Simple Demographics Often Identify People Uniquely

Latanya Sweeney
Carnegie Mellon University
latanya@andrew.cmu.edu

This work was funded in part by H. John Heinz III School of Public Policy and Management at Carnegie Mellon University and by a grant from the U.S. Bureau of Census.

1. Abstract

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently. It was found that combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are publicly available in this form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

2. Introduction

Data holders often collect person-specific data and then release derivatives of collected data on a public or semi-public basis after removing all explicit identifiers, such as name, address and phone number. Evidence is provided in this document that this practice of de-identifying data and of ad hoc generalization are not sufficient to render data anonymous because combinations of attributes often combine uniquely to re-identify individuals.

2.1. Linking to re-identify de-identified data

In this subsection, I will demonstrate how linking can be used to re-identify de-identified data. The National Association of Health Data Organizations (NAHDO) reported that 44 states have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth [1]. These data collections often include the patient's *ZIP* code, *birth date*, *gender*, and *ethnicity* but no explicit identifiers like *name* or *address*. The leftmost circle in Figure 1 contains some of the data elements collected and shared.

For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes [2]. The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using *ZIP*, *birth date* and *gender* to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals. The question that remains of course is how unique would such linking be.

In general I can say that the greater the number and detail of attributes reported about an entity, the more likely that those attributes combine uniquely to identify the entity. For example, in the voter list, there were 2 possible values for gender and 5 possible five-digit ZIP codes; birth dates were within a range of 365 days for 100 years. This gives 365,000 unique values, but there were only 54,805 voters.

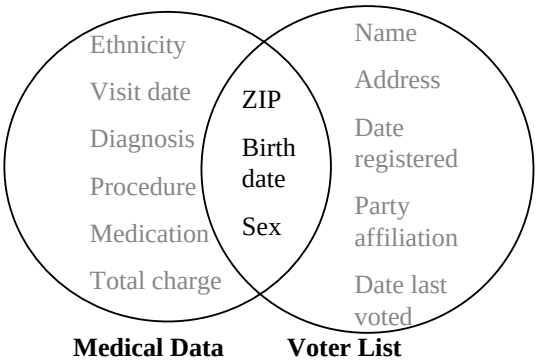

Figure 1 Linking to re-identify data

2.2. Publicly and semi-publicly available health data

As mentioned in the previous subsection, most states (44 of 50 or 88%) collect hospital discharge data [3]. Many of these states have subsequently distributed copies of these data to researchers, sold copies to industry and made versions publicly available. While there are many possible sources of patient-specific data, these represent a class of data collections that are often publicly and semi-publicly available.

#	Field description	Size
1	HOSPITAL ID NUMBER	12
2	PATIENT DATE OF BIRTH (MMDDYYYY)	8
3	SEX	1
4	ADMIT DATE (MMDYYYY)	8
5	DISCHARGE DATE (MMDDYYYY)	8
6	ADMIT SOURCE	1
7	ADMIT TYPE	1
8	LENGTH OF STAY (DAYS)	4
9	PATIENT STATUS	2
10	PRINCIPAL DIAGNOSIS CODE	6
11	SECONDARY DIAGNOSIS CODE - 1	6
12	SECONDARY DIAGNOSIS CODE - 2	6
13	SECONDARY DIAGNOSIS CODE - 3	6
14	SECONDARY DIAGNOSIS CODE - 4	6
15	SECONDARY DIAGNOSIS CODE - 5	6
16	SECONDARY DIAGNOSIS CODE - 6	6
17	SECONDARY DIAGNOSIS CODE - 7	6
18	SECONDARY DIAGNOSIS CODE - 8	6
19	PRINCIPAL PROCEDURE CODE	7
20	SECONDARY PROCEDURE CODE - 1	7
21	SECONDARY PROCEDURE CODE - 2	7
22	SECONDARY PROCEDURE CODE - 3	7
23	SECONDARY PROCEDURE CODE - 4	7
24	SECONDARY PROCEDURE CODE - 5	7
25	DRG CODE	3
26	MDC CODE	2
27	TOTAL CHARGES	9
28	ROOM AND BOARD CHARGES	9
29	ANCILLARY CHARGES	9
30	ANESTHESIOLOGY CHARGES	9
31	PHARMACY CHARGES	9
32	RADIOLOGY CHARGES	9
33	CLINICAL LAB CHARGES	9
34	LABOR-DELIVERY CHARGES	9
35	OPERATING ROOM CHARGES	9
36	ONCOLOGY CHARGES	9
37	OTHER CHARGES	9
38	NEWBORN INDICATOR	1
39	PAYER ID 1	9
40	TYPE CODE 1	1
41	PAYER ID 2	9
42	TYPE CODE 2	1
43	PAYER ID 3	9
44	TYPE CODE 3	1
45	PATIENT ZIP CODE	5
46	Patient Origin COUNTY	3
47	Patient Origin PLANNING AREA	3
48	Patient Origin HSA	2
49	PATIENT CONTROL NUMBER	
50	HOSPITAL HSA	2

Figure 2 IHCCCC Research Health Data

The Illinois Health Care Cost Containment Council (IHCCCC) is the organization in the State of Illinois that collects and disseminates health care cost data on hospital visits in Illinois. IHCCCC reports more than 97% compliance by Illinois hospitals in providing the information

[4]. Figure 2 contains a sample of the kinds of fields of information that are not only collected, but also disseminated.

Of the states mentioned in the NAHDO report, 22 of these states contribute to a national database called the State Inpatient Database (*SID*) sponsored by the Agency for Healthcare Research and Quality (AHRQ). A copy of each patient's hospital visit in these states is sent to AHRQ for inclusion in *SID*. Some of the fields provided in *SID* are listed in Figure 3 along with the compliance of the 13 states that contributed to *SID*'s 1997 data [5].

Field	Comments	#states	%states
Patient Age	years	13	100%
Patient Date of birth	month, year	5	38%
Patient Gender		13	100%
Patient Racial background		11	85%
Patient ZIP	5-digit	9	69%
Patient ID	encrypted (or scrambled)	3	23%
Admission date	month, year	8	62%
Admission day of week		12	92%
Admission source	emergency, court/law, etc	13	100%
Birth weight	for newborns	5	38%
Discharge date	month, year	7	54%
Length of stay		13	100%
Discharge status	routine, death, nursing home, etc	13	100%
Diagnosis Codes	ICD9, from 10 to30	13	100%
Procedure Codes	from 6 to 21	13	100%
Hospital ID	AHA#	12	92%
Hospital county		12	92%
Primary payer	Medicare, insurance, self-pay, etc	13	100%
Charges	from 1 to 63 categories	11	85%

Figure 3 Some data elements for AHRQ's State Inpatient Database (13 participating states)

State	Month and Year of Birth date	Age
Arizona	Yes	Yes
California		Yes
Colorado		Yes
Florida		Yes
Iowa	Yes	Yes
Massachusetts		Yes
Maryland		Yes
New Jersey		Yes
New York	Yes	Yes
Oregon	Yes	Yes
South Carolina		Yes
Washington		Yes
Wisconsin	Yes	Yes

Figure 4 Age information provided by states to SID

Figure 4 lists the states reported in Figure 3 that provide the *month and year of birth* and the *age* for each patient.

The remainder of this document provides experimental results from summary data that show how demographics often combine to make individuals unique or almost unique in data like these.

2.3. A single attribute

The frequency with which a single characteristic occurs in a population can help identify individuals based on unusual or outlying information. Consider a frequency distribution of birth years found in the list of registered voters. It is not surprising to see fewer people present with earlier birth years. Clearly, a person born in 1900 is unusual and by implication less anonymous in data.

2.4. More than one attribute

What may be more surprising is that combinations of characteristics can combine to occur even less frequently than the characteristics appear alone.

ZIP	Birth	Gender	Race
60602	7/15/54	m	Caucasian
60140	2/18/49	f	Black
62052	3/12/50	f	Asian

Figure 5 Data that looks anonymous

Consider Figure 5. If the three records shown were part of a large and diverse database of information about Illinois residents, then it may appear reasonable to assume that these three records would be anonymous. However, the 1990 federal census [6] reports that the ZIP (postal code) 60602 consisted primarily of a retirement community in the Near West Side of Chicago and therefore, there were very few people (less than 12) of an age under 65 living there. The ZIP code 60140 is the postal code for Hampshire, Illinois in Dekalb county and reportedly there were only two black women who resided in that town. Likewise, 62052 had only four Asian families. In each of these cases, the uniqueness of the combinations of characteristics found could help re-identify these individuals.

Race	Birth	Gender	ZIP	Problem
Black	09/20/65	m	02141	short of breath
Black	02/14/65	m	02141	chest pain
Black	10/23/65	f	02138	hypertension
Black	08/24/65	f	02138	hypertension
Black	11/07/64	f	02138	obesity
Black	12/01/64	f	02138	chest pain
White	10/23/64	m	02138	chest pain
White	03/15/65	f	02139	hypertension
White	08/13/64	m	02139	obesity
White	05/05/64	m	02139	short of breath
White	02/13/67	m	02138	chest pain
White	03/21/67	m	02138	chest pain

Figure 6 De-identified data

As another example, Figure 6 contains de-identified data. Each row contains information about a distinct person, so information about 12 people is reported. The table contains the following fields of information {*Race/Ethnicity, Date of Birth, Gender, ZIP, Medical Problem*}.

In Figure 6, there is information about an equal number of African Americans (listed as *Black*) as there are Caucasian Americans (listed as *White*) and an equal number of men (listed as *m*) as there are women (listed as *f*), but in combination, there appears only one Caucasian female.

2.5. Learned from the examples

These examples demonstrate that in general, the frequency distributions of combinations of characteristics have to be examined in combination with respect to the entire population in order to determine unusual values and cannot be generally predicted from the distributions of the characteristics individually. Of course, obvious predictions can be made from extreme distributions --such as values that do not appear in the data will not appear in combination either.

3. Background of definitions and terms

Definition (informal). Person-specific data Collections of information whose granularity of details are specific to an individual are termed *person-specific data*. More generally, in *entity-specific data*, the granularity of details is specific to an entity.

Example. Person-specific data

Figure 5 and Figure 6 provide examples of person-specific data. Each row of these tables contains information related to one person.

The idea of anonymous data is a simple one. The term "anonymous" means that the data cannot be linked or manipulated to confidently identify the individual who is the subject of the data.

Definition (informal). Anonymous data *Anonymous data* implies that the data cannot be manipulated or linked to confidently identify the entity that is the subject of the data.

Most people understand that there exist explicit identifiers, such as name and address, which can provide a direct means to communicate with the person. I term these *explicit identifiers*; see the informal definition below.

Definition (informal). Explicit identifier An *explicit identifier* is a set of data elements, such as {*name, address*} or {*name, phone number*}, for which there exists a direct communication method, such as email, telephone, postal mail, etc., where with no additional information, the designated person could be directly and uniquely contacted.

A common incorrect belief is that removing all explicit identifiers such as name, address and phone number from the data renders the result anonymous. I refer to this instead as *de-identified data*; see the informal definition below.

Definition (informal). De-identified data *De-identified data* result when all explicit identifiers, such as name, address, or phone number are removed, generalized or replaced with a made-up alternative.

Example. De-identified data

Figure 5 and Figure 6 provide examples of de-identified person-specific data. There are no explicit identifiers in these data.

Because a combination of characteristics can combine uniquely for an individual, it can provide a means of recognizing a person and therefore serve as an identifier. In the literature, such combinations were nominally introduced as *quasi-identifiers* [7] and *identificates* [3-58] with no supporting evidence provided as to how identifying specific combinations might be. Extending beyond the literature and its casual use in the literature, I term such a combination a quasi-identifier and informally define it below. I then examine specific quasi-identifiers found within publicly and semi-publicly available data and compute their general ability to uniquely associate with particular persons in the U.S. population.

Definition (informal). Quasi-identifier A quasi-identifier is a set of data elements in entity-specific data that in combination associates uniquely or almost uniquely to an entity and therefore can serve as a means of directly or indirectly recognizing the specific entity that is the subject of the data.

Example. Quasi-identifier

A quasi-identifier whose values are unique for all the records in Figure 6 is {*ZIP*, *gender*, *Birth*}.

In the next section, I will show that {*ZIP*, *gender*, *Birth*} is a unique quasi-identifier for most people in the U.S. population.

The term *table* is really quite simple and is synonymous with the casual use of the term data collection. It refers to data that are conceptually organized as a 2-dimensional array of rows (or records) and columns (or fields). A database is considered to be a set of one or more tables.

Definition (informal). Table, tuple and attribute A *table* conceptually organizes data as a 2-dimensional array of rows (or records) and columns (or fields). Each row (or record) is termed a *tuple*. A tuple contains a relationship among the set of values associated with an entity. Tuples within a table are not necessarily unique. Each column (also known as a field or data element) is called an *attribute* and denotes a field or semantic category of information that is a set of possible values; therefore, an attribute is also a domain. Attributes within a table are unique. So by observing a table, each row is an ordered *n*-tuple of values $<d_1, d_2, \ldots, d_n>$ such that each value d_j is in the domain of the *j*-th column, for $j=1, 2, \ldots, n$ where *n* is the number of columns.

In mathematical set theory, a relation corresponds with this tabular presentation; the only difference is the absence of column names. Ullman provides a detailed discussion of relational database concepts [9].

Examples of tables

Figure 5 provides an example of a person-specific table with attributes {*ZIP*, *Birth*, *Gender*, *Race*}. Each tuple concerns information about a single person. Figure 6 provides an example of a person-specific table with attributes {*Race*, *Birth*, *Gender*, *ZIP*, *Problem*}.

Unfortunately, the terminology with respect to data collections is not the same across communities and diverse communities have an interest in this work. In order to accommodate these different vocabularies, I provide the following thesaurus of interchangeable terms. In general, *data collection*, *data set* and *table* refer to the same representation of information though a data collection may have more than one table. The terms *record*, *row* and *tuple* all refer to same kind of information. Finally, the terms *data element*, *field*, *column* and *attribute* refer to the same kind of information. For brevity, from this point forward, I will use the more formal database terms of table, tuple and attribute. I do allow the tuples of a table to appear in a "sorted" order on occasion and such cases pose a slight deviation from its more formal meaning. These uses are explicitly noted.

4. Methods

4.1. Census Tables

Information from the 1990 US Census made available on the Web [10] and on CDROM [11] and from the U.S. Postal Service [12] was loaded into Microsoft Access and the following tables produced and used with Microsoft Excel.

1. ZIP census table provides 1990 federal census information summarized by each ZIP (postal code) in the United States.

2. Place census table provides 1990 federal census information summarized by place name (town, city, municipality, or postal facility name).

3. County census table provides 1990 federal census information summarized by US counties.

Figure 7 contains a list of attributes (or data elements) for each of these tables. The name and description of each attribute is listed and a "yes" appears in the column that associates the attribute to the ZIP, Place or County table in which the attribute appears. Information for all 50 states and the District of Columbia were provided. For example, values associated with the attribute *Tot_pop* in the ZIP table are the total numbers of individuals reported as living in each corresponding *ZIP*. Each tuple (or row) in the table corresponds to a unique ZIP.

Given a particular geographical specification such as ZIP, place or county, the number of people reported as residing in the noted geographical area is reported by age subdivision in the ZIP, Place and County tables. The age subdivisions are: under 12 years of age (denoted as *Aunder12*), between 12 and 18 years of age (denoted as *A12to18*), between 19 and 24 years of age (denoted as *A19to24*), between 25 and 34 years of age (denoted as *A25to34*), between 35 and 44 years of age (denoted as *A35to44*), between 45 and 54 years of age (denoted as *A45to54*),

between 55 and 64 years of age (denoted as *A55to64*) or more than 65 years of age (denoted as *A65Plus*).

Field	Description	ZIP	PLACE	COUNTY
StateID	State Code	yes	yes	yes
ZIP	5-digit ZIP	yes	**NO**	**NO**
Place	Name of Incorporated Place	**NO**	yes	**NO**
CoName	County Name	**NO**	**NO**	yes
Tot_Pop	Total Population	yes	yes	yes
AUnder12	Population Under Age 12 Years	yes	yes	yes
A12to18	Population Age 12-18 Years	yes	yes	yes
A19to24	Population Age 19-24 Years	yes	yes	yes
A25to34	Population Age 25-34 Years	yes	yes	yes
A35to44	Population Age 35-44 Years	yes	yes	yes
A45to54	Population Age 45-54 Years	yes	yes	yes
A55to64	Population Age 55-64 Years	yes	yes	yes
A65Plus	Population Age 65 Years and up	yes	yes	yes

Figure 7 1990 Census attributes in ZIP, Place, County tables

4.2. ZIPNameGIS Table

ZIP information provided from the U.S. Postal Service included place, which is a name of a town, city, municipality or postal facility uniquely assigned to a ZIP code. This information was loaded directly to provide the ZIPNameGIS table. The attributes (or data elements) for the ZIPNameGIS table are {*StateID, ZIP, State, POName, longitude, latitude, population*}.

The Place table was constructed by linking the ZIP table to the ZIPNameGIS table on *ZIP*. Results were then grouped by *POName* (respecting state designations) so that population information from multiple ZIP codes were grouped together by the city or town in which the ZIP code referred. Finally, the Place table was generated by collapsing these groupings into single entries that contained the sum of the population values reported for all ZIP codes corresponding to the same place.

During the process, 3 ZIP codes were found to cross state lines and therefore, be listed in two states. To avoid this duplication, the following assignments were made: (1) ZIP code 32530 refers to Pinetta in both Florida and Georgia. The Georgia entry was removed from Place; (2) ZIP code 42223 refers to Fort Campbell in both Kentucky and Tennessee. The Tennessee entry was removed from Place; and, (3) ZIP code 63673 refers to Saint Mary in both Illinois and Missouri. The Missouri entry was removed from Place.

4.2.1. Schemas of shared data

Figure 2 and Figure 3 contain descriptions of publicly and semi-publicly available hospital discharge data. Below are some quasi-identifiers found in those data that also appear in the census data. The experiments reported in this document estimate the uniqueness of values associated with these quasi-identifiers given the occurrences reported in the census data.

1. Illinois Research Health Data.

The Illinois Research Health Data (R_{rod}) is described in Figure 2. Among the attributes listed there, I consider QI_{rod} = {*date of birth, gender, 5-digit ZIP*} to be a quasi-identifier within R_{rod}.

2. AHRQ's State Inpatient Database

The Agency for Healthcare Research and Quality's State Inpatient Database (R_{SID}) is described in part in Figure 3. Among the attributes listed there, I consider QI_{SID1} = {*month and year of birth, gender, 5-digit ZIP*} to be a quasi-identifier within data released by some states and I consider QI_{SID2} = {*age, gender, 5-digit ZIP*} to be a quasi-identifier within data released by other states.

4.3. Design and procedures

The experiments reported in the next section can be generally described in terms of values attributes can assume. Let $T(A_1,\ldots,A_n)$ be an entity-specific table and let Q_T be a quasi-identifier of T. Q_T is represented as a finite set of attributes $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$. I write $|A_m|$ to represent the finite number of values A_m can assume. So, the number of distinct possible values that be assigned to Q_T, written $|Q_T|$, is: $|Q_T| = |A_i| * |A_{i+1}| * \ldots * |A_j|$.

Example.
Given Q_{dob}={*date of birth, gender*}, then $|Q_{dob}| = 365 * 76 * 2 = 55{,}480$ because there are 365 days in a year, an expected lifetime of 76 years, and 2 genders.

In this document, I am concerned with a person-specific table $T(A_1,\ldots,Z,\ldots,A_n)$ that includes a *geographic attribute Z*. Values assigned to a *geographic attribute* are specific to the residences of people. Examples of geographic attributes include 5-digit ZIP codes, names of cities and towns, and names of counties in which people reside. Let *U* be the universe of all people and the person-specific table $Geo[z_i, A_r, \ldots, A_s)$ contain all or almost all of the people of *U* having $Z=z_i$. I say Geo_{zi} is a *population register* for z_i. And, $T[A_1,\ldots Z_i,\ldots A_n]$ is a pseudo-random sample drawn from $Geo[z_i, A_r, \ldots, A_s]$. Unique and unusual combinations of characteristics found in Geo with respect to z_i can be no less unique or unusual when recorded in T. Therefore, the probability distribution of combinations of characteristics found in Geo limits the values those combinations of characteristics can assume in T. Determining unique and unusual combinations of characteristics within a residential domain is a counting problem.

Theorem. Generalized Dirichlet drawer principle [13]
(also known as the Generalized pigeonhole principle)

If *N* objects are distributed in *k* boxes, then there is at least one box containing at least $\lceil N / k \rceil$ objects.

Proof.
Suppose that none of the boxes contain more than $\lceil N / k \rceil$ -1 objects. Then, the total number of objects is at most: $k * (\lceil N / k \rceil - 1) < k * (((N / k) + 1) - 1) = N$
This has the inequality $\lceil N / k \rceil < (N / k) + 1$
This is a contradiction because there are a total of *N* objects.

Example.
Given a random sample of 500 people, there are at least $\lceil 500 / 365 \rceil = 2$ people with the same birthday because there are 365 possible birthdays.

Let z_i be a 5-digit ZIP code. I write *population*(z_i) to denote the number of people who reside in z_i and *population*(z_i) ≡ |Geo$_{zi}$|. If *population*(Z_i) > |Q_{dob}|, then by the generalized pigeonhole principle, a tuple $t \in R_{rod}$[*date of birth*, *gender*, z_i] would not uniquely correspond to one person. In these cases, I say t[A_1, …, *date of birth*, *gender*, z_i, …, *An*] is not likely to be uniquely identifiable. On the other hand, if *population*(z_i) ≤ |Q_{dob}| then by the generalized pigeonhole principle, a tuple $t \in R_{rod}$[*date of birth*, *gender*, z_i] would likely relate to only one person. In these cases, I say t[A_1, …, *date of birth*, *gender*, z_i, …, *An*] is likely to be uniquely identifiable. This is the general approach to the experiments reported in the next section though each differs in terms of attribute specification.

4.3.1. Subdivision analyses

The analyses of the identifiability of geographically situated populations are based on age-based divisions within a geographic attribute. Let age subdivision *a* be either *Aunder12*, *A12to18*, *A19to24*, *A25to34*, *A35to44*, *A45to54*, *A55to64*, or *A65Plus*. The quasi-identifier Q_a has the same attributes as Q_{dob} but values which *date of birth* can assume are limited by *a*. That is, |Q_a| is the number of possible distinct values that can be assigned to Q_a. I say |Q_a| is the *threshold* for Q_{dob} with respect to age subdivision *a*.

Example.
Given Q_{dob}={*date of birth*, *gender*} and age subdivision *a* = *A19to24*, then |Q_a| = 365 * 2 * 6 = 4380 because there are 365 birthdays, 2 genders and 6 years between the ages of 19 to 24, inclusive.

Number of subjects uniquely identified in a subdivision of a geographical area (ID_{aZi})

Given a value for a geographic attribute, written z_i, and an age subdivision *a*, I write *population*(z_i, *a*) as the number of people residing in z_i with an age within *a*. The number of people considered uniquely identified by *a* and Z_i, written ID_{aZi}, is determined by the rule:

if *population*(z_i, *a*) ≥ |Q_a|, then ID_{aZi}= *population*(z_i, *a*)
else ID_{aZi} = 0.

By extension, the percentage of people residing in z_i considered uniquely identified (written ID_{zi}) with respect to the set of age subdivisions is computed as:

$$ID_{Zi} = \frac{population(z_i) - \sum_{a=AUnder12}^{A65Plus} ID_{aZi}}{population(z_i)}$$

4.3.2. Statistics on geographical areas

Statistics are reported on geographic regions. Given a geographic attribute *Z*, let $Region_Z$ = {z_i | z_i ∈ *Z* } and *AgeDivs* = {*Aunder12*, *A12to18*, *A19to24*, *A25to34*, *A35to44*, *A45to54*, *A55to64*, *A65Plus*}. That is, $Region_Z$ is a set of values that can be assigned to the geographic

attribute Z and *AgeDivs* is a set of age subdivisions. $Region_Z$ is partitioned into NotIDSet and IDSet based on age subdivision a∈*AgeDivs* such that:

$$\text{NotIDSet}_{Za} = \{(z_i,a) \mid z_i \in Region_Z \text{ and } population(z_i, a) > |Q_a| \}$$
$$\text{IDSet}_{Za} = \{(z_i,a) \mid z_i \in Region_Z \text{ and } population(z_i, a) \le |Q_a| \}$$

The population of NotIDSet_Z is not considered uniquely identifiable by values of Q_{dob}. The population of IDSet_Z is considered uniquely identifiable by values of Q_{dob}. In the experiments, the following statistics are reported.

Maximum subpopulation(NotIDSet_{Za}) = max(*population*(z_1, *a*), …, *population*(z_y, *a*)), where $(z_i,a) \in \text{NotIDSet}_{Za}$

Maximum subpopulation(IDSet_{Za}) = max(*population*(z_1, *a*), …, *population*(z_y, *a*)), where $(z_i,a) \in \text{IDSet}_{Za}$

Minimum subpopulation(NotIDSet_{Za}) = min(*population*(z_1, *a*), …, *population*(z_y, *a*)), where $(z_i,a) \in \text{NotIDSet}_{Za}$

Minimum subpopulation(IDSet_{Za}) = min(*population*(z_1, *a*), …, *population*(z_y, *a*)), where $(z_i,a) \in \text{IDSet}_{Za}$

$$Average\ \ subpopulation(\text{NotIDSet}_{Za}) = \frac{\sum_{(z_i,a)\in NotIDSet} population(z_i, a)}{|\text{NotIDSet}_{Za}|}$$

$$Average\ \ subpopulation(\text{IDSet}_{Za}) = \frac{\sum_{(z_i,a)\in IDSet} population(z_i, a)}{|\text{IDSet}_{Za}|}$$

$$Number\ of\ \ geographical\ areas(\text{NotIDSet}_{Za}) = |\text{NOTIDSet}_{Za}|$$

$$Number\ of\ \ geographical\ areas(\text{IDSet}_{Za}) = |\text{IDSet}_{Za}|$$

$$Percentage\ of\ \ geographical\ areas(\text{NotIDSet}_{Za}) = \frac{|\text{NotIDSet}_{Za}|}{|\text{NotIDSet}_{Za}| + |\text{IDSet}_{Za}|}$$

$$Percentage\ of\ \ geographical\ areas(\text{IDSet}_{Za}) = \frac{|\text{IDSet}_{Za}|}{|\text{NotIDSet}_{Za}| + |\text{IDSet}_{Za}|}$$

State		AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	4,040,587	699,554	425,425	369,639	652,466	585,299	422,565	363,033	522,606
AK	544,698	123,789	53,662	46,478	111,790	101,699	55,887	29,236	22,157
AZ	3,665,228	678,439	352,557	333,055	639,702	530,192	354,711	299,372	477,200
AR	2,350,725	410,665	246,486	197,424	361,268	328,397	244,096	212,573	349,816
CA	29,755,274	5,436,303	2,722,076	2,904,739	5,738,645	4,645,553	2,955,455	2,231,171	3,121,332
CO	3,293,771	599,278	305,595	282,268	617,333	570,797	340,276	249,924	328,300
CT	3,287,116	517,724	275,158	295,271	588,185	509,760	360,488	294,866	445,664
DE	666,168	113,963	58,980	64,726	119,782	100,110	68,367	59,570	80,670
DC	606,900	80,760	45,404	71,605	122,777	94,984	62,648	51,050	77,672
FL	12,686,788	1,931,088	1,041,486	1,010,156	2,102,614	1,778,994	1,283,728	1,235,820	2,302,902
GA	6,478,847	1,171,969	659,386	623,625	1,182,367	1,014,579	678,987	495,259	652,675
HI	1,108,229	195,278	98,594	104,537	203,466	178,406	109,493	93,778	124,677
ID	1,006,749	207,979	115,708	81,770	154,087	149,338	98,910	77,819	121,138
IL	11,429,942	2,012,780	1,102,499	1,021,458	2,003,217	1,702,509	1,179,345	974,035	1,434,099
IN	5,543,954	975,582	568,654	510,374	919,924	819,577	572,585	481,329	695,929
IA	2,776,442	487,879	271,630	240,359	430,947	397,287	272,959	249,594	425,787
KS	2,474,885	457,755	236,911	216,092	416,003	363,571	234,451	208,146	341,956
KY	3,673,969	626,236	383,356	337,585	610,721	549,204	380,791	320,712	465,364
LA	4,219,973	836,481	458,677	387,821	710,773	606,119	412,186	340,483	467,433
ME	1,226,626	210,082	117,015	104,754	205,713	194,139	123,745	108,198	162,980
MD	4,771,143	812,147	409,957	431,840	901,956	774,414	528,246	395,946	516,637
MA	6,011,978	933,306	506,033	613,116	1,104,645	914,852	605,951	514,398	819,677
MI	9,295,222	1,671,777	930,841	850,016	1,583,364	1,408,199	950,316	793,711	1,106,998
MN	4,370,288	815,963	409,705	377,084	783,562	666,480	428,315	343,315	545,864
MS	2,573,216	495,074	298,599	240,546	403,754	351,197	249,684	213,117	321,245

Figure 8 Population by state and age group, part 1

State		AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	5,113,266	897,590	490,067	436,468	855,640	734,252	524,756	457,095	717,398
MT	799,065	150,406	83,457	57,351	123,913	128,067	81,522	67,930	106,419
NE	1,577,600	294,659	156,790	130,613	259,709	229,478	148,720	134,711	222,920
NV	1,201,833	208,695	100,891	102,609	223,599	192,324	138,893	107,621	127,201
NH	1,109,252	195,970	98,977	100,411	205,815	183,649	111,387	88,059	124,984
NJ	7,730,188	1,217,936	681,960	664,059	1,366,267	1,200,167	850,983	718,589	1,030,227
NM	1,515,069	307,898	160,598	123,983	259,975	229,577	149,712	120,808	162,518
NY	17,990,026	2,891,618	1,615,696	1,664,461	3,148,965	2,720,452	1,944,539	1,642,487	2,361,808
NC	6,628,637	1,074,691	637,603	662,849	1,152,229	1,008,277	705,099	585,832	802,057
ND	637,713	119,767	65,036	57,151	104,833	90,808	56,215	53,132	90,771
OH	10,846,581	1,899,661	1,064,732	957,750	1,805,063	1,619,291	1,115,355	978,701	1,406,028
OK	3,145,585	563,941	318,809	267,411	514,663	452,308	326,770	278,089	423,594
OR	2,842,321	495,834	265,630	225,488	455,371	476,343	297,101	235,423	391,131
PA	11,881,643	1,892,957	1,074,128	1,041,626	1,918,168	1,739,212	1,224,867	1,160,974	1,829,711
RI	1,003,211	155,439	86,271	102,680	174,149	146,571	97,958	89,156	150,987
SC	3,486,703	616,373	363,140	339,600	596,534	526,103	357,747	291,077	396,129
SD	695,133	137,110	71,070	56,976	109,919	96,063	61,962	59,623	102,410
TN	4,896,046	812,832	484,155	452,701	823,042	740,485	530,654	433,773	618,404
TX	16,984,748	3,320,887	1,776,426	1,578,004	3,118,515	2,548,657	1,649,538	1,284,825	1,707,896
UT	1,722,850	430,959	226,933	167,637	275,853	224,715	139,656	107,405	149,692
VT	562,758	99,365	53,099	53,049	95,880	92,804	57,274	45,118	66,169
VA	6,184,493	1,030,088	564,690	616,835	1,147,609	991,563	670,457	500,955	662,296
WA	4,866,692	878,141	444,693	417,468	861,441	804,413	504,238	380,725	575,573
WV	1,792,969	279,885	192,881	148,808	262,961	270,784	191,957	176,960	268,733
WI	4,891,452	887,426	472,270	437,743	825,056	726,753	478,819	412,492	650,893
WY	453,588	92,123	49,716	33,980	75,462	74,182	45,541	35,539	47,045
USA	248,418,140	43,454,102	23,694,112	22,614,049	43,429,692	37,582,954	25,435,905	21,083,554	31,123,772

Figure 9 Population by state and age group, part 2

Different experiments have different age and geographic attributes. See Figure 11 for a list of all 13 experiments identified as A through M. So, Q_{dob} and Z_i, as used above, are representative of several quasi-identifiers that have varying specifications. In experiment B through experiment E, $Z_i \in$ {ZIP codes in USA in which people reside}. In experiment F through experiment I, $Z_i \in$ {Cities, municipalities, towns and recognized post office names in the USA}. Finally, in experiment J through experiment M, $Z_i \in$ {Counties in the USA}. Similarly, in experiments B, F, and J, Q_{dob} = {*date of birth*, *gender*}. In experiments C, G and K, Q_{dob} = {*month and year of birth*, *gender*}. In experiments D, H and L, Q_{dob} = {*year of birth*, *gender*}. Finally, in experiments E, I and M, Q_{dob} = {*2 year age subdivision*, *gender*}.

For completeness, Figure 8 and Figure 9 report the total population per state of each age group. These values are used to compute percentages throughout this document unless otherwise noted.

4.4. Special data elements

This section compares age and year of birth values, as well as, 5-digit ZIP codes, places and counties.

4.4.1. Age versus Year of Birth

Values for an *age* attribute do not necessarily translate to known values for a *year of birth* attribute. There are two cases to consider. If there exists a date to which values for *age* can be referenced, then corresponding values for *year of birth* can be confidently computed. For example, in SID, states calculate the patient's age in years at the time of admission [14]. Because both the computed *age* and the *date of admission* are released, the patient's year of birth can be confidently determined. In experiment D, H and L, I examine *age* as providing a distinct year of birth, and so QI_{SID2} = {*age, gender, 5-digit ZIP*} can be considered as QI_{SID2} = {*year of birth, gender, 5-digit ZIP*}.

On the other hand, if values for *date of admission* were not released, values for *age* would be calendar year specific. In such cases, data are collected with respect to a particular calendar year (that is known) but not a particular day within that year. As a result, each value for *age* corresponds to two possible values for each person's *year of birth*. During any given calendar year, a person reports two ages. The first age occurs before the person's birthday and the second occurs on and after the person's birthday. Because each person's birthday can appear at any time during the calendar year (in contrast to societies in which everyone's "birthday", in terms of determining age, occurs on the same day), two values can be inferred for *year of birth* from a recorded value for *age*. In the experiment E, I and M, I examine {*2 yr age subdivision, gender, 5-digit ZIP*} in which the birth year is within a known 2-year range.

4.4.2. Comparison of 5-digit ZIP codes, Places and Counties

Figure 10 shows a comparison of 5-digit ZIP codes, places and counties in the United States. There are a total of 29,343 ZIP codes, 25,688 places and 3,141 counties. The state having the largest number of counties was Texas (with 254). The District of Columbia had the fewest number of counties (with 1). The average number of counties per state was 62 and the standard deviation was 47.

State	Number 5-digit ZIPs	Number Places	Number Counties	State	Number digit ZIPs	Number Places	Number Counties
AL	567	511	67	MO	993	899	115
AK	195	183	25	MT	315	309	57
AZ	270	178	15	NE	572	518	93
AR	578	563	75	NV	104	66	17
CA	1,515	1,071	58	NH	218	212	10
CO	414	330	63	NJ	540	490	21
CT	263	224	8	NM	276	258	33
DE	53	46	3	NY	1,594	1,369	62
DC	24	2	1	NC	705	624	100
FL	804	463	67	ND	387	384	53
GA	636	561	159	OH	1,007	854	88
HI	80	70	5	OK	586	511	77
ID	244	233	44	OR	384	344	36
IL	1,236	1,147	102	PA	1,458	1,369	67
IN	675	597	92	RI	69	52	5
IA	922	889	99	SC	350	313	46
KS	713	646	105	SD	383	377	66
KY	810	772	120	TN	583	505	95
LA	469	408	64	TX	1,672	1,234	254
ME	410	408	16	UT	205	181	29
MD	419	378	24	VT	243	243	14
MA	473	404	14	VA	820	729	136
MI	875	768	83	WA	484	397	39
MN	877	809	87	WV	655	646	55
MS	363	342	82	WI	714	666	72
				WY	141	135	23
				USA	29,343	25,688	3,141
				max	1,672	1,369	254
				min	24	2	1
				avg	575	504	62
				stdev	401	337	47

Figure 10 Number of 5-digit ZIP codes, Places and Counties by State

5. Results

In the previous sections, I defined terminology and introduced the materials that will be used. In this section, I report on experiments I conducted to estimate the number of unique occurrences for various combinations of demographic attributes that are typically released in publicly and semi-publicly available data.

Experiment A: Uniqueness of {*ZIP*, *gender*, *date of birth*} assume uniform age distribution
Experiment B: Uniqueness of {*ZIP*, *gender*, *date of birth*} based on actual age distribution
Experiment C: Uniqueness of {*ZIP*, *gender*, *month and year of birth*}
Experiment D: Uniqueness of {*ZIP*, *gender*, *age*}
Experiment E: Uniqueness of {*ZIP*, *gender*, *2yr age range*}
Experiment F: Uniqueness of {*place/city*, *gender*, *date of birth*}
Experiment G: Uniqueness of {*place/city*, *gender*, *month and year of birth*}
Experiment H: Uniqueness of {*place/city*, *gender*, *age*}
Experiment I: Uniqueness of {*place/city*, *gender*, *2yr age range*}
Experiment J: Uniqueness of {*county*, *gender*, *date of birth*}
Experiment K: Uniqueness of {*county*, *gender*, *month and year of birth*}
Experiment L: Uniqueness of {*county*, *gender*, *age*}
Experiment M: Uniqueness of {*county*, *gender*, *2yr age range*}

Figure 11 List of 13 experiments

A total of 13 experiments were conducted [15]. These are identified below. Only experiment B, C, D, F and J are briefly reported in this document. Figure 32 contains a summary of results from all 13 experiments.

5.1. Experiment B: Uniqueness of {*ZIP, gender, date of birth*}

Recall, Illinois Research Health Data named ROD provides an example of shared data that contains demographic attributes; in particular, QI_{rod} = {*date of birth*, *gender*, *5-digit ZIP*}. This experiment shows that medical conditions included in these data can be attributed uniquely to one person in most cases.

5.1.1. Experiment B Design

Step 1. Use ZIP table for each of the 50 states and the District of Columbia. Step 2. Figure 12 contains the thresholds for *Q*={*gender*, *date of birth*} specific to each age subdivision. Step 3. Report statistical measurements computed from the table in step 1 using the thresholds determined in step 2. Figure 13 and Figure 14 report the results.

Q = {*gender, date of birth*}		
$\|Q_{AUnder12}\|$	= 2 * 365 * 12	= 8,760
$\|Q_{A12to18}\|$	= 2 * 365 * 7	= 5,110
$\|Q_{A19to24}\|$	= 2 * 365 * 6	= 4,380
$\|Q_{A25to34}\|$	= 2 * 365 * 10	= 7,300
$\|Q_{A35to44}\|$	= 2 * 365 * 10	= 7,300
$\|Q_{A45to54}\|$	= 2 * 365 * 10	= 7,300
$\|Q_{A55to64}\|$	= 2 * 365 * 10	= 7,300
$\|Q_{A65Plus}\|$	= 2 * 365 * 12	= 8,760

Figure 12 Number of possible values for each age subdivision {*gender, date of birth*}

5.1.2. Experiment B Results

Figure 13 and Figure 14 show the results from applying the 3 steps of experiment B to each state, the District of Columbia and the entire United States. The percentages computed for each locale appear in the column named "RANGE %ID_pop." The last row in Figure 14 reports the results of applying the 3 steps of experiment B to all ZIP codes in the United States. As shown, 87.1% of the population of the United States is likely to be uniquely identified by values of {*gender*, *date of birth*, *ZIP*} when age subdivisions are considered.

During the analysis of experiment B, many interesting ZIP codes were found. Here are a few. The ZIP code 11794 in the State of New York is small and extremely homogenous. 4666 of its total population of 5418 (or 86%) are in the age subdivision of 19 to 24. This is the home of the State University of New York at Sony Brook. The ZIP code 10475 in the State of New York reportedly has a larger population of 37077, but people are distributed somewhat evenly across the age subdivisions making the population in each range less than its corresponding threshold. The ZIP code 01701 in the Commonwealth of Massachusetts reportedly has a population of 65,001, which is the largest population for a ZIP code in the state. In experiment A, any person residing in that ZIP code would NOT have been considered likely to be uniquely identified by {*gender*, *date of birth*, *ZIP*}; however, only the subpopulation between the ages of 19 and 44 in

that ZIP code is large enough not to be considered uniquely identified by {*gender*, *date of birth*, *ZIP*}. Persons residing in that ZIP code, who are not in that age subdivision, are less common and considered likely to be uniquely identified by {*gender*, *date of birth*, *ZIP*} even though the population in the entire ZIP code is the largest in the state.

State	#ZIPs	Population	RANGE %population	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	567	4,040,587	99%	100.0%	100.0%	89.7%	98.7%	100.0%	100.0%	100.0%	100.0%
AK	195	544,698	100%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
AZ	270	3,665,228	82%	82.3%	90.1%	67.4%	64.3%	88.8%	100.0%	100.0%	80.7%
AR	578	2,350,725	98%	97.8%	100.0%	87.1%	95.3%	100.0%	100.0%	100.0%	100.0%
CA	1,515	29,755,274	71%	62.4%	73.1%	54.9%	47.2%	70.0%	96.8%	99.6%	96.8%
CO	414	3,293,771	92%	89.7%	96.2%	85.0%	81.1%	92.1%	100.0%	100.0%	100.0%
CT	263	3,287,116	91%	94.3%	98.1%	76.1%	76.2%	88.9%	100.0%	100.0%	97.8%
DE	53	666,168	91%	100.0%	100.0%	72.0%	66.7%	100.0%	100.0%	100.0%	100.0%
DC	24	606,900	64%	62.0%	74.9%	32.5%	47.6%	55.3%	100.0%	84.9%	85.1%
FL	804	12,686,788	91%	93.9%	95.8%	87.5%	85.2%	94.3%	98.6%	99.2%	83.6%
GA	636	6,478,847	90%	90.4%	93.5%	80.4%	77.8%	87.6%	100.0%	100.0%	100.0%
HI	80	1,108,229	74%	62.5%	94.4%	56.7%	55.9%	71.9%	100.0%	100.0%	83.7%
ID	244	1,006,749	99%	100.0%	100.0%	85.6%	100.0%	100.0%	100.0%	100.0%	100.0%
IL	1,236	11,429,942	75%	73.0%	76.4%	59.2%	60.1%	73.9%	90.3%	93.9%	86.7%
IN	675	5,543,954	94%	94.3%	95.2%	80.4%	85.4%	94.7%	100.0%	100.0%	100.0%
IA	922	2,776,442	98%	100.0%	100.0%	78.9%	98.0%	100.0%	100.0%	100.0%	100.0%
KS	713	2,474,885	98%	100.0%	100.0%	83.1%	94.1%	100.0%	100.0%	100.0%	100.0%
KY	810	3,673,969	98%	100.0%	100.0%	85.7%	97.5%	98.6%	100.0%	100.0%	100.0%
LA	469	4,219,973	91%	89.8%	91.7%	80.4%	83.6%	93.0%	100.0%	100.0%	100.0%
ME	410	1,226,626	98%	100.0%	100.0%	86.3%	96.3%	100.0%	100.0%	100.0%	100.0%
MD	419	4,771,143	83%	84.8%	94.1%	79.2%	63.7%	80.2%	93.8%	100.0%	88.7%
MA	473	6,011,978	91%	95.7%	97.9%	73.5%	74.8%	92.8%	100.0%	100.0%	98.8%
MI	875	9,295,222	85%	80.5%	84.7%	72.5%	74.5%	83.2%	98.2%	99.1%	98.3%
MN	877	4,370,288	95%	96.2%	100.0%	81.8%	87.7%	97.4%	100.0%	100.0%	100.0%
MS	363	2,573,216	98%	98.2%	98.1%	88.3%	100.0%	97.8%	100.0%	100.0%	100.0%

Figure 13 Uniqueness of {*ZIP*, *Gender*, *Date of birth*} respecting age distribution, part 1

State	#ZIPs	Population	RANGE %population	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	993	5,113,266	94%	94.4%	98.8%	86.9%	86.8%	92.1%	100.0%	100.0%	97.3%
MT	315	799,065	98%	100.0%	100.0%	78.9%	100.0%	100.0%	100.0%	100.0%	100.0%
NE	572	1,577,600	99%	100.0%	100.0%	90.2%	100.0%	100.0%	100.0%	100.0%	100.0%
NV	104	1,201,833	86%	79.5%	94.3%	79.5%	66.9%	88.3%	94.6%	100.0%	100.0%
NH	218	1,109,252	97%	100.0%	100.0%	94.1%	88.5%	100.0%	100.0%	100.0%	100.0%
NJ	540	7,730,188	92%	92.6%	93.1%	88.0%	79.8%	92.9%	99.1%	100.0%	94.1%
NM	276	1,515,069	88%	86.1%	89.0%	88.6%	71.6%	82.4%	100.0%	100.0%	100.0%
NY	1,594	17,990,026	76%	74.3%	77.3%	64.1%	60.0%	72.1%	88.3%	93.4%	85.5%
NC	705	6,628,637	94%	98.1%	96.4%	77.5%	86.4%	96.5%	98.8%	100.0%	100.0%
ND	387	637,713	96%	100.0%	100.0%	68.5%	91.9%	100.0%	100.0%	100.0%	100.0%
OH	1,007	10,846,581	92%	92.2%	94.7%	82.4%	82.5%	93.6%	100.0%	100.0%	98.5%
OK	586	3,145,585	97%	96.7%	100.0%	85.2%	93.5%	96.7%	100.0%	100.0%	100.0%
OR	384	2,842,321	97%	100.0%	100.0%	89.5%	90.6%	93.1%	100.0%	100.0%	100.0%
PA	1,458	11,881,643	91%	90.5%	94.0%	80.1%	82.2%	90.3%	99.3%	99.4%	94.3%
RI	69	1,003,211	92%	94.4%	100.0%	71.1%	84.2%	94.9%	100.0%	100.0%	94.2%
SC	350	3,486,703	91%	90.0%	95.1%	74.8%	79.5%	95.0%	97.9%	100.0%	100.0%
SD	383	695,133	96%	92.7%	100.0%	81.4%	91.6%	100.0%	100.0%	100.0%	100.0%
TN	583	4,896,046	93%	93.7%	94.8%	80.5%	87.1%	93.5%	100.0%	100.0%	100.0%
TX	1,672	16,984,748	88%	85.0%	89.1%	78.8%	76.5%	90.0%	100.0%	100.0%	100.0%
UT	205	1,722,850	87%	75.8%	80.0%	78.0%	90.2%	92.6%	100.0%	100.0%	100.0%
VT	243	562,758	98%	100.0%	100.0%	80.1%	100.0%	100.0%	100.0%	100.0%	100.0%
VA	820	6,184,493	87%	88.2%	91.6%	71.9%	75.5%	82.7%	97.8%	100.0%	100.0%
WA	484	4,866,692	92%	94.6%	100.0%	82.8%	82.5%	87.2%	100.0%	100.0%	100.0%
WV	655	1,792,969	97%	96.7%	96.4%	90.2%	95.7%	96.4%	100.0%	100.0%	96.5%
WI	714	4,891,452	92%	88.9%	97.7%	77.6%	86.4%	92.6%	100.0%	100.0%	100.0%
WY	141	453,588	98%	100.0%	100.0%	79.2%	100.0%	100.0%	100.0%	100.0%	100.0%
USA	29,343	248,418,140	87%	85.8%	90.2%	75.0%	75.1%	87.0%	97.8%	99.0%	95.3%

Figure 14 Uniqueness of {*ZIP*, *Gender*, *Date of birth*} respecting age distribution, part 2

Figure 15 plots the percentage of the population considered identifiable in each ZIP code in the United States based on experiment B's criteria. The horizontal axis represents the

population that resides in the ZIP code. The vertical axis represents the percentage of the population considered uniquely identified by values of *Q*={*date of birth, gender, 5-digit ZIP*} for a particular ZIP code. The criteria for computing the percentage of the population considered identifiable in experiment B is based on binary decisions, where each decision considers whether a sufficient number of people in a particular age subdivision reside in a particular ZIP code. If so, that sub-population is not considered identifiable; otherwise, its entire sub-population is considered identifiable.

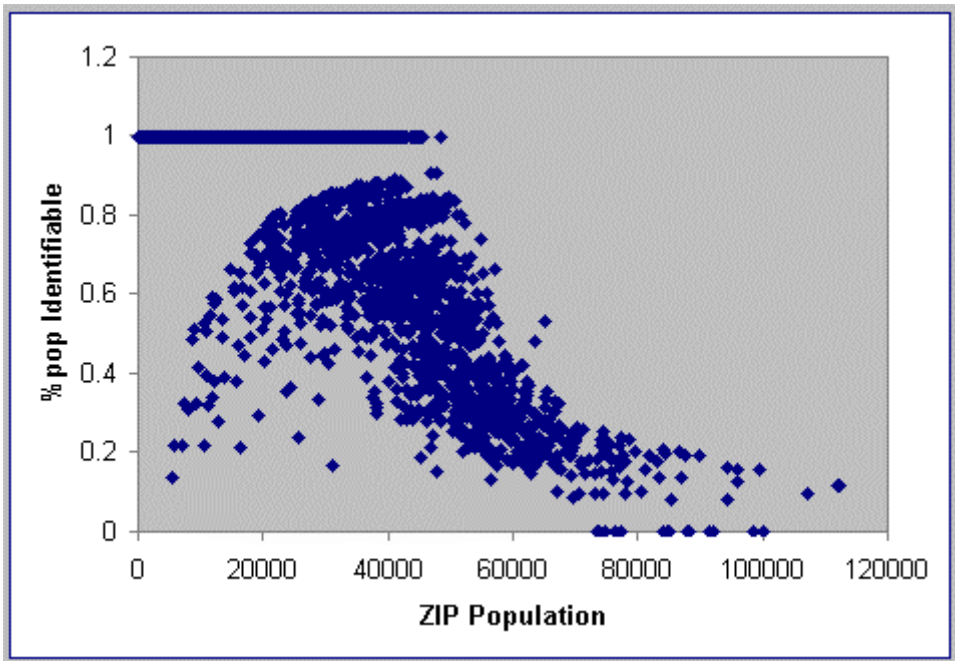

Figure 15 Percentage of Population Identifiable Based on Age subdivisions in ZIP Population

Most ZIP codes (27697 of 29212 or 95%) in the United States that have people listed as residing within them do not have enough people in any age subdivision to consider any such sub-population as identifiable. This is evidenced in Figure 15 by the appearance of dots where the *%pop identifiable* is 1. The largest population having *%pop identifiable* = 1 consists of 48,549 total people. There are very few ZIP codes (15 of 29212) in Figure 15 having sufficient numbers of people in each age subdivision that each such sub-population is not considered uniquely identifiable. This is evidenced in Figure 15 by the appearance of dots where the *%pop identifiable* is 0. The largest population having *%pop identifiable* = 0 has 99,995 people and the smallest has 73,321.

The ZIP code having the largest population, ZIP 60623 with 112,167 people, has a percentage of its population considered identifiable in Figure 15 as being only 11%. It is not 0% because there are insufficient numbers of people above the age of 55 living there despite the large number of people residing in the ZIP code. The point representing this ZIP code in Figure 15 is the rightmost point shown.

The lowest leftmost point shown in Figure 15 corresponds to ZIP 11794, which was discussed earlier. It has a total population of 5418 people and consists primarily of people between the ages of 19 and 24 (4666 of 5418 or 86%). Despite having a small population, the people residing there are very homogenous in terms of age and so the percentage of its population

considered identifiable based on experiment B's criteria is only 13%. It is clear from these examples that population size alone is not an absolute predictor of the identifiability of the people residing within. Care must be taken to model the population as precisely as possible to insure privacy protection.

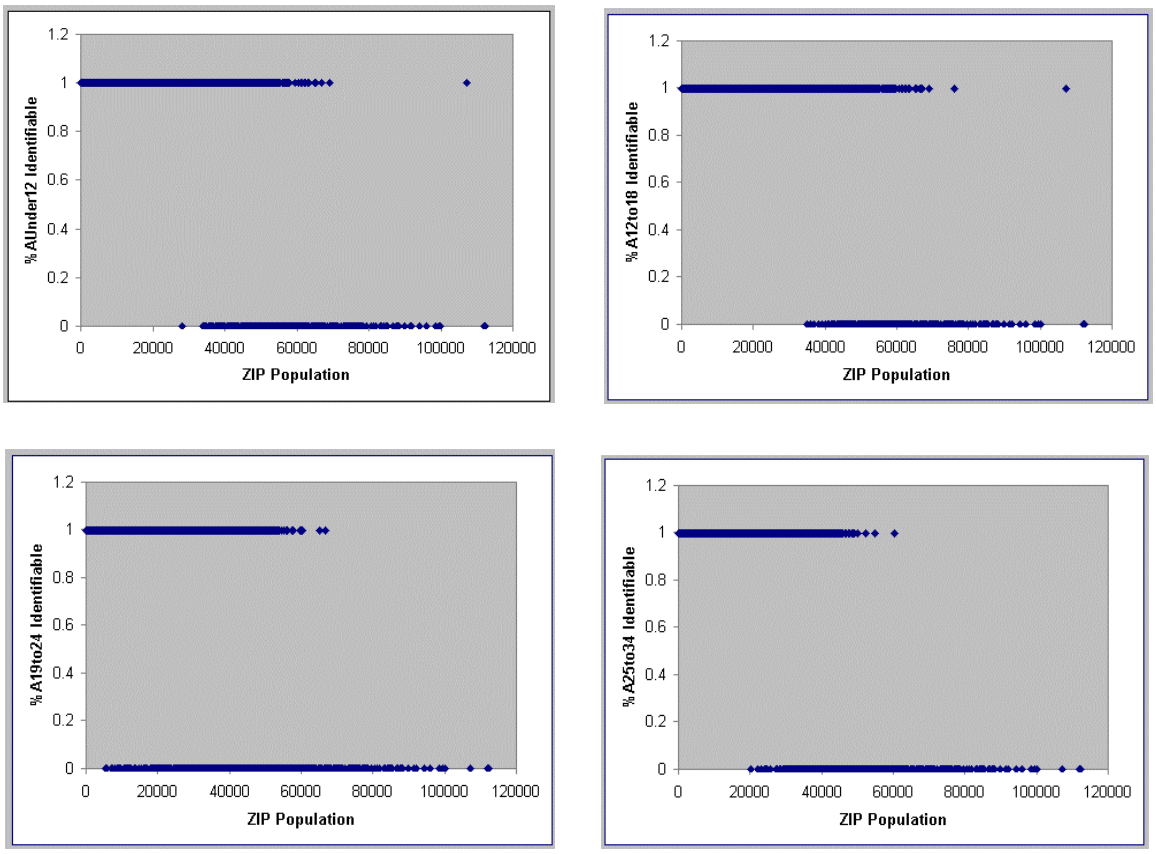

Figure 16 Percentage of Age-based Populations Identifiable within ZIP Population, Part 1

Recall the computation of the percentage of the population considered uniquely identified by values of *Q*={*date of birth*, *gender*, *5-digit ZIP*} for a particular ZIP code in experiment B is based on a composite of binary decisions. Each binary decision concerns the number of people residing within a specific ZIP code in a particular age subdivision. Figure 16 and Figure 17 show plots of the percentage of sub-populations considered identifiable in each ZIP code in the United States based on experiment B's criteria. The horizontal axis represents the population that resides in the ZIP code. The vertical axis represents the percentage of the population considered uniquely identified by values of *Q*={*date of birth*, *gender*, *5-digit ZIP*} for a particular ZIP code and a particular age subdivision. If a sufficient number of people within an age subdivision are reported as residing in a particular ZIP code, then that sub-population is considered identifiable; otherwise, the entire sub-population is not considered identifiable.

Figure 18 provides statistical highlights from the plots in Figure 16 and Figure 17. The topmost table provides statistics on ZIP codes in which the number of people within the noted age subdivision is less than or equal to the threshold for that subdivision. In these cases, the sub-population within the ZIP code is considered uniquely identifiable; that is, *%pop_Identifiable* = 1 for that age subdivision and ZIP code. The bottom table provides statistics in cases where *%pop_Identifiable* < 1. In these ZIP codes, the number of people within the noted age subdivision

is greater than the threshold for that subdivision; therefore, this subdivision is not considered uniquely identifiable.

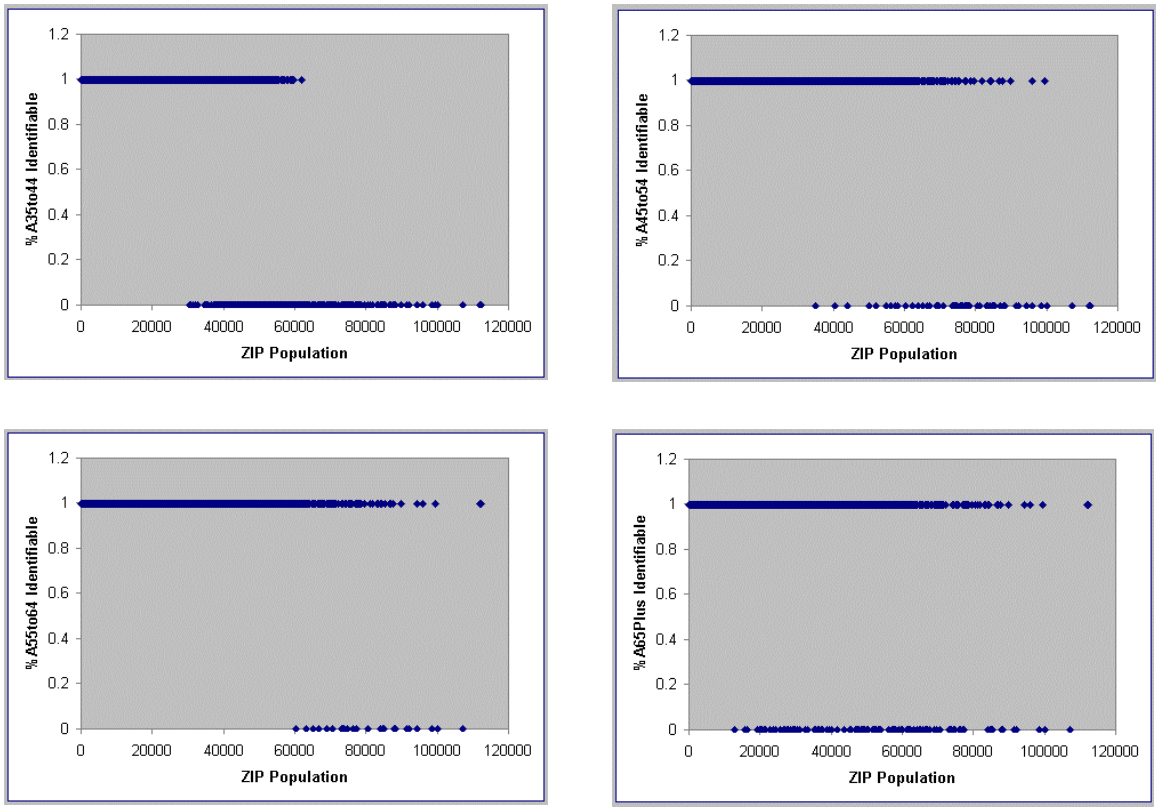

Figure 17 Percentage of Age-based Populations Identifiable within ZIP Population, Part 2

Sub-population considered uniquely identifiable (<= *threshold, IDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP population	107197	107197	66722	60388	62031	99420	112167	112167
Min ZIP population	1	1	1	1	1	1	1	1
Average ZIP population	7615	7873	7332	6911	7596	8358	8442	8311
standard deviation	19452	10915	10070	9227	10393	11938	12165	11956
Number of ZIP codes	28675	28860	28352	28105	28665	29148	29187	29081
Percentage ZIP codes	98.2%	98.8%	97.1%	96.2%	98.1%	99.8%	99.9%	99.6%

Sub-population NOT considered uniquely identifiable (> *threshold, NotIDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP population	112167	112167	112167	112167	112167	112167	107197	107197
Min ZIP population	28294	35092	5418	20211	30577	34860	60388	12890
Average ZIP population	55958	60254	47153	48944	56072	74798	80513	51313
standard deviation	12770	13036	17178	12681	13157	15961	12304	20367
Number of ZIP codes	537	352	860	1107	547	64	25	131
Percentage ZIP codes	1.8%	1.2%	2.9%	3.8%	1.9%	0.2%	0.1%	0.4%

Figure 18 Statistical highlights from Figure 16 and Figure 17

5.2. Experiment C: Uniqueness of {*ZIP, gender, month and year of birth*}

This experiment (referred to as experiment C) is motivated by the Agency for Healthcare Research and Quality's State Inpatient Database (R_{SID}), which is described in part in Figure 3. Among the attributes listed there, I consider QI_{SID1} = {*month and year of birth, gender, 5-digit ZIP*} to be a quasi-identifier within data released by some states. This experiment attempts to characterize the identifiability of QI_{SID1}.

5.2.1. Experiment C Design

Step 1. Use ZIP table for each of the 50 states and the District of Columbia. Step 2. Figure 19 contains the thresholds for *Q*={*gender, month and year of birth*} specific to each age subdivision. Step 3. Report statistical measurements computed from the table in step 1 using the thresholds determined in step 2. Figure 20 and Figure 21 report the results.

Q3 = {*gender, month and year of birth*}		
$\|Q3_{AUnder12}\|$	= 2 * 12 * 12	= 288
$\|Q3_{A12to18}\|$	= 2 * 12 * 7	= 168
$\|Q3_{A19to24}\|$	= 2 * 12 * 6	= 144
$\|Q3_{A25to34}\|$	= 2 * 12 * 10	= 240
$\|Q3_{A35to44}\|$	= 2 * 12 * 10	= 240
$\|Q3_{A45to54}\|$	= 2 * 12 * 10	= 240
$\|Q3_{A55to64}\|$	= 2 * 12 * 10	= 240
$\|Q3_{A65Plus}\|$	= 2 * 12 * 12	= 288

Figure 19 Number of possible values for each age subdivision for {*gender, month and year of birth*}

5.2.2. Experiment C Results

Figure 20 and Figure 21 show the results of applying the 3 steps of experiment C to each state, the District of Columbia (as just reported) and the entire United States. The percentage of people residing in each locale likely to be uniquely identifiable based on {*gender, month and year of birth, ZIP*} appear in the column named "MonYr %ID_pop." For example, 18.1% of the population of Iowa (see Figure 20) and 26.5% of the population of North Dakota (see Figure 21) are likely to be uniquely identifiable based on {*gender, month and year of birth, ZIP*}.

The next to last row in Figure 21 labeled "USA" reports the results of applying the 3 steps of experiment C to all ZIP codes in the United States. As shown, 3.7% of the population of the United States is likely to be uniquely identified by values of {*gender, month and year of birth, ZIP*}. The last row in Figure 21 labeled "%ID_pop" displays the percentage of people in each age subdivision who are likely to be uniquely identified by values of {*gender, month and year of birth, ZIP*}. For example, it reports that 5% of the population of persons residing in the United States between the ages of 45 and 54 are likely to be uniquely identifiable based on {*gender, month and year of birth, ZIP*}.

Figure 22 plots the percentage of the population considered identifiable in each ZIP code in the United States based on experiment C's criteria. The horizontal axis represents the population that resides in the ZIP code. The vertical axis represents the percentage of the

population considered uniquely identified by values of QI_{SID1} = {*month and year of birth*, *gender*, *5-digit ZIP*} for a particular ZIP code. This is the same as the approach used in experiment B.

State	MonYr %ID_pop	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	3.8%	22,253	11,325	10,982	18,197	19,285	22,443	24,806	24,254
AK	10.7%	12,416	6,542	4,826	9,045	7,633	6,253	5,522	6,063
AZ	1.4%	6,804	3,888	4,386	6,786	5,968	7,091	7,095	8,120
AR	11.4%	41,221	23,185	20,274	34,340	35,164	35,248	35,440	42,675
CA	0.8%	33,588	19,440	16,982	27,467	26,335	31,331	33,500	34,743
CO	3.7%	18,174	10,214	8,764	14,721	14,523	16,946	17,965	21,333
CT	1.2%	5,203	2,845	3,097	4,102	3,675	5,104	7,135	7,514
DE	0.9%	867	557	257	653	652	960	715	1,627
DC	0.2%	275	72	26	180	95	66	57	404
FL	0.6%	10,862	6,777	6,548	8,311	9,208	11,647	11,760	13,330
GA	2.7%	19,935	11,272	11,318	18,321	22,193	26,345	31,161	34,905
HI	1.6%	1,767	1,242	1,602	1,911	1,795	2,797	3,645	3,469
ID	8.9%	11,922	7,146	6,950	11,657	11,988	12,404	12,220	15,587
IL	4.4%	75,604	42,727	40,364	62,012	63,393	68,919	70,997	77,971
IN	4.0%	28,592	16,297	17,739	25,328	25,849	33,632	34,730	36,884
IA	18.1%	82,724	44,905	34,644	70,040	64,634	65,878	65,808	72,916
KS	12.1%	46,345	25,207	20,797	36,178	38,319	40,822	41,630	49,544
KY	8.3%	48,404	24,728	23,501	37,727	39,465	41,358	43,680	46,346
LA	2.8%	15,800	8,567	8,553	13,180	13,922	17,090	18,399	22,675
ME	15.5%	29,727	16,098	14,462	23,099	23,470	26,896	26,041	30,713
MD	2.1%	14,087	7,843	8,086	11,105	11,093	13,739	16,099	20,297
MA	1.1%	8,446	5,949	5,540	6,291	6,191	10,006	12,702	12,847
MI	2.4%	27,008	16,914	18,153	22,223	25,106	33,248	37,570	40,591
MN	9.0%	59,128	34,860	28,225	49,369	52,048	54,780	53,583	60,926
MS	4.4%	12,939	7,915	8,487	12,557	14,378	17,937	18,845	20,676

Figure 20 Uniqueness of {*ZIP, Gender, Month and year of birth*} respecting age distribution, part 1

Of the ZIP codes reported in Figure 22, about half (13,871 of 29,212 or 47%) have sufficient numbers of people in each age subdivision so that values of QI_{SID1} = {*month and year of birth*, *gender*, *5-digit ZIP*} are not likely to be uniquely identifying; in these cases, *%pop identifiable* = 0. Values of QI_{SID1} for about one third (9103 of 29212 or 31%) of the ZIP codes are considered uniquely identifying in all age subdivisions; in these cases, *%pop identifiable* = 1. The remaining ZIP codes (6238 of 29212 or 21%) have sub-populations in which values of QI_{SID1} are uniquely identifiable for some age subdivisions but not for others.

Figure 23 provides statistical highlights from the plot in Figure 22. The topmost table provides statistics on ZIP codes in which the number of people within the noted age subdivision is less than or equal to the threshold for that subdivision. In these cases, the sub-population within the ZIP code is considered uniquely identifiable; that is, *%pop_Identifiable* = 1 for that age subdivision and ZIP code. The bottom table provides statistics in cases where *%pop_Identifiable* < 1. In these ZIP codes, the number of people within the noted age subdivision is greater than the threshold for that subdivision; therefore, this subdivision is not considered uniquely identifiable. The method for computing these statistics was described earlier in the Methods section (on page 11).

State	MonYr %ID_pop	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	8.2%	65,966	37,847	31,629	52,566	53,596	57,098	56,566	65,194
MT	15.5%	18,771	11,741	7,717	16,581	16,326	16,280	16,432	19,924
NE	18.2%	46,646	27,556	17,763	38,678	40,574	34,699	37,697	43,232
NV	2.0%	4,320	2,035	1,983	3,341	2,977	2,516	2,705	4,256
NH	7.5%	11,934	7,545	6,001	8,773	7,859	12,067	13,156	15,851
NJ	0.6%	6,760	4,693	3,510	3,811	4,642	5,846	8,238	8,142
NM	5.2%	11,169	6,307	5,208	10,048	10,235	10,844	11,340	14,141
NY	2.3%	54,792	33,243	31,443	45,160	49,560	61,882	68,223	76,979
NC	2.4%	22,064	11,906	10,595	17,177	16,987	23,559	27,726	31,714
ND	26.5%	28,362	16,090	9,492	22,535	22,563	20,666	22,226	27,314
OH	2.2%	28,645	14,449	18,930	24,301	24,283	37,395	43,814	47,838
OK	7.1%	32,749	20,178	16,901	26,174	29,484	29,507	32,320	35,238
OR	4.2%	18,614	9,286	8,839	15,741	14,495	15,766	15,778	19,684
PA	3.5%	58,144	32,516	32,758	47,305	45,996	62,507	66,894	75,584
RI	0.9%	1,085	642	500	764	1,417	1,025	1,487	1,996
SC	2.3%	9,342	5,171	5,813	8,643	8,309	12,372	13,670	16,738
SD	25.9%	27,699	17,147	11,054	25,496	24,375	22,171	23,721	28,405
TN	3.4%	24,172	12,553	13,053	18,105	19,074	22,832	25,898	30,553
TX	2.3%	51,615	29,794	30,883	45,082	50,060	58,173	62,784	68,838
UT	3.4%	8,496	4,844	4,042	7,026	7,447	8,832	8,293	10,307
VT	21.9%	19,797	11,196	8,334	16,536	17,312	16,075	16,093	18,066
VA	4.4%	41,345	23,241	20,634	30,706	33,035	35,263	40,117	47,007
WA	2.6%	18,736	11,083	9,104	14,925	15,043	17,563	19,665	21,650
WV	15.5%	43,535	25,866	21,381	36,753	37,676	34,584	35,731	42,582
WI	5.4%	32,406	21,664	21,855	31,257	30,297	40,576	43,567	44,714
WY	10.1%	8,492	3,943	2,743	6,058	5,943	6,251	5,893	6,684
USA	**3.7%**	1,329,747	759,051	676,728	1,098,342	1,125,947	1,269,289	1,351,139	1,529,041
%ID_pop		3.1%	3.2%	3.0%	2.5%	3.0%	5.0%	6.4%	4.9%

Figure 21 Uniqueness of {*ZIP, Gender, Month and year of birth*} respecting age distribution, part 2

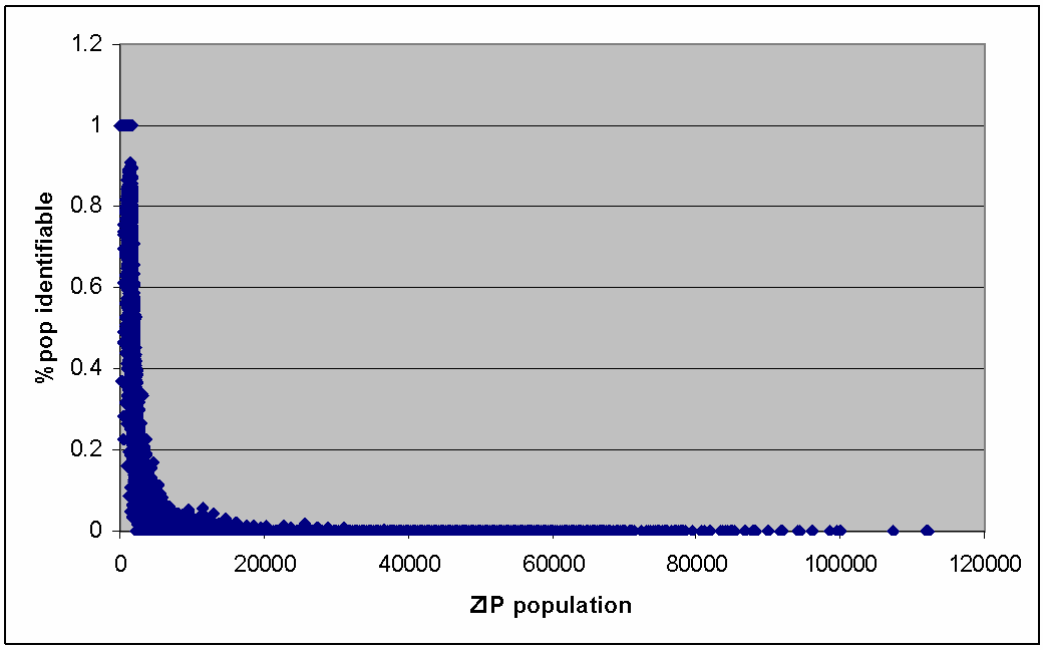

Figure 22 Percentage of Population Identifiable Based on Uniform Distribution of Ages in ZIP Population

The values reported as ZIP populations in Figure 22 are not the total number of people within the reported age subdivision residing in those ZIP codes but are just the numbers of people

residing in the ZIP code. For example, consider the values appearing in the "Aunder12" column in Figure 22. They report information about children under the age of 12 residing in 10,852 ZIP codes in the United States that had insufficient numbers of children to render corresponding values of QI_{SID1} = {*month and year of birth, gender, 5-digit ZIP*} uniquely identifiable. Of these ZIP codes, the largest number of children of under the age of 12, residing in a ZIP code was 287. Some ZIP codes, who had people residing within them, had no children in this age. The average number of children in these ZIP codes was 123 with a standard deviation of 80.

Sub-population considered uniquely identifiable (<= *threshold, IDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP sub-population	287	167	143	239	239	239	239	287
Min ZIP sub-population	0	0	0	0	0	0	0	0
Average ZIP sub-population	123	71	53	101	102	96	95	118
standard deviation	80	47	40	66	66	66	66	80
Number of ZIP codes	10852	10725	12760	10883	11045	13202	14220	12905
Percentage ZIP codes	37.1%	36.7%	43.7%	37.3%	37.8%	45.2%	48.7%	44.2%

Sub-population NOT considered uniquely identifiable (> *threshold, NotIDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP sub-population	26914	15352	27123	24587	19543	15544	12205	25799
Min ZIP sub-population	288	168	144	240	240	240	240	288
Average ZIP sub-population	2294	1241	1333	2309	2007	1509	1316	1815
standard deviation	2530	1327	1690	2632	2096	1419	1174	1860
Number of ZIP codes	18360	18487	16452	18329	18167	16010	14992	16307
Percentage ZIP codes	62.9%	63.3%	56.3%	62.7%	62.2%	54.8%	51.3%	55.8%

Figure 23 Statistical highlights from Figure 20 and Figure 21

5.3. Experiment D: Uniqueness of {*ZIP, gender, age*}

In this experiment, I examine the identifiability of {*year of birth, gender, 5-digit ZIP*} in the United States. Progressing through the results from the last three experiments, values referring to age became less specific and as expected, the values became less uniquely identifying. What may be surprising however is that these values remained uniquely identifying for some people.

The Agency for Healthcare Research and Quality's State Inpatient Database (SID; see Figure 3) motivated this experiment as well as experiment C. In addition to QI_{SID1} used in experiment C, SID also includes QI_{SID2} = {*age, gender, 5-digit ZIP*} for some states in those data. Recall in section 4.4.1, I examine *age* as providing a distinct year of birth, and so QI_{SID2} = {*age, gender, 5-digit ZIP*} can be considered as QI_{SID2} = {*year of birth, gender, 5-digit ZIP*}.

5.3.1. Experiment D Design

Step 1. Use ZIP table for each of the 50 states and the District of Columbia. Step 2. Figure 24 contains the thresholds for Q={*gender, date of birth*} specific to each age subdivision. Step 3. Report statistical measurements computed from the table in step 1 using the thresholds determined in step 2. Figure 25 and Figure 26 report the results.

Q4 = {*gender, year of birth*}		
$\|Q4_{AUnder12}\|$	= 2 * 12	= 24
$\|Q4_{A12to18}\|$	= 2 * 7	= 14
$\|Q4_{A19to24}\|$	= 2 * 6	= 12
$\|Q4_{A25to34}\|$	= 2 * 10	= 20
$\|Q4_{A35to44}\|$	= 2 * 10	= 20
$\|Q4_{A45to54}\|$	= 2 * 10	= 20
$\|Q4_{A55to64}\|$	= 2 * 10	= 20
$\|Q4_{A65Plus}\|$	= 2 * 12	= 24

Figure 24 Number of possible values for each age subdivision for {*gender, year of birth*}

5.3.2. Experiment D Results

Figure 25 and Figure 26 show the results of applying the 3 steps of experiment D to each state, the District of Columbia (as just reported) and the entire United States. The percentage of people residing in each locale likely to be uniquely identifiable based on {*gender, year of birth, ZIP*} appears in the column named “BirthYr %ID_pop” and the number of people represented by the percentage appears in the column named "BirthYr #ID_pop". For example, 0.89% (or 5703 people) of the population of Iowa (see Figure 26) are likely to be uniquely identifiable by values of {*gender, year of birth, ZIP*}.

State	BirthYr %ID_pop	BirthYr Total	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	0.02%	918	105	53	89	97	112	125	158	179
AK	0.70%	3,809	227	223	227	223	315	631	804	1,159
AZ	0.02%	638	68	31	23	53	98	98	96	171
AR	0.09%	2,121	452	138	264	208	248	312	349	150
CA	0.01%	4,229	541	319	362	461	336	540	678	992
CO	0.08%	2,752	287	224	346	336	201	447	426	485
CT	0.01%	474	69	55	36	52	30	108	63	61
DE	0.02%	158	18	13	21	28	36	5	10	27
DC	0.01%	46	6	-	-	-	-	-	16	24
FL	0.00%	512	76	63	9	5	43	90	121	105
GA	0.01%	780	83	29	91	101	56	120	182	118
HI	0.01%	165	28	11	9	33	42	12	20	10
ID	0.19%	1,943	259	148	205	255	258	310	248	260
IL	0.01%	1,401	167	111	148	141	123	246	255	210
IN	0.01%	746	82	27	54	88	84	89	131	191
IA	0.11%	3,106	278	305	647	182	249	583	535	327
KS	0.22%	5,482	575	446	924	571	594	1,017	750	605
KY	0.13%	4,722	671	309	280	528	448	697	966	823
LA	0.02%	870	118	48	75	118	84	135	169	123
ME	0.19%	2,296	293	217	190	287	228	280	331	470
MD	0.03%	1,275	152	119	96	156	179	187	194	192
MA	0.01%	499	83	50	51	35	25	58	100	97
MI	0.01%	920	124	133	134	151	71	133	120	54
MN	0.06%	2,709	365	214	439	421	265	326	335	344
MS	0.02%	462	54	23	21	39	26	57	136	106

Figure 25 Uniqueness of {*ZIP, Gender, Year of birth*} respecting age distribution, part 1

The next to last row in Figure 26 labeled "USA" reports the results of applying the 3 steps of experiment D to all ZIP codes in the United States. As shown, 0.04% (or 105,016 people) of the population of the United States is likely to be uniquely identified by values of {*gender, year of birth, ZIP*}. The last row in Figure 26 labeled "%ID_pop" displays the percentage of people in each age subdivision who are likely to be uniquely identified by values of {*gender, year of birth, ZIP*}. For example, it reports that 0.08% of the population of persons residing in the

United States between the ages of 55 and 64 are likely to be uniquely identified by values of {*gender, year of birth, ZIP*}.

State	BirthYr %ID_pop	BirthYr Total	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	0.07%	3,403	451	320	402	312	371	549	531	467
MT	0.43%	3,465	399	263	405	433	362	534	492	577
NE	0.23%	3,560	241	241	717	325	387	676	455	518
NV	0.04%	439	77	35	39	47	47	62	57	75
NH	0.07%	777	154	62	106	56	81	111	100	107
NJ	0.01%	728	125	62	41	61	51	96	114	178
NM	0.22%	3,302	343	276	237	395	350	569	644	488
NY	0.03%	5,460	714	469	533	720	445	804	818	957
NC	0.02%	1,032	133	94	74	134	103	177	168	149
ND	0.89%	5,703	586	476	832	675	639	932	787	776
OH	0.00%	377	34	25	30	37	33	38	96	84
OK	0.06%	1,963	220	135	248	219	274	336	237	294
OR	0.07%	1,900	369	140	172	258	124	214	315	308
PA	0.03%	3,099	501	201	324	413	348	429	440	443
RI	0.01%	92	-	-	-	9	10	30	19	24
SC	0.01%	443	87	16	41	66	63	85	45	40
SD	0.63%	4,408	489	291	607	544	516	632	597	732
TN	0.02%	836	201	14	125	70	53	165	128	80
TX	0.03%	5,483	815	383	443	641	661	717	794	1,029
UT	0.08%	1,323	78	59	146	189	151	230	230	240
VT	0.20%	1,117	76	63	171	54	81	166	150	356
VA	0.06%	3,754	572	286	350	423	445	483	638	557
WA	0.03%	1,227	164	85	145	138	122	142	220	211
WV	0.30%	5,360	746	316	433	614	605	874	869	903
WI	0.02%	881	80	101	135	130	79	103	103	150
WY	0.41%	1,851	213	157	232	165	195	361	223	305
USA	**0.04%**	105,016	13,049	7,879	11,729	11,697	10,747	16,121	16,463	17,331
%ID_pop			0.03%	0.03%	0.05%	0.03%	0.03%	0.06%	0.08%	0.06%

Figure 26 Uniqueness of {*ZIP, Gender, Year of birth*} respecting age distribution, part 2

Most ZIP codes (25,705 of 29,212 or 88%) have sufficient numbers of people in each age subdivision so that values of QI_{SID2} = {*year of birth, gender, 5-digit ZIP*} are not likely to be uniquely identifying; in these cases, *%pop identifiable* = 0. Values of QI_{SID2} for about one third (353 of 29212 or 1%) of the ZIP codes are considered uniquely identifying in all age subdivisions; in these cases, *%pop identifiable* = 1. The remaining ZIP codes (3154 of 29212 or 11%) have sub-populations in which values of QI_{SID2} are uniquely identifiable for some age subdivisions but not for all.

Figure 27 provides statistical highlights. The topmost table provides statistics on ZIP codes in which the number of people within the noted age subdivision is less than or equal to the threshold for that subdivision. In these cases, the sub-population within the ZIP code is considered uniquely identifiable; that is, *%pop_Identifiable* = 1 for that age subdivision and ZIP code. The bottom table provides statistics in cases where *%pop_Identifiable* < 1. In these ZIP codes, the number of people within the noted age subdivision is greater than the threshold for that subdivision; therefore, this subdivision is not considered uniquely identifiable.

Sub-population considered uniquely identifiable (<= *threshold, IDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP sub-population	24	14	12	20	20	20	20	24
Min ZIP sub-population	0	0	0	0	0	0	0	0
Average ZIP sub-population	11	6	5	10	9	10	9	11
standard deviation	8	5	4	7	7	7	7	8
Number of ZIP codes	1200	1342	2309	1210	1150	1651	1798	1584
Percentage ZIP codes	4.1%	4.6%	7.9%	4.1%	3.9%	5.7%	6.2%	5.4%

Sub-population NOT considered uniquely identifiable (> *threshold, NotIDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP sub-population	26914	15352	27123	24587	19543	15544	12205	25799
Min ZIP sub-population	25	15	13	21	21	21	21	25
Average ZIP sub-population	1551	850	840	1551	1339	922	768	1126
standard deviation	2291	1212	1460	2372	1914	1284	1057	1652
Number of ZIP codes	28012	27870	26903	28002	28062	27561	27414	27628
Percentage ZIP codes	95.9%	95.4%	92.1%	95.9%	96.1%	94.3%	93.8%	94.6%

Figure 27 Statistical highlights from Figure 25 and Figure 26

5.4. Experiment F: Uniqueness of {*place*/*city, gender, date of birth*}

This experiment examines the identifiability of {*date of birth, gender, place*}. While the number of places is expected to be less than the number of ZIP codes, the difference is not as dramatic as one would expect.

State	DOB %ID_pop	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	74.31%	510,294	316,271	246,921	455,646	425,871	340,085	290,787	416,713
AK	67.62%	86,943	36,668	30,801	72,365	66,328	34,744	18,336	22,157
AZ	30.18%	207,821	117,371	79,857	154,789	150,173	121,318	116,660	158,254
AR	85.73%	355,634	221,013	144,471	278,355	286,099	217,119	197,637	314,833
CA	35.99%	1,705,016	1,032,675	785,915	1,266,384	1,411,260	1,494,618	1,350,466	1,663,710
CO	40.41%	221,248	124,459	100,826	189,908	196,192	164,375	145,456	188,554
CT	66.44%	355,973	208,871	144,966	296,959	320,087	299,897	249,481	307,714
DE	68.04%	78,966	40,675	32,116	63,018	69,766	49,625	52,013	67,054
DC	0.00%	-	-	26	-	-	-	-	-
FL	44.12%	866,146	523,124	416,970	743,419	783,719	697,379	690,365	875,685
GA	62.62%	737,096	425,884	331,861	601,348	569,614	501,763	393,910	495,444
HI	49.94%	89,975	69,406	41,139	80,566	82,216	68,636	55,413	66,056
ID	76.93%	147,599	93,691	50,482	116,729	113,067	83,114	66,524	103,285
IL	60.16%	1,205,138	698,921	490,199	976,815	965,017	842,088	731,266	966,748
IN	63.45%	610,004	362,468	272,124	485,926	499,979	431,504	366,413	488,978
IA	77.50%	375,417	218,025	141,276	310,173	302,724	238,696	219,669	345,691
KS	66.77%	295,043	167,547	111,512	236,104	229,189	182,750	160,132	270,086
KY	78.76%	513,045	319,232	234,139	451,331	419,197	325,073	277,950	369,257
LA	58.86%	474,999	271,968	196,903	380,395	336,651	278,656	233,811	310,514
ME	94.22%	201,167	117,015	82,913	184,342	184,857	123,745	108,198	153,502
MD	63.22%	542,516	299,174	256,363	432,696	456,506	379,792	307,456	341,639
MA	73.33%	738,432	409,915	351,483	610,144	673,586	526,058	440,426	658,804
MI	56.68%	912,385	535,570	393,345	760,515	737,677	656,494	551,937	720,202
MN	71.55%	582,951	327,576	213,712	462,644	439,233	358,955	299,529	442,243
MS	81.12%	386,515	232,392	164,750	307,447	278,994	231,718	197,189	288,516

Figure 28 Uniqueness of {*Place, Gender, Date of birth*} respecting age distribution, part 1

Step 1. Use ZIP table for each of the 50 states and the District of Columbia. Step 2. Figure 12 contains the thresholds for Q={*gender, date of birth*} specific to each age subdivision. Step 3. Report statistical measurements computed from the table in step 1 using the thresholds

determined in step 2. Figure 28 and Figure 29 report the results of applying the 3 steps of experiment F to each state, the District of Columbia and the entire United States.

The percentage of people residing in each locale likely to be uniquely identifiable by values of {*gender, date of birth, place*} appear in the column named "DOB %ID_pop." For example, 94.22% of the population of Maine (see Figure 28) and 74.99% of the population of Pennsylvania (see Figure 29) are likely to be uniquely identifiable by values of {*gender, date of birth, place*}. Vermont had the largest percentage of its population identifiable (98.12%). The District of Columbia had 0% identified. The state having the smallest percentage was Nevada with 26.48%. The average was 64.54% and the standard deviation was 17.88%.

State	DOB %ID_pop	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	65.98%	575,534	345,340	253,443	490,825	454,370	395,626	347,346	509,243
MT	78.05%	111,323	63,624	30,390	86,536	94,856	73,526	67,930	95,497
NE	60.86%	173,370	100,557	63,607	137,330	136,238	98,945	92,145	157,885
NV	26.48%	48,890	29,379	17,274	44,040	48,251	49,077	36,910	44,428
NH	83.26%	164,556	84,043	75,108	158,945	156,196	94,268	79,579	110,913
NJ	75.46%	916,586	513,909	459,760	887,738	910,504	705,604	615,918	823,232
NM	58.82%	185,741	103,241	70,980	125,794	127,320	94,373	78,403	105,250
NY	50.89%	1,510,307	893,370	734,124	1,331,293	1,394,790	1,103,058	955,471	1,231,836
NC	66.99%	748,655	434,802	352,507	670,230	637,726	523,682	455,492	617,381
ND	89.24%	108,831	59,803	33,455	83,627	83,251	56,215	53,132	90,771
OH	65.65%	1,218,515	726,779	536,583	1,009,900	1,059,754	865,805	737,419	965,782
OK	64.24%	349,375	209,852	141,980	280,350	266,557	233,933	212,063	326,461
OR	64.29%	318,531	186,694	120,253	251,227	266,919	224,214	180,088	279,439
PA	74.99%	1,427,475	829,811	674,412	1,324,556	1,288,682	1,002,535	960,527	1,401,861
RI	55.57%	83,379	52,128	46,137	74,615	83,775	73,597	65,732	78,157
SC	67.65%	404,179	259,598	178,853	347,400	357,955	263,798	240,827	306,073
SD	81.02%	108,221	62,338	36,113	80,508	80,733	53,059	51,721	90,508
TN	64.98%	529,152	319,932	243,251	474,021	459,452	388,946	320,903	433,014
TX	44.27%	1,410,090	792,176	561,715	1,100,437	1,053,590	840,761	735,749	1,025,466
UT	56.43%	208,964	117,137	81,156	132,730	134,699	106,448	84,198	106,867
VT	98.12%	99,365	53,099	42,494	95,880	92,804	57,274	45,118	66,169
VA	58.50%	588,706	358,361	294,519	565,454	531,480	468,056	357,966	453,394
WA	53.56%	458,232	257,086	168,811	372,536	382,178	319,725	272,740	375,511
WV	90.95%	260,338	178,947	125,468	232,443	242,711	184,384	169,168	237,233
WI	68.27%	584,155	333,763	235,969	497,263	483,528	372,939	334,139	497,585
WY	79.05%	67,039	36,679	20,714	52,859	53,145	45,541	35,539	47,045
USA	**58.38%**	24,859,832	14,572,359	10,914,146	20,826,555	20,879,466	17,343,591	15,107,247	20,512,640
%ID_pop		57.2%	61.5%	48.3%	48.0%	55.6%	68.2%	71.7%	65.9%

Figure 29 Uniqueness of {*Place, Gender, Date of birth*} respecting age distribution, part 2

The next to last row in Figure 29 labeled "USA" reports the results of applying the 3 steps of experiment F to all places in the United States. As shown, 58.38% of the population of the United States is likely to be uniquely identified by values of {*gender, date of birth, place*}. The last row in Figure 29 labeled "%ID_pop" displays the percentage of people in each age subdivision who are likely to be uniquely identified by values of {*gender, date of birth, place*}. For example, it reports that 71.7% of the population of persons residing in the United States between the ages of 55 and 64 are likely to be uniquely identifiable based on {*gender, date of birth, place*}.

The place having the largest population was Chicago, Illinois, with 2,451,767 people. The place having the smallest population was Crooked Creek, Alaska that reports only one person of age 65 or more resides there. The average population for a place is 9,710 and the standard deviation is 44,149. There are a total of 25,585 places.

5.5. Experiment J: Uniqueness of {*county, gender, date of birth*}

This experiment examines the identifiability of {*date of birth*, *gender*, *county*}. Recall, there are a total of 29,343 ZIP codes, 25,688 places and 3,141 counties.

Step 1. Use ZIP table for each of the 50 states and the District of Columbia. Step 2. Figure 12 contains the thresholds for *Q*={*gender*, *date of birth*} specific to each age subdivision. Step 3. Report statistical measurements computed from the table in step 1 using the thresholds determined in step 2. Figure 30 and Figure 31 report the results of applying the 3 steps of experiment J to each state, the District of Columbia and the entire United States.

The percentage of people residing in each locale likely to be uniquely identifiable by values of {*gender*, *date of birth*, *county*} appear in the column named "DOB %ID_pop." For example, 58% of the population of Mississippi (see Figure 30) and 52% of the population of Nebraska (see Figure 31) are likely to be uniquely identifiable by values of {*gender*, *date of birth*, *county*}. Wyoming had the largest percentage of its population identifiable (75%). Connecticut, Delaware, the District of Columbia and New Jersey had 0% identified. The average was 28% and the standard deviation was 22%.

The next to last row in Figure 31 labeled "USA" reports the results of applying the 3 steps of experiment J to all counties in the United States. As shown, 18.1% of the population of the United States is likely to be uniquely identified by values of {*gender*, *date of birth*, *county*}. The last row in Figure 31 labeled "%ID_pop" displays the percentage of people in each age subdivision who are likely to be uniquely identified by values of {*gender, date of birth*, *county*}. For example, it reports that 25.84% of the population of persons residing in the United States between the ages of 55 and 64 are likely to be uniquely identifiable based on {*gender, date of birth*, *county*}.

State	DOB %ID_pop	DOB Total	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
AL	31%	1,239,261	203,418	119,887	100,859	147,718	149,568	145,639	165,943	206,229
AK	43%	231,537	39,695	25,282	18,424	47,769	29,382	31,533	17,293	22,159
AZ	5%	168,352	23,995	13,659	8,351	15,873	23,248	23,557	26,589	33,080
AR	55%	1,286,703	204,611	126,862	85,675	165,982	171,952	153,203	149,635	228,783
CA	2%	482,182	74,362	42,716	33,536	62,826	50,565	61,321	70,890	85,966
CO	16%	530,181	94,650	50,001	38,332	85,809	86,915	60,219	46,794	67,461
CT	0%	-	-	-	-	-	-	-	-	-
DE	0%	-	-	-	-	-	-	-	-	-
DC	0%	-	-	-	-	-	-	-	-	-
FL	5%	680,438	109,084	74,526	59,719	96,106	93,589	80,169	64,489	102,756
GA	36%	2,335,158	385,475	236,121	182,875	311,416	297,509	294,860	257,992	368,910
HI	2%	24,302	-	4,985	3,356	-	20	5,039	4,127	6,775
ID	50%	504,176	84,045	54,338	27,716	64,270	70,098	61,874	63,762	78,073
IL	15%	1,733,651	294,307	164,151	119,585	237,212	225,134	210,334	189,098	293,830
IN	33%	1,805,518	310,118	183,259	129,393	268,623	228,630	204,738	205,590	275,167
IA	57%	1,574,848	267,585	153,138	102,462	208,798	216,811	168,181	161,950	295,923
KS	45%	1,117,968	187,792	105,602	71,548	150,530	142,864	128,522	120,241	210,869
KY	55%	2,015,672	339,649	199,166	162,837	287,814	264,521	239,055	215,956	306,674
LA	26%	1,103,759	166,000	99,616	76,791	129,159	151,255	132,897	154,279	193,762
ME	24%	289,549	45,914	25,233	25,821	33,214	34,481	37,839	34,151	52,896
MD	6%	288,043	36,084	19,602	20,508	36,667	32,605	29,983	51,224	61,370
MA	1%	30,080	2,997	1,179	914	2,739	3,651	8,515	7,345	2,740
MI	14%	1,270,356	187,954	101,271	84,202	147,645	144,700	167,106	186,504	250,974
MN	35%	1,545,738	264,233	169,559	98,392	209,122	217,857	169,995	152,615	263,965
MS	58%	1,503,027	258,287	160,447	109,981	202,539	185,987	179,021	161,836	244,929

Figure 30 Uniqueness of {*County, Gender, Date of birth*} respecting age distribution, part 1

The county having the largest population was Los Angeles County in California, with 8,863,164 people. The county having the smallest population was Yellowstone County in

Montana where only 52 people reside. The average population for a county is 79,182 and the standard deviation is 263,813. There are a total of 3,141 counties.

State	DOB %ID_pop	DOB Total	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
MO	35%	1,777,250	299,822	177,231	117,495	235,542	226,741	199,449	201,418	319,552
MT	58%	461,847	79,868	50,920	26,385	52,524	57,155	54,121	58,303	82,571
NE	52%	827,590	148,680	81,858	48,439	114,003	115,534	80,505	86,132	152,439
NV	17%	205,707	39,060	19,336	13,922	34,721	32,791	23,853	18,433	23,591
NH	14%	158,917	18,460	14,248	10,218	17,034	17,684	29,898	27,505	23,870
NJ	0%	13,203	-	0	-	-	-	7,089	6,114	-
NM	31%	466,942	55,935	41,285	37,298	48,782	59,912	68,042	65,578	90,110
NY	4%	714,072	86,136	44,198	48,241	39,884	79,329	130,425	139,808	146,051
NC	26%	1,718,318	242,446	149,044	126,104	202,459	209,646	232,090	249,455	307,074
ND	63%	401,471	65,977	37,393	19,272	49,281	47,612	47,773	53,295	80,868
OH	14%	1,536,542	244,518	135,966	102,380	185,611	200,338	199,829	190,210	277,690
OK	44%	1,395,889	214,447	135,344	106,030	180,771	170,655	171,464	164,825	252,353
OR	16%	468,933	58,089	37,189	25,490	51,118	50,034	77,348	76,465	93,200
PA	7%	868,774	143,074	81,634	65,841	120,867	115,691	110,104	109,679	121,884
RI	4%	36,592	7,442	4,146	-	-	7,157	5,220	4,929	7,698
SC	23%	792,897	115,127	71,978	53,250	100,618	97,592	104,465	111,669	138,198
SD	73%	506,465	96,431	52,085	32,146	70,960	65,236	51,201	50,256	88,150
TN	37%	1,832,875	296,158	180,822	134,829	239,766	227,196	226,279	223,498	304,327
TX	19%	3,185,236	555,868	314,582	220,496	408,489	396,535	357,970	372,889	558,407
UT	17%	296,513	58,729	33,397	21,901	43,107	40,697	30,917	26,743	41,022
VT	59%	329,450	48,194	35,514	24,136	42,551	42,821	44,422	36,277	55,535
VA	35%	2,186,920	327,643	195,729	180,037	286,163	280,550	300,469	262,255	354,074
WA	11%	523,874	66,444	57,010	36,219	58,899	62,605	75,825	83,264	83,608
WV	59%	1,059,753	168,623	100,661	72,214	144,775	151,174	140,705	128,935	152,666
WI	25%	1,211,247	190,779	110,977	71,189	162,807	159,047	141,883	157,036	217,529
WY	75%	338,752	57,064	36,055	20,545	52,035	52,251	38,218	35,539	47,045
USA	**18.1%**	45,076,528	7,265,269	4,329,202	3,175,354	5,854,598	5,787,325	5,543,164	5,448,813	7,672,803
%ID_pop			16.72%	18.27%	14.04%	13.48%	15.40%	21.79%	25.84%	24.65%

Figure 31 Uniqueness of {*County, Gender, Date of birth*} respecting age distribution, part 2

6. Discussion

Figure 32 contains a summary of the results reported in the previous section. A description of each reported percentage is provided in the following paragraphs. These percentages demonstrate how combinations of characteristics can combine to narrow the number of possible people under consideration as the subject of de-identified person-specific data.

County	18.1	0.04	0.00004	0.00000*
Place	58.4	3.6	0.04	0.01
ZIP	87.1	3.7	0.04	0.01
	DOB	**Mon/Year**	**BirthYear**	**2yr Age**

Figure 32 Percentage of US population identified with gender as geography and age vary

Experiment B reported that 87.1% (216 million of 248 million) of the population in the United States had characteristics that were likely made them unique based only on {*5-digit ZIP*, *gender*, *date of birth*}. Experiment C reported that 3.7% of the population in the United States had characteristics that were likely made them unique based only on {*5-digit ZIP*, *gender*, *Month and year of birth*}. Experiment D reported that 0.04% of the population in the United States had characteristics that were likely made them unique based only on {*5-digit ZIP*, *gender*, Year *of birth*}. Experiment E reported that 0.01% of the population in the United States had characteristics that were likely made them unique based only on {*5-digit ZIP*, *gender*, *2year age range*}.

Experiment F reported that 58.4% of the population in the United States had characteristics that were likely made them unique based only on {*Place, gender, date of birth*}. Experiment G reported that 3.6% of the population in the United States had characteristics that were likely made them unique based only on {*Place, gender, Month and year of birth*}. Experiment H reported that 0.04% of the population in the United States had characteristics that were likely made them unique based only on {*Place, gender, Year of birth*}. Experiment I reported that 0.01% of the population in the United States had characteristics that were likely made them unique based only on {*Place, gender, 2year age range*}.

Experiment J reported that 18.1% of the population in the United States had characteristics that were likely made them unique based only on {*County, gender, date of birth*}. Experiment K reported that 0.04% of the population in the United States had characteristics that were likely made them unique based only on {*County, gender, Month and year of birth*}. Experiment L reported that 0.00004% of the population in the United States had characteristics that were likely made them unique based only on {*County, gender, Year of birth*}. Experiment M reported that 0.00000% of the population in the United States had characteristics that were likely made them unique based only on {*County, gender, 2year age range*}, but despite it being a very small number, it is not 0.[*]

As the number of possible values a quasi-identifier can assume decreases, the percentage of the population in the United States who had characteristics that were likely unique based on those values decreases. This is evidenced by each row in Figure 32. Moving from left to right within each row of Figure 32, the numbers of possible combinations decrease and the corresponding percentages decrease. Aggregating the geographical specification to county resulted in far fewer possible combinations than available with place or ZIP codes. This is evidenced within each column in Figure 32. Notice however that the differences between the number of places and the number of ZIP codes are not as dramatic, and as a result, neither are the corresponding percentages.

6.1. Predicting the number of people that can be identified in a release

It was already shown that de-identified releases of person-specific data that contain no explicit identifiers such as name, address or phone number, is not necessarily anonymous [16]. The maximum number of patients who could be identified in a public or semi-public release of health data is the number of patients who were hospitalized and whose information is therefore included in the data. Many possible combinations of attributes can combine to form a quasi-identifier useful for linking the de-identified data to explicitly identified data. The number of hospitalizations reported in the IHCCCC's R_{rod} data (see Figure 2) in one year is estimated to be 1 million based on the average statistic that 1:12 people are hospitalized each year.

However, the actual number of patients that could be re-identified in publicly and semi-publicly released health data is not necessarily every patient and the actual number is likely to differ among releases due to varying quasi-identifiers available. The results from the experiments reported in this document can help predict a minimum level of identifiability based on a combination of three demographics.

[*] In Loving County, Texas, 6 of 107 people are likely to be uniquely identified by values of {*gender, 2yr age range, county*}. All of these 6 people are between the ages of 12 and 18 years.

6.1.1. Illinois Research Health Data

As shown in Figure 2, R_{rod} includes the full date of birth, gender, and the patient's 5-digit residential ZIP. Figure 13 reports that 75.3% of the population of Illinois is likely to be uniquely identified by {*5-digit ZIP*, *gender*, *date of birth*}. That corresponds to 753,000 patients being identified per year in R_{rod}.

6.1.2. AHRQ's State Inpatient Database

As shown in Figure 3, SID includes the month and year of birth, gender, and the patient's 5-digit residential ZIP for some states. Figure 33 estimates that 112,595 patients per year are likely to be uniquely identified by {*ZIP*, *Gender*, *Month and year of birth*} in SID. The five states known to report the month and year of the birth date of each patient to SID were introduced in **Error! Reference source not found.**. The populations for each of these states according to the 1990 Census data [17] were reported in Figure 8 and Figure 9. It is estimated that 1:12 people are hospitalized each year. These values are summarized in Figure 33.

State	Population	Hospitalized	Unique	PopID
AZ	3,665,228	305,436	1.4%	4,276
IA	2,776,442	231,370	18.1%	41,878
NY	17,990,026	1,499,169	2.3%	34,481
OR	2,842,321	236,860	4.2%	9,948
WI	4,891,452	407,621	5.4%	22,012
			Total per year	***112,595***

Figure 33 Estimated Uniqueness of {*ZIP, Gender, Month and year of birth*} in SID

State	Population	Hospitalized	Unique	PopID
AZ	3,665,228	305,436	0.02%	61
CA	29,755,274	2,479,606	0.01%	248
CO	3,293,771	274,481	0.08%	220
FL	12,686,788	1,057,232	0.00%	42
IA	2,776,442	231,370	0.11%	255
MA	6,011,978	500,998	0.01%	50
MD	4,771,143	397,595	0.03%	119
NJ	7,730,188	644,182	0.01%	64
NY	17,990,026	1,499,169	0.03%	450
OR	2,842,321	236,860	0.07%	166
SC	3,486,703	290,559	0.01%	29
WA	4,866,692	405,558	0.03%	122
WI	4,891,452	407,621	0.02%	82
			Total per year	***1,907***

Figure 34 Estimated Uniqueness of {*ZIP, Gender, Year of birth*} in SID

As shown in Figure 3, SID includes the year of birth (by way of age[18]), gender, and the patient's 5-digit residential ZIP for some states. Figure 34 estimates that 1,907 patients per year are likely to be uniquely identified by {*ZIP*, *Gender*, *Year of birth*} in SID. The 13 states known to report the year of the birth date of each patient to SID were introduced in **Error! Reference source not found.**. The populations for each of these states according to the 1990 Census data [19] were reported in Figure 8 and Figure 9. It is estimated that 1:12 people are hospitalized each year. These values are summarized in Figure 34.

There are many ways to misunderstand these values. These values are not to be considered an estimate of the uniqueness of R_{rod} or SID. There may exist other quasi-identifiers that may consist of more and different attributes that can link to other available data and thereby render the released health data even more identifiable. Such quasi-identifiers may use the hospital identifying number or discharge status or payment information. The estimates reported in this document are just approximations based on the demographic quasi identifiers stated. Therefore, these estimates should be viewed as a minimal estimate of the identifiability of these data. Clearly, these data are not anonymous.

6.2. Unique and unusual information found in data

A significant problem with producing anonymous data concerns unique and unusual information appearing within the data themselves. Instances of uniquely occurring characteristics found within the original data can be used by a reporter, private investigator and others to discredit the anonymity of the released data even when these instances are not unique in the general population. Unusual cases are often unusual in other sources of data as well making them easier to identify.

Importantly, close examination of the particulars of a database provides the best basis for determining uniquely identifying information and quasi-identifiers. In this document, I have examined outside information without examining the values of the released data themselves. The analysis is based on the fact that a combination of characteristics that makes one unique in a geographic population, for example, results in uniqueness in all other data that includes that geographic specification. An examination of the data however can reveal other kinds of unusual information that can be found in other sources of data making more patients easier to identify.

In an interview, for example, a janitor may recall an Asian patient whose last name was Chan and who worked as a stockbroker because the patient gave the janitor some good investing tips. Any single uniquely occurring value or group of values can be used to identify an individual. Remember that the unique characteristic may not be known beforehand. It could be based on diagnosis, treatment, birth year, visit date, or some other little detail or combination of details available to the memory of a patient or a doctor, or knowledge about the database from some other source.

As another example, consider the medical records of a pediatric hospital in which only one patient is older than 45 years of age. Suppose a de-identified version of the hospital's records is to be released for public-use that includes age and city of residence but not birth date or zip code. Many may believe the resulting data would be anonymous because there are thousands of people of age 45 living in that city. However, the rare occurrence of a 45 year-old pediatric patient at that facility can become a focal point for anyone seeking to discredit the anonymity of the data. Nurses, clerks and other hospital personnel will often remember unusual cases and in interviews may provide additional details that help identify the patient.

6.3. Future Work

Below are proposed projects of varying degrees of difficulties and skill requirements that extend this work.

In this document, I have demonstrated how combinations of characteristics can combine to narrow the number of possible people under consideration. However, knowing that there exist a one or a few people that share particular characteristics and explicitly identifying those people are not exactly the same. These combinations of characteristics must be linked to explicitly identified information to reveal the identities of the individuals. Further demonstrate the identifiability of these data by providing population registers to which the data could be linked to re-identify the noted individuals.

In an earlier document [20], privacy risk measures were computed on the data sets R_{rod} and SID based on the assumption that the entire populations within those data were identifiable. While that may be correct, use the findings reported in this document, which are based only on basic demographic attributes and do not include other attributes within those data that could be used for re-identification, and re-compute the measures of risk for those collections. Make an argument as to why these re-computed risk measurements should be considered "minimal" risk values.

7. References

1 National Association of Health Data Organizations, *NAHDO Inventory of State-wide Hospital Discharge Data Activities* (Falls Church: National Association of Health Data Organizations, May 2000).

2 Cambridge Voters List Database. *City of Cambridge, Massachusetts*. Cambridge: February 1997.

3 Supra note 1 NAHDO.

4 State of Illinois Health Care Cost Containment Council, *Data release overview.* (Springfield: State of Illinois Health Care Cost Containment Council, March 1998).

5 Agency for Healthcare Research and Quality, *Healthcare Cost and Utilization Project: Central Distributor* (April, 2000) available at http://www.ahcpr.gov/data/hcup/hcup-pkt.htm.

6 1990 U.S. Census Data, Database C90STF3B. *U.S. Bureau of the Census*. Available at http://venus.census.gov and http://www.census.gov. Washington: 1993.

7 T. Dalenius. Finding a needle in a haystack – or identifying anonymous census record. *Journal of Official Statistics*, 2(3):329-336, 1986.

8 G. Smith. Modeling security-relevant data semantics. In *Proceedings of the 1990 IEEE Symposium on Research in Security and Privacy*, May 1990.

9 J. Ullman. *Principles of Database and Knowledge Base Systems*. Computer Science Press, Rockville, MD. 1988.

10 Supra note 6 U.S. Bureau of the Census.

11 "Census Counts 90," *U.S. Bureau of the Census*. (Available on CDROM.) Washington: 1993.

12 "1996 National Five-Digit ZIP Code and Post Office Directory," *U.S. Postal Service*. Washington: 1996. Also available at http://www.usps.gov.

13 Brualdi, R.A., *Introductory Combinatorics*, North-Holland, New York, 1977.

14 Supra note 5 AHRQ.

15 L. Sweeney. *Inferences from unusual values in statistical data*. Carnegie Mellon Data Privacy Center Working Paper 3.

16 L. Sweeney. Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine and Ethics, 25(2-3):98-110, 1997.

17 Supra note 6 and note 11 U.S. Census.

18 Supra section 4.4.1 Age versus Year of Birth

19 Supra note 6 and note 11 U.S. Census.

20 L. Sweeney. *Towards all the data on all the people*. Formal publication forthcoming. Earlier version available as Carnegie Mellon Data Privacy Center Working Paper 2.