**1.2 Testing of a Dataset**

To experience an anonymization as well as deanonymization for ourselves, we created a fake data

set of 10,000 records. The attributes included are:

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Data Type |
| *\*\*Michael to replace with his info from the notebook\*\** | *Feel Free to change format* |  |

This data set was generated randomly, and provides us with several opportunities to try common anonymization

techniques discussed in this paper. The attributes were chose to reflect common member records that would go

under de-identification methods such as 1) aggregating full addresses to produce ambiguity, 2) changing specific ages to an age range, 3) removing all names and replacing with one field of an assigned identification number, 4) dividing the columns into less comprehensive tuples, 5) removing classes which do not meet a defined “K” threshold, and 6) perturbing data, making values difficult to discern while keeping distributions consistent.

*1.2.1 Anonymizing Data*

**Removing PII***\*\*Michael’s Section\*\**

**K-Anonymization**

Attributes Age, Ethnicity, Blood Type, State, Hair Color, Eye Color, and Hispanic\_Latino were all great candidates for examination with K-Anonymization, removing those elements that have a low repetition count. The chosen "K" typically depends on the sample size of the original dataset, and contextual knowledge of the domain. We chose to set our "K" threshold for K-Anonymization to 5% of the original sample size (i.e. 500 observations). This means, that for every attribute "class" we analyzed the frequency of that class, and for every class with less than 500 observations in the original dataset we removed all observations in that class. During the K-Anonymization process, we went from a sample size of 10,000 down to 7,662 removing classes such as:

* Age(after converting to age ranges): “>=80”
* Ethnicity: "American Indian or Alaskan Native", "Asian American", "Native Hawaiian or Other Pacific Islander"
* Blood Type: “AB+”, “AB-“, “B-“
* Hair Color: “Grey”, “Light Blond”, “Light Brown”
* Eye Color: “Black”, “Grey”, “Light Blue”, “Light Brown”, “Other”

Attributes State and Hispanic\_Latino were untouched during this process due to equal distribution amongst classes or the “K” threshold met for all classes. One notable disadvantage of this process, is that insights made on this dataset may not be extended to these demographics, as models created on the shared data may not hold true to the extended population.

**Stratification of the Response**

We were concerned about the distributions of each classifier matching publicly available data. For example, the American Red Cross shares publicly blood type distributions by ethnicity type[[1]](#endnote-1). We feared, that information like this, and similar information for other attributes may make it possible to identify data even after perturbation (as discussed in the next section). By pulling equal distributions of each ZVirus type, we slightly skewed the predictable distributions based on available population data. This also helps with prediction for our open source contributors, as we help eliminate bias towards dominating classes.

When identifying how many records are remaining in each ZVirus category, we found that within the remaining dataset, our least frequent ZVirus type was "Infected" with a frequency of 1528. In order to stratify, this means we could not have a sample size larger than 1528 \* 3 = 4584. Our chosen stratification was random samples without replacement taking 1,250 samples within each ZVirus type for a total sample size of 3,750.

**Perturbation**

*\*\*Michael’s Section\*\**

*1.2.2 Deanonymizing Data*

**The Attacker**

Bob…

**References**

1. **Blood Type Distributions by Ethnicity Type:**

   <http://www.redcrossblood.org/learn-about-blood/blood-types> [↑](#endnote-ref-1)