REVISITING THE UNIQUENESS OF

SIMPLE DEMOGRAPHICS

IN THE US POPULATION

Project Report

CSC 591 – Data Driven Privacy

Aditya Shirode

Kushan Kunal Prasad

Mentor: Dr. Jessica Staddon

# Abstract

According to a famous study based on 1990 Census Data, 87% of US population is uniquely identifiable given three attributes – Gender, Date of Birth, and Zip Code. A paper in 2006 revisited these demographics for the 2000 Census Data, and found that 63% of US population is uniquely identifiable; a considerable decrease from the previous result. In this report, we try to replicate the methods delineated by aforementioned studies, and revisit the uniqueness of US population by observing 2010 census data. Our results are nearly similar to those found in the 2006 study, and no significance difference is observed over various characteristics. We conclude that even after applications of data declassification, there remains a certain data which cannot be completely anonymized without incurring considerable information loss.

Table of Contents

[Abstract 2](#_Toc437911466)

[Introduction 4](#_Toc437911467)

[Methods 5](#_Toc437911468)

[Data Collection 5](#_Toc437911469)

[Data Analysis 6](#_Toc437911470)

[Results 7](#_Toc437911471)

[Anonymity, by age, given {Gender, Location, Full date of birth} 7](#_Toc437911472)

[Anonymity, by age, given {Gender, Location, Year of birth} 9](#_Toc437911473)

[Conclusion 11](#_Toc437911474)

[Appendix I – Java program code used for analysis 12](#_Toc437911475)

[Bibliography 17](#_Toc437911476)

[Acknowledgements 18](#_Toc437911477)

# Introduction

A famous study of the 1990[[1]](#footnote-1) census data showed that 87% (216 million of 248 million) of the population in the United States reported characteristics that likely made them unique based only on gender, 5-digit ZIP code and full date of birth (day, month and year). The study further reported that 53% of the U.S. population is uniquely identified only by {Gender, Place, Date of Birth}, where “place” is basically the city, town, or municipality in which the person resides. Even at the county level, {gender, county, date of birth} uniquely identifies 18% of the U.S. population. In general, it shows that “few characteristics are needed to uniquely identify a person.”

A paper published in 2006[[2]](#footnote-2) tried to revisit the findings on the recent decennial census at the time (2000 Census), and found that the results put forth by previous paper held their ground to an extent. But the uniqueness of the population had fallen down to 63% (177 million of 281 million). The paper clearly explained the data collection and methodology applied so that future privacy researchers can easily replicate their results. It also offered a fine-grained characterization by providing k-anonymity values for the demographics.

In this report, we try to replicate the methodology described in the 2006 paper on the latest available census data from 2010 Census. We try to find the anonymity values for population based on three attributes {Gender, Zip Code/ County, Date/Year of Birth}. The observations are made across two different granularity levels, 5-Digit Zip Code and County. As people tend to give out year of birth more frequently (by mentioning their age) rather than exact date (birthdate), we observe the data based on only year of birth as well.

Our results match to the ones found in 2006 study. No significant difference is observed in the statistics. Although, with advent of modern data protection and anonymization techniques, we had hypothesized that this difference would be significant. The rise of social media and the broadcasting of general information with constant updates might lead to easy re-identification of the population, proving these numbers wrong, which might be a cause for concern.

# Methods

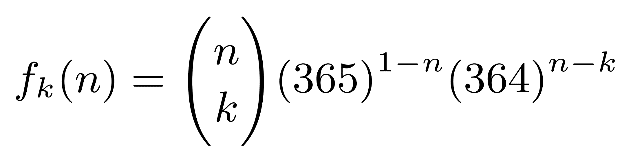
## Data Collection

Gender, Location (Zip Code/ County), and Age (Either Year of Birth, or the complete date of birth) are three most commonly disclosed demographic characteristics by people via registration forms, surveys, and other files. These simple characteristics were chosen to analyze the data from the 2010 census.

The demographic information can be easily downloaded from the Census Bureau’s official website[[3]](#footnote-3) free of charge. We are interested in the PCT12 database, a subset of the available census data. This table/set of data uniquely provides access to ‘lowest common denominator’ single year of age detail from under 1 year to 100 years of age for wide-ranging geographic areas enabling users to construct age cohort data not otherwise available. Table PCT12 is replicated for several race/ethnic combinations adding to analytical possibilities. We ignore the ethnic and racial information from the database in this report and only focus on the Age and Gender of the given population.

The retrieved data covers 3221 counties and county equivalents (boroughs, census tracts, parishes, independent cities and Municipals), and 33120 Zip Code Tabulation Areas (ZCTAs) in all 50 States, the District of Columbia and Puerto Rico. The 2010 census does not give a ‘wall-to-wall’ coverage of United States; there are certain ZCTAs which were left as ‘holes’, either because lack of human habitat or were covered in some other way.

Table PCT12 gives the year of birth, but not the month or day (no other table in the census data gives such detailed information, precisely because it is a threat to privacy). Hence, to consider one’s full date of birth (date, month, and year) to calculate its privacy implications, we must estimate the number of individuals who live in a given location and were born on a given day, month and year. The 2006 paper gave a statistical way of computing a precise estimate of the number of individuals born on a given day and month of the year as follows –



Where, *f­k(n)* = expected number of dates that contain *i* individuals

*n* individuals are distributed randomly and independently across *N* dates containing *i* individuals each.

Now we have obtained all of the required data we need to work on – Gender, Location, and Date of Birth.

## Data Analysis

While the Census Bureau Fact Finder Tool can be used to "call-up" and display/download the PCT12 data, it is difficult to use the data directly, or even open with a tool such as Excel, and easily view the data. The age detail is iterated by gender and it is cumbersome to aggregate by gender. To compare aggregated gender data for one area to another is also tedious. Labeling the age detail can be time consuming and challenging. Computing percentages requires additional steps.

The data provided in PCT12 format, downloaded from the Census site, is a csv (Comma Separated Values) file with about 3221 rows and 209 columns of data for the county level granularity (3.5MB size). The file contains 33120 rows and 209 columns over zip code granularity (31.3MB size). Both of the files can be opened and accessed using MS Excel.

For our analysis, we had to aggregate all this data, divided into various age groups (0-99, 100-104, 105-109, 110+), over genders (Male and Female), into a single table with one column for individuals of a certain age living in a county/ specific zip code. We found out the sum over these attributes as required using excel formulas. Then we ran a Java code to find further required attributes.

We used basic Mathematical functions provided by the Java language to perform basic operations over the csv data. Using the code provided in the appendix, you can find median value of the dataset, k-anonymity value over present granularity, and percentage of population uniquely identifiable. As US population covered by the 2010 Census (312,462,997 individuals) is out of bounds for normal data types for numbers provided by Java (Integer, Long, Double, etc), a special data type, *Big Integer*, had to be used. This came with certain restrictions, modifications, and restricted capabilities.

The Java program was fed the csv files downloaded from the Census site, and it generated two new csv files, one for each granularity level. These resultant files contained aggregation of data over certain aspects, which are mentioned in the *Results* section below. The Java code used for analysis is mentioned in the *Appendix I*.

# Results

As compared to the 2000 Census, the 2010 Census was bigger in size and details. The population increased by about 11% in the 10 years that passed, with a particular observable shift found in the age group of 30-50. This group of population, which was 30-50 years older during the 2000 Census, can be observed with a similar graph visible in the 40-60 year spectrum in the 2010 Census (Figure 1).

Table 1

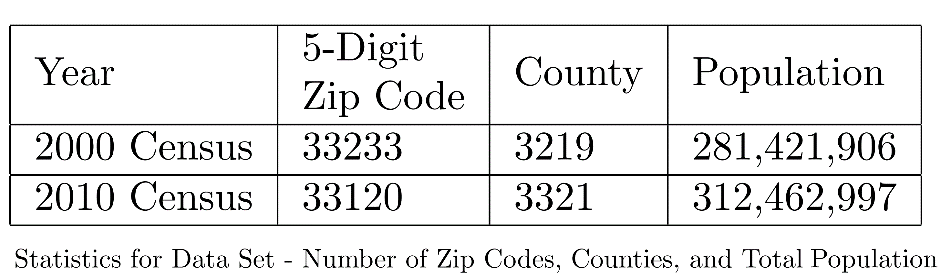
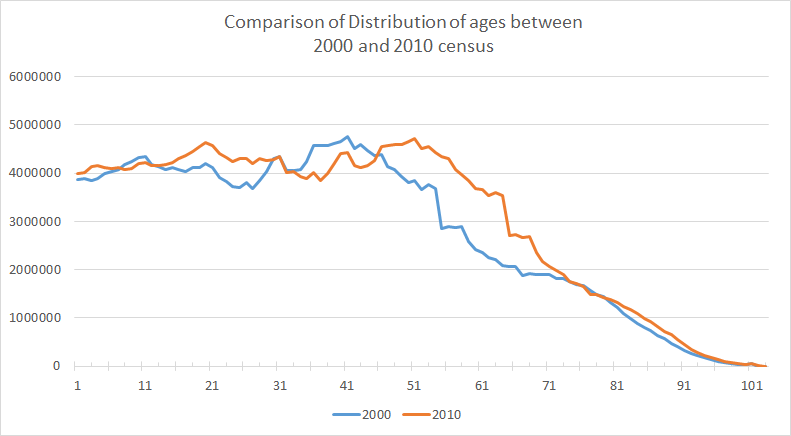


Figure 1



First, we compute the percentage of the U.S. population which is uniquely identifiable by {gender, location, date of birth}, where location is either a 5-digit ZIP code or a county, and date of birth is either the year of birth only, or the year and month of birth, or the full date of birth (year, month and day). Our results are summarized in Table 2.

Table 2

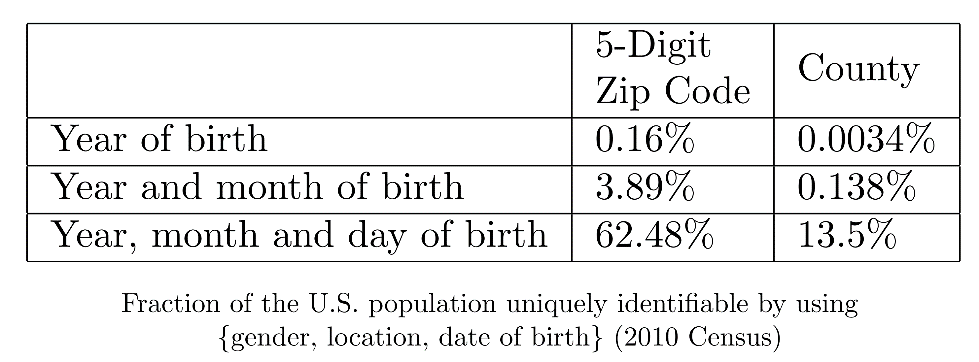
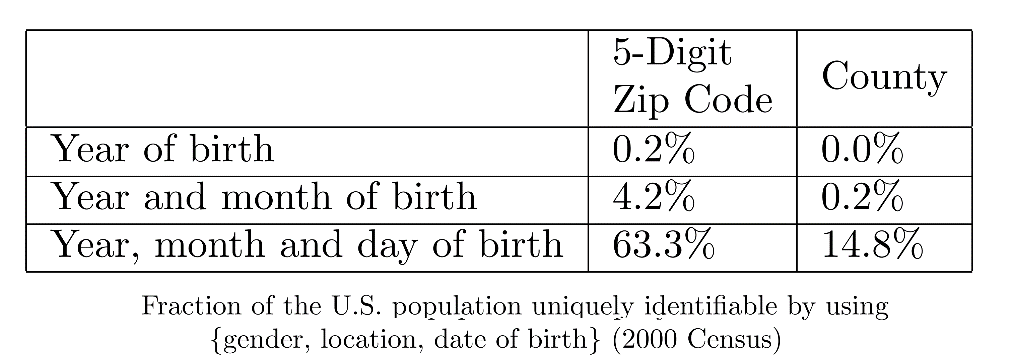


Table 3



Comparing them to the results obtained over 2000 census, summarized in Table 3, we can see that the numbers don’t differ that much. Although we can observe slight decrease in every statistic, the significance is not high enough.

About 62.8% of US Population in 2010 is uniquely identifiable by {gender, ZIP code, full date of birth}, as compared to 63.3% in the 2000, whereas it was found 87% uniquely identifiable by the same characteristics in 1990.

## Anonymity, by age, given {Gender, Location, Full date of birth}

Figures 2 and 3 give a more fine-grained view of the degree of anonymity of the US population, by age, given {gender, location, full date of birth}, where location is either a 5-digit ZIP code (Figure 2) or a county (Figure 3). These graphs show that the privacy threat of disclosing simple demographics is fairly uniform between the ages of 0 and 50, then rises rapidly after 50. The graphs also show that even individuals who are not uniquely identifiable enjoy very little anonymity.

For example, disclosing {gender, county, full date of birth} leaves 13.5% of the population uniquely identifiable (1-anonymous) but also leaves 34.9% of the population 5-anonymous or less (i.e. hidden indistinguishably in a group of size 5 or less). The proportion of individuals who are 5-anonymous or less rises to 52% for people over the age of 60 (see Figure 3).

Figure 2

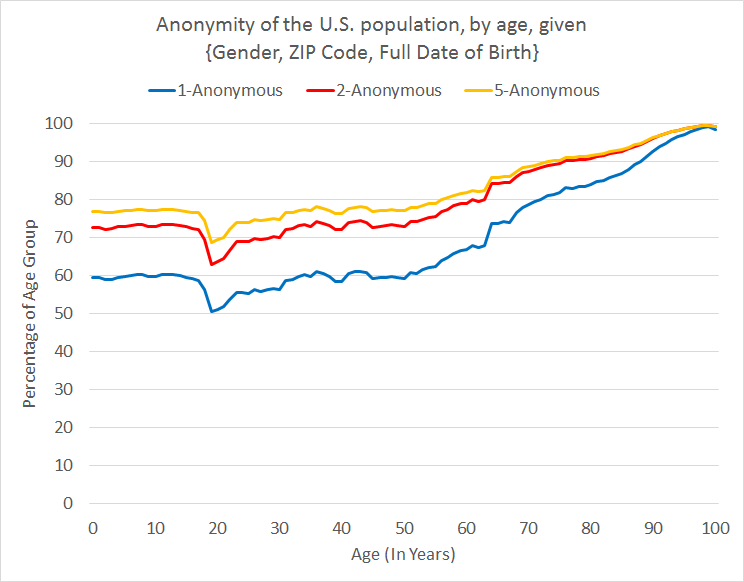
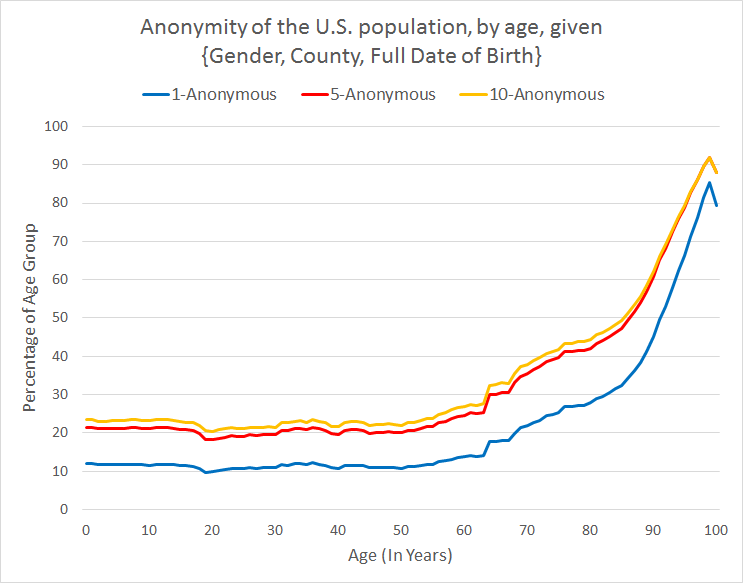


Figure 3



## Anonymity, by age, given {Gender, Location, Year of birth}

Table 1 shows that revealing one’s {gender, location, year of birth} allows for unique identification of only 0.2% of individuals. We analyze in more detail the privacy implications of disclosing one’s {gender, location, year of birth}. Figures 3 and 4 show the degree of anonymity of the US population, by age, given {gender, location, year of birth}, where location is either a 5-digit ZIP code (Figure 4) or a county (Figure 5). In both graphs, the green curve (the middle curve) shows the median degree of anonymity by age (the degree of anonymity of the 50-th percentile).

We see for example that the median anonymity of the population under age 50 after disclosing {gender, ZIP code, year of birth} is 184-anonymity (the median is 2927-anonymity if the county is disclosed instead of the ZIP code).

The sharp spikes around age of 20 are due to a reason observed in the previous study. It comes from college and university towns with high concentrations of individuals between the ages of 18 and 22. In our analysis, the top three entries, who enjoyed k-anonymity of about 4000, hailed from University towns. In fact, we found out they belonged to same Zip Code (77840, College Station, Texas). The top three unique zip codes (77840, 47906, 32304) with highest anonymity, all belonged to university towns.

Figure 4

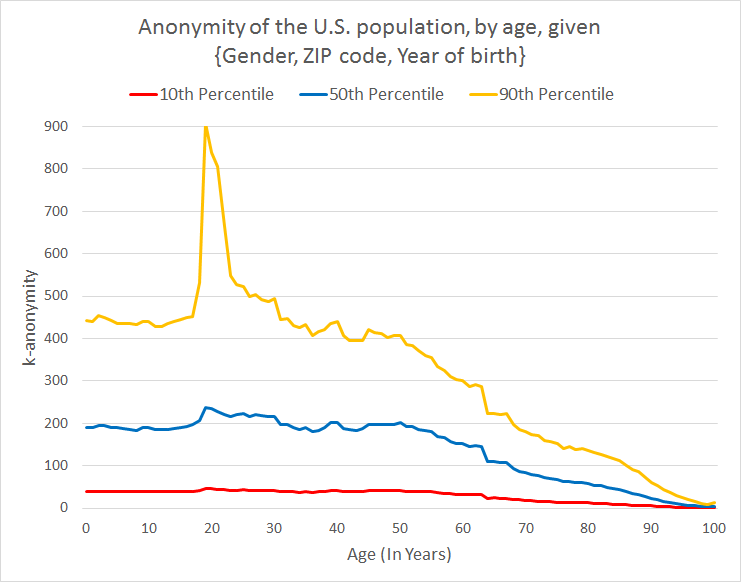
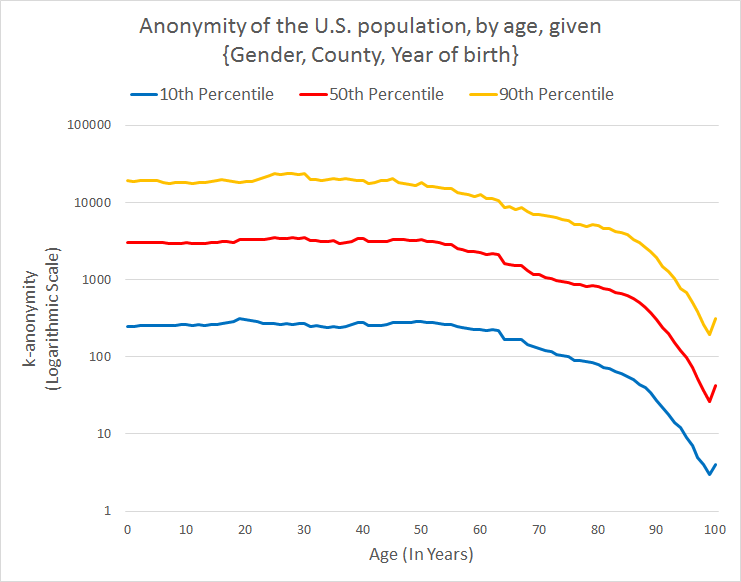


Figure 5



# Conclusion

This short project report revisits the uniqueness of simple demographics in the US population based on the most recent census data (the 2010 census). We offer a follow-up to the 2006 paper that visited the same demographical characteristics for the 2000 census. We found out that the uniqueness values for the population have gone down, but only slightly. With the advent of information available to a third party attacker due to rise of social media, this might pose a bigger threat of re-identification of unique individuals due to publicly released datasets in future.

# Appendix I – Java program code used for analysis

Classes –

Age.java

ZipYear.java

CountyYear.java

**Age.java**

import java.util.Arrays;

import java.math.\*;

public class Age{

int val[];

int pos;

int total;

int median;

int median9;

int median1;

Age(){

val=new int[33178\*2];//3219

pos=0;

total=0;

}

void add(int n){

val[pos]=n;

total+=n;

pos++;

}

void median(){

Arrays.sort(val);

int mid = this.total/2;

int runningsum=0;

int i=0;

while(runningsum<mid){

runningsum+=val[i++];

}

median=val[i-1];

mid = (int)((double)this.total \* 0.1);

runningsum=0;

i=0;

while(runningsum<mid){

runningsum+=val[i++];

}

median1=val[i-1];

mid = (int)((double)this.total \* 0.9);

runningsum=0;

i=0;

while(runningsum<mid){

runningsum+=val[i++];

}

median9=val[i-1];

}

double percentK(int k){

double sum=0;

for(int i=0;i<pos;i++)

sum+=threshold(val[i], k);

return (sum/(double)total)\*100;

}

double nonpercentK(int k){

double sum=0;

for(int i=0;i<pos;i++)

sum+=threshold(val[i], k);

return sum;

}

double threshold(int n,int k){

if(k<1)

return 0;

MathContext mc=new MathContext(20);//precision

BigDecimal result=new BigDecimal(1);

for(int i=0;i<k;i++)

result=result.multiply(new BigDecimal(n-i), mc);

result=result.divide(new BigDecimal(factorial(k)), mc);

result=result.multiply(new BigDecimal(364).pow(n-k, mc),mc);

result=result.divide(new BigDecimal(365).pow(n-1, mc),mc);

return result.doubleValue()+threshold(n,k-1);

}

int factorial(int n){

int fact=1;

for(int i=2;i<=n;i++)

fact\*=i;

return fact;

}

}

**ZipYear.java**

import java.io.\*;

import java.math.BigDecimal;

import java.math.MathContext;

import java.util.Arrays;

public class ZipYear {

public static void main(String args[])throws IOException{

BufferedReader br = new BufferedReader(new FileReader("zip.csv"));

int ctr=0,i,k=0;

String line=br.readLine();

String ip[]=null;

line=br.readLine();

Age ages[]=new Age[103];

for(i=0;i<103;i++)

ages[i]=new Age();

/\*\*

\* Parse dataset

\*/

while((line=br.readLine()) != null){

ip=line.split(",");

k=0;

for(i=5;i<108;i++){

ages[k++].add(Integer.parseInt(ip[i]));

}

i++;

k=0;

for(;i<212;i++){

ages[k++].add(Integer.parseInt(ip[i]));

}

}

/\*\*

\* Calculate percentile and median

\*/

int medians[]=new int[103];

PrintWriter writer = new PrintWriter("outputZip.csv", "UTF-8");

for(i=0;i<103;i++){

ages[i].median();

medians[i] = ages[i].median;

writer.println(i+","+ages[i].median+","+ages[i].median1+","+ages[i].median9);

System.out.println(ages[i].median);

}

writer.close();

/\*\*

\* Calculate overall median (~166)

\*/

Arrays.sort(medians);

double median;

median = (double) medians[medians.length/2];

System.out.println(median);

/\*\*

\* K-annonymity by birthdate

\*/

PrintWriter writer = new PrintWriter("outputZip.csv", "UTF-8");

for(i=0;i<103;i++){

writer.println(i+","+ages[i].percentK(1)+","+ages[i].percentK(2)+","+ages[i].percentK(5));

}

writer.close();

/\*\*

\* Overall unique

\*/

BigDecimal total=new BigDecimal(0),sum=new BigDecimal(0);

for(i=0;i<103;i++){

total=total.add(new BigDecimal(ages[i].total));

sum=sum.add(new BigDecimal(ages[i].nonpercentK(1)));

}

MathContext mc= new MathContext(20);

System.out.println(total+" stuff "+sum);

double percent = (sum.divide(total,mc).doubleValue())\*100;

System.out.println(percent);

}

}

**CountyYear.java**

import java.io.\*;

import java.util.Arrays;

import java.math.\*;

public class CountyYear {

public static void main(String args[])throws IOException{

BufferedReader br = new BufferedReader(new FileReader("county.csv"));

int ctr=0,i,k=0;

String line=br.readLine();

String ip[]=null;

line=br.readLine();

Age ages[]=new Age[103];

for(i=0;i<103;i++)

ages[i]=new Age();

/\*\*

\* Parse dataset

\*/

while((line=br.readLine()) != null){

ip=line.split(",");

k=0;

for(i=6;i<109;i++){

ages[k++].add(Integer.parseInt(ip[i]));

}

i++;

k=0;

for(;i<213;i++){

ages[k++].add(Integer.parseInt(ip[i]));

}

}

/\*\*

\* Calculate percentile and median

\*/

int medians[]=new int[103];

PrintWriter writer = new PrintWriter("outputCounty.csv", "UTF-8");

for(i=0;i<103;i++){

ages[i].median();

medians[i] = ages[i].median;

writer.println(i+","+ages[i].median+","+ages[i].median1+","+ages[i].median9);

System.out.println(ages[i].median);

}

writer.close();

/\*\*

\* Calculate overall median (~2515)

\*/

Arrays.sort(medians);

double median;

median = (double) medians[medians.length/2];

System.out.println(median);

/\*\*

\* K-annonymity by birthdate

\*/

PrintWriter writer = new PrintWriter("outputCounty.csv", "UTF-8");

for(i=0;i<103;i++){

writer.println(i+","+ages[i].percentK(1)+","+ages[i].percentK(5)+","+ages[i].percentK(10));

}

writer.close();

/\*\*

\* Overall unique

\*/

BigDecimal total=new BigDecimal(0),sum=new BigDecimal(0);

for(i=0;i<103;i++){

total=total.add(new BigDecimal(ages[i].total));

sum=sum.add(new BigDecimal(ages[i].nonpercentK(1)));

}

MathContext mc= new MathContext(20);

System.out.println(total+" stuff "+sum);

double percent = (sum.divide(total,mc).doubleValue())\*100;

System.out.println(percent);

}

}

# Bibliography

1. L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000
2. Philippe Golle. 2006. Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM workshop on Privacy in electronic society (WPES '06)
3. U.S. Census Bureau FactFinder - <http://factfinder.census.gov>

# Acknowledgements

We would like to thank our Project Mentor and Instructor, Prof Jessica Staddon, for her valuable help and guidance during this project. Without her inputs and follow-ups, completing this would have been a tough task. We would also like to thank our fellow classmates from the CSC 591 – Data Driven Privacy class of Fall 2015, for their challenging questions, which always gave us a new dimension to add to the project.

1. L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data

   Privacy Working Paper 3. Pittsburgh 2000 [↑](#footnote-ref-1)
2. Philippe Golle. 2006. Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM workshop on Privacy in electronic society (WPES '06) [↑](#footnote-ref-2)
3. U.S. Census Bureau FactFinder - http://factfinder.census.gov [↑](#footnote-ref-3)