$;
- Weight: $w = 10, 100, 1000$,

where c_p, c_q, and $\mathbf{d}$ are as defined in Alg. 1. After we explored these parameters (see Fig. 1 through 4), we used $t = 0.6$, $g = 32$, $r_p = 1\%$, $w = 100$ for FER-2013, and $t = 0.8$, $g = 128$, $r_p = 1\%$, $w = 10$ for AudioMNIST.



(a) $g = 32, w = 100, r_p = 1\%, r_q = 0\%$ (b) $g = 128, w = 10, r_p = 1\%, r_q = 0\%$

Fig. 1: Varying set purity t. Higher t leads to better attribute-of-interest recognition accuracy as each original data record $\mathbf{d}$ is combined with more records sharing $\mathbf{d}$'s attribute value.

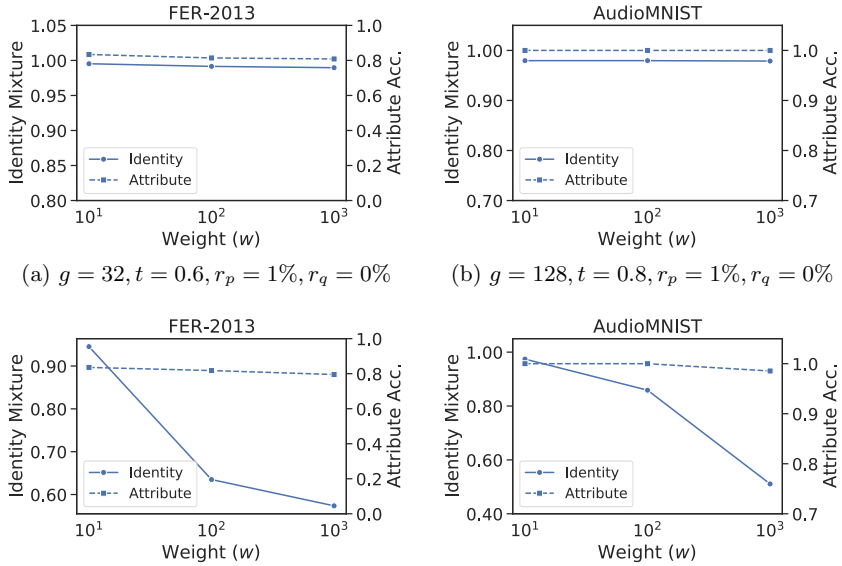
5.2 Results

We organize our results as follows. First, we report the attribute recognition accuracies on the original unanonymized dataset as baselines. Next, we perform ablation studies on the parameters of our method and discuss the result. We then take a closer examination of the quality of the anonymization achieved by our method. Please note that all these above results on FER-2013 are obtained by using the FaceGraph feature representation. So lastly, we experiment with applying our method on all four different feature representations on FER-2013, as discussed in Sec. 5.1, and report our findings.

Performance on Unanonymized Data. Before any discussion on the anonymized biometric data, we first establish a reference point by obtaining the accuracy of the classification model for the attribute of interest on original unanonymized data. We expect this classification result to be reasonably accurate because otherwise it would be difficult to assess the level of utility preservation if the original biometric dataset already had low utility to begin with. For the attributes of interest on FER-2013 and AudioMNIST, the random-forest classifier achieved 77% and 90.6% recognition accuracy, respectively.

Effects of Parameters. Figures 1 through 4 show how each parameter affects the data transformation's identity mixture and preservation of the attribute of interest. In each of these experiments, we tune a single parameter while keeping the rest fixed at the optimal configuration we obtained empirically.

First, we examined the influence of the set purity t, which determines the percentage of the data records sharing the same attribute value as the target in each random set, as defined in Sec. 4.1. As shown in Fig. 1, the set purity and the recognition accuracy of the attribute of interest on the anonymized data is positively correlated, which demonstrates that our method can indeed preserve the attribute of interest effectively. On the other hand, varying the purity level does not affect the level of identity mixtures.



(a) $g = 32, t = 0.6, r_p = 1\%, r_q = 0\%$ (b) $g = 128, t = 0.8, r_p = 1\%, r_q = 0\%$

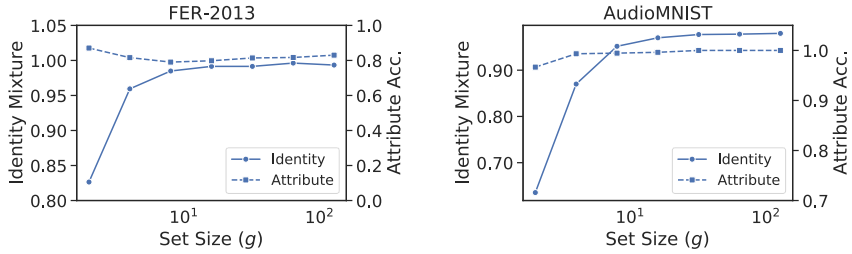
(c) $g = 32, t = 0.6, r_p = 50\%, r_q = 0\%$ (d) $g = 128, t = 0.8, r_p = 50\%, r_q = 0\%$

Fig. 2: Varying weight w. Under $r_p = 1\%$, our method works well regardless of the weight since only 1% of features are retrained. On the other hand, with $r_p = 50\%$, the identity mixture decreases when w increases because the anonymized data record is now much closer to the original one because of the large portion of features being retained via a higher r_p and anchored in place via a higher w.

The weight w controls how much a data record is anchored in place during transformation in terms of its retained features. Its other features would still be blended with the other data records. As shown in Fig. 2, when we set the feature retention to only keep $r_p = 1\%$ of features, even with very small weight, we can still achieve high recognition accuracy for the attribute of interest and high identity mixture on anonymized data. On the other hand, when we retain $r_p = 50\%$ features, the larger weight w results in lower identity mixture as the anonymized data is now much too similar to the original data.

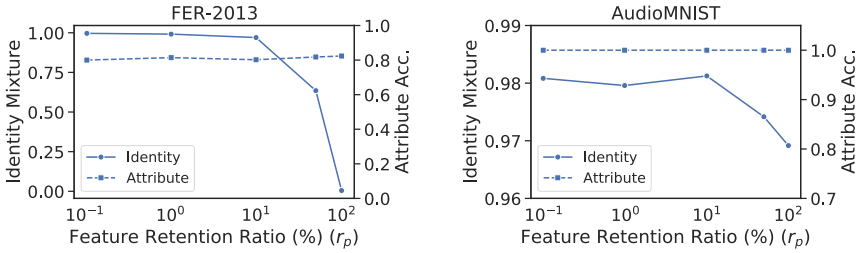
The set size g is related to the size of the population each data record is to be mixed with. Therefore, a larger g would lead to a more diverse set for our method to increase the level of identity mixture. On the other hand, as we can control the set purity t, the result set will affect identity mixture more than it does the attribute of interest. As shown in Fig. 3, identity mixture improves when set size increases, whereas the recognition accuracy of the attribute of interest remains relatively unchanged.

For the feature retention ratio r_p for the attribute of interest, retaining more features would lead to smaller difference between the original data record and its anonymized version, resulting in lower identity mixture, as can be observed



(a) $t = 0.6, w = 100, r_p = 1\%, r_q = 0\%$ (b) $t = 0.8, w = 10, r_p = 1\%, r_q = 0\%$

Fig. 3: Varying set size g. With larger set size, identity mixture increases as mixing more data leads to better anonymization without affecting the recognition of the attribute of interest.



(a) $g = 32, t = 0.6, w = 100, r_q = 0\%$ (b) $g = 128, t = 0.8, w = 10, r_q = 0\%$

Fig. 4: Varying feature retention ratio r_p. Retaining more features increases the similarity between the original and anonymized data. Therefore, it helps increase attribute recognition accuracy, but lead to lower identity mixture. Hence, a trade-off needs to be made here.

in Fig. 4. On the other hand, thanks to feature ranking, even if only $r_p = 1\%$ of features are retained, the recognition accuracy of the attribute of interest remains unaffected even though the identity mixture drops significantly. As we expect identity mixture to be high in an anonymized dataset, we can use such experimental parameter space exploration to help locate desirable configuration. For example, for the feature retention ratio in the range $r_p \in [0.1\%, 10\%]$, we observe, for both FER-2013 and AudioMNIST, both high levels of identity mixture and high attribute recognition accuracy—both are desirable characteristics for utility-preserving anonymization.

Additional Attribute. We next demonstrate, using FER-2013, the preservation of not only the attribute of interest, but also an additional attribute, while performing anonymization. The results are shown in Fig. 5. It can be seen that when we retain more features related to the additional attribute, the recognition accuracy of the attribute of interest stays the same while the recognition accu-

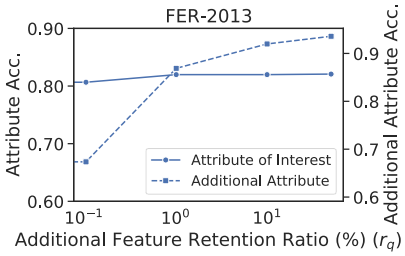


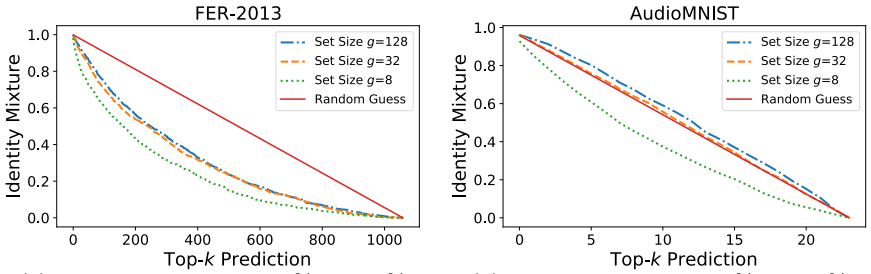
Fig. 5: Varying feature retention ratio r_q for the additional attribute for FER-2013 ($g = 32, t = 0.6, w = 100, r_p = 1\%$). The corresponding identity mixtures are 0.99, 0.99, 0.97, 0.70, from left to right. Again, a trade-off can be made here that achieves good recognition accuracy for both the attribute of interest and the additional attribute, as well as a high degree of identity mixture.

racy for the additional attribute enjoys a drastic boost. For example, when we retain just 1% of the features for the addition attribute, its recognition accuracy increases by ∼15% without decreasing identity mixture, which is at 0.99. This clearly demonstrates that our method can effectively preserve multiple attributes when performing anonymization.

Anonymization Quality. We have so far been judging the quality of anonymization via identity mixture. While an informative metric, identity mixture does not paint the whole picture, as it is based only on the binary hit-or-miss results of identity classification models. A "perfect" anonymization would reduce an attacker's re-identification attempts to random guesses, which means the attacker gains zero information with the attacks. Therefore, we use two methods to take a deeper look into the anonymization quality achieved by our proposed approach, with different set sizes. *i)* The level of identity mixture over top-k predictions, and *ii)* The KL divergence between the predicted probability and that of random guesses. Results are shown in Fig. 6 and Fig. 7.

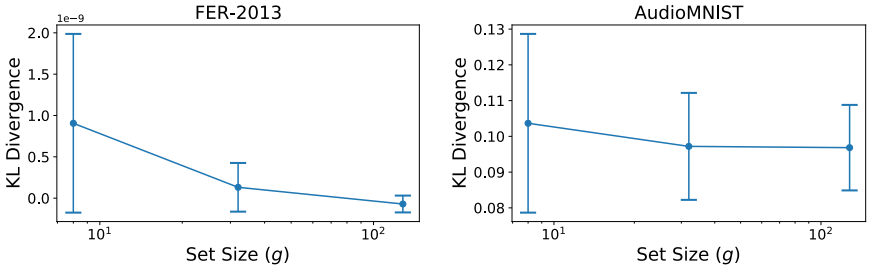
The level of identity mixture over top-k prediction means that a re-identification attack is considered successful if the true identity is contained in the attacker's top k candidate matches. As shown in Fig. 6, if the curve is below and close to that of the random guess, it implies that the data are anonymized in a way that the attacker can only achieve random guess in re-identification attacks. Moreover, if the curve is above the random guess, it means the anonymized data can actually fool the attacker better than random guess, in which case the attacker might as well try guessing randomly.

We also measure how far the predicted distribution deviates from that of random guesses. A value close to zero means that the attacker won't be able to do better than random guess. We compute each KL divergence from random guess for each data record and then average across the whole dataset. Results are shown in Fig. 7, where we observe that *i)* the overall KL divergence values are already close to zero, indicating good anonymization qualities, and *ii)* with



(a) $t = 0.6, w = 100, r_p = 1\%, r_q = 0\%$ (b) $t = 0.8, w = 10, r_p = 1\%, r_q = 0\%$

Fig. 6: The level of identity mixture over top-k predictions, where a re-identification attack is considered successful if the true identity is contained in the attacker's top k candidate matches. The straight lines correspond to random guesses.



(a) $t = 0.6, w = 100, r_p = 1\%, r_q = 0\%$ (b) $t = 0.8, w = 10, r_p = 1\%, r_q = 0\%$

Fig. 7: KL Divergence from random guess at different set size. For FER-2013, the KL divergence from random guess is almost 0, whereas the overall KL divergence remain very small for AudioMNIST.

larger set size, re-identification attempts tend to behave increasingly more like random guesses.

Different Data Representations. Lastly, we experiment with multiple different biometric feature extractions and data representations. Results are shown in Fig. 8. First, it can be observed that our method is applicable to different data representations. For example, when setting the feature retention ratio to $r_p \in [0.1\%, 1\%]$, good identity mixture is observed for all different data representations, even though they do show varying recognition accuracies for the attribute of interest. In this particular example, our FaceGraph representation happens to give the best result among all. We also observe that the Vggfeats performs the worst, likely due to the low resolution of the images and the fact that the domain of images is different from that of the pretrained model. In general, the optimal data representation as well as parameter configuration can

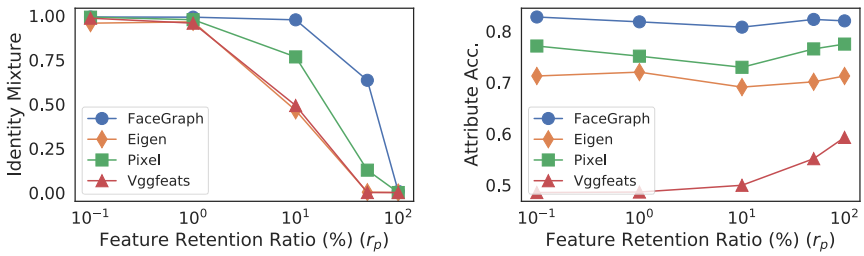


Fig. 8: Different data representations for FER-2013. The rest of the parameter setting is $g = 32, t = 0.6, w = 100, r_q = 0\%$. Our method is applicable to different data representations, and with only 1% features, the method can increase the identity mixture to almost 1.0 while retaining good accuracy on the attribute of interest.

always be found by our empirical data-driven approach, to tailor for the specific data type, utility-preserving needs, and anonymization requirements.

6 Related Work

Closely related research can be considered those efforts to apply or adapt the data truthfulness preserving anonymity techniques, for example k-anonymity, ℓ-diversity, and t-closeness to various data sources, ranging from categorical data that might appear in for example relational database tables, to location data and biometric data. Typically, most applications seek to find that balance between anonymizing the data effectively while also retaining utility to some degree [18, 19]. The fundamental difference of our proposed method from these existing techniques is that *we do not generalize*, as each transformed biometric data record remains different from the others, which opens up the possibility of more interesting attributes being preserved through anonymization. Next, we briefly survey these techniques.

Categorical data was one of the first sources for application of k-anonymity [20] and ℓ-diversity [21]. k-Anonymity, which our dynamically assembling a random set technique is inspired by, is a property induced in the data by generalization and suppression, which means each record is indistinguishable from $k-1$ other records. ℓ-Diversity addresses the weakness of k-anonymity with regards to attribute disclosure, by demanding that there are at least ℓ well-represented values of the attribute within each group of indistinguishable records (equivalence class). Achieving the property of k-anonymity in a database can be a challenging task. In fact, Bayardo and Agrawal [22] highlight that achieving optimal k-anonymity on a database is an NP-hard computational problem and they propose an optimization technique to achieve a given level of k-anonymity automatically. t-Closeness extends the protection of k-anonymity by requiring that the distribution of a sensitive attribute in an equivalence class be similar to the global attribute distribution, so no information is leaked by a potentially

altered cluster-specific distribution, which is another concept we borrowed when dynamically assembling a random set in our proposed pipeline.

Biometric image data is most relevant to the contribution in this paper. Initially obfuscation of image data was achieved by blurring, blacking out or pixelating salient regions of the face [23]. While these methods achieve admirable anonymization of the biometric data, they do not retain any degree of utility. Anonymization of facial image biometric data while retraining utility has been addressed by several prior studies [11, 24–26], which introduced the k-same family of algorithms. In general, this family of algorithms work as follows with different variations: Firstly, the biometric data in the database is partitioned into clusters, usually with k individuals per cluster, for the required level of k-anonymity. The centroids of the clusters are computed and the k individuals in a cluster are replaced by their corresponding cluster centroid. In this way every individual in the cluster shares the same de-identified face (i.e., the centroid). The algorithms vary in how individuals are assigned to clusters, e.g., using label information or not, and in what space the analysis is performed, e.g., pixel space, parameter space. The k-same family of algorithms were not extended to enforce ℓ-diversity in attributes that could be considered sensitive. Furthermore, many of the instantiations of k-same operate in pixel-space, leading to degradation in the utility of the anonymity representations, e.g., via excessive blurring induced by the centroid computation. Recent works have explored more advanced anonymization models such as neural networks for face de-identification [27–29]. While showing impressive clarity in generating fake faces for replacing real faces, these techniques require large amounts of training data and are also difficult to interpret or reason about, making it difficult to audit the models for industrial applications. It is also in question as to whether the focus should be on accurate reproduction of life-like anonymized images in pixel-space or a focus on generating highly anonymized abstract representations that can retain utility for other tasks, we follow the latter approach in this paper. The survey article [23] has further related work on face de-identification for the interested reader.

7 Conclusions

In this paper we introduce a biometric data anonymizing transformation framework that aims at stripping away personally identifiable information while at the same time preserving the utility of the biometrics by leaving intact its other characteristics such that downstream tasks such as machine learning-based analytics could still extract useful and valuable attributes from the anonymized biometric data. We present our end-to-end algorithm design, which uses dynamically assembled random set and selective weighted-mean to transform biometrics. We experimentally evaluated our method using publicly available facial image and voice audio datasets and observed that our proposed method could effectively anonymize the different modalities of biometrics, while at the same time successfully preserve other interesting attributes for downstream analytics.

Disclaimer

This paper was prepared for information purposes by the teams of researchers from the various institutions identified above, including the Global Technology Applied Research group of JPMorgan Chase Bank, N.A.. This paper is not a product of the Research Department of JPMorgan Chase Bank, N.A. or its affiliates. Neither JPMorgan Chase Bank, N.A. nor any of its affiliates make any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, but limited to, the completeness, accuracy, reliability of information contained herein and the potential legal, compliance, tax or accounting effects thereof. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

References

1. Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE access*, vol. 7, pp. 5994–6009, 2018.
2. N. Ortiz, R. D. Hernández, R. Jimenez, M. Mauledeoux, and O. Avilés, "Survey of biometric pattern recognition via machine learning techniques," *Contemporary Engineering Sciences*, vol. 11, no. 34, pp. 1677–1694, 2018.
3. M. Barni, R. Donida Labati, A. Genovese, V. Piuri, and F. Scotti, "Iris deidentification with high visual realism for privacy protection on websites and social networks," *IEEE Access*, vol. 9, pp. 131 995–132 010, 2021, 2169-3536.
4. P. Datta, S. Bhardwaj, S. N. Panda, S. Tanwar, and S. Badotra, "Survey of security and privacy issues on biometric system," in *Handbook of Computer Networks and Cyber Security*. Springer, 2020, pp. 763–776.
5. R. Donida Labati, V. Piuri, and F. Scotti, "Biometric privacy protection: guidelines and technologies," in *Communications in Computer and Information Science*, M. S. Obaidat, J. Sevillano, and F. Joaquim, Eds. Springer, 2012, vol. 314, pp. 3–19, 978-3-642-35754-1.
6. S. Garfinkel, *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology, 2015.
7. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *NIPS*, 2013.
8. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.
9. S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.
10. Apple, "Vision framework: Apply computer vision algorithms to perform a variety of tasks on input images and video," https://developer.apple.com/documentation/vision, 2021.
11. E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE TKDE*, 2005.
12. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

13. W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
14. L. Breiman, "Random forests," *Machine learning*, 2001.
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.
16. G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4202–4210.
17. Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'11. Arlington, Virginia, USA: AUAI Press, 2011, p. 266–273.
18. S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, no. 6, pp. 793–817, December 2012.
19. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity," in *Secure Data Management in Decentralized Systems*, T. Yu and S. Jajodia, Eds. Springer-Verlag, 2007.
20. P. Samarati, "Protecting Respondents' Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 13, no. 6, pp. 1010–1027, November/December 2001.
21. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM TKDD*, 2007.
22. R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *ICDE*, 2005.
23. S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content," *Image Commun.*, 2016.
24. R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *PET*, 2005.
25. R. Gross, L. Sweeney, F. de la Torre, and S. Baker, "Semi-supervised learning of multi-factor models for face de-identification," in *CVPR*, 2008.
26. Z. Sun, L. Meng, and A. Ariyaeeinia, "Distinguishable de-identified faces," in *FG*, 2015.
27. B. Meden, Z. Emersic, V. Struc, and P. Peer, "κ-same-net: neural-network-based face deidentification," in *IWOBI*, 2017.
28. Y.-L. Pan, M.-J. Haung, K.-T. Ding, J.-L. Wu, and J.-S. Jang, "K-same-siamese-gan: K-same algorithm with generative adversarial network for facial image de-identification with hyperparameter tuning and mixed precision training," in *AVSS*, 2019.
29. T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *CVPR Workshops*, 2019.