Revisiting the Uniqueness of Simple Demographics in the US Population

Philippe Golle
Palo Alto Research Center
pgolle@parc.com

ABSTRACT

According to a famous study [10] of the 1990 census data, 87% of the US population can be uniquely identified by gender, ZIP code and full date of birth. This short paper revisits the uniqueness of simple demographics in the US population based on the most recent census data (the 2000 census). We offer a detailed, comprehensive and up-to-date picture of the threat to privacy posed by the disclosure of simple demographic information. Our results generally agree with the findings of [10], although we find that disclosing one's gender, ZIP code and full date of birth allows for unique identification of fewer individuals (63% of the US population) than reported in [10]. We hope that our study will be a useful reference for privacy researchers who need simple estimates of the comparative threat of disclosing various demographic data.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Measurement

Keywords

Privacy, Anonymity, Census Data

1. INTRODUCTION

A famous study [10] of the 1990 census data showed that 87% (216 million of 248 million) of the population in the United States reported characteristics that likely made them unique based only on gender, 5-digit ZIP code and full date of birth (day, month and year). The study further reported that 53% of the U.S. population is uniquely identified only by {gender, place, date of birth}, where "place" is basically the city, town, or municipality in which the person resides. Even at the county level, {gender, county, date of birth} uniquely identifies 18% of the U.S. population. In general [10] shows that "few characteristics are needed to uniquely identify a person."

The results of this study are influential and widely cited. They informed the design of U.S. standards of privacy for health information [6, 12] and for the release of Census, research and statistical information [3, 5]. They shed light on the surprising privacy behavior of individuals [2], and motivated research in efficient algorithms to protect the anonymity of data [1, 7].

This paper updates, extends and in one instance finds a discrepancy with the results of [10]. Specifically, we make the following contributions:

- **We study the uniqueness of simple demographics using the most recent census data (the 2000 census [4]).** Our results generally confirm the findings of [10], although we find that disclosing one's gender, ZIP code and date of birth allows for unique identification of fewer individuals than reported in [10]. We found that in 1990 (resp. 2000), only 61% (resp. 63%) of the US population was uniquely identifiable by {gender, ZIP code, full date of birth}, whereas [10] reported that the same attributes allowed for unique identification of 87% of the US population in 1990. Unfortunately, we can not explain this discrepancy because we lack detailed information about the data collection and analysis techniques of [10]. Our data collection and methodology are explained in detail in the following sections, so our results can be replicated and verified.

- **We offer a fine-grained characterization of the privacy threat of disclosing simple demographics.** We define precisely the degree of privacy of individuals on a scale that goes from uniquely identifiable to k-anonymous (i.e. hidden indistinguishably in a group of size k). Compared to [10], this offers a finer-grained view of the degree of privacy of individuals who are not uniquely identifiable.

- **We study the privacy implications, by age, of releasing simple demographics.** We show that the privacy threat is smallest for individuals around age 20, then rises rapidly. This has important implications for the release of data about individuals in specific age ranges. For example, much greater care must be taken with the medical data of elderly populations.

WPES'06, October 30, 2006, Alexandria, Virginia, USA.

2. DATA COLLECTION

The focus of this paper is on the threat to privacy of disclosing one's gender, location (ZIP code or county) and age (either year of birth only, or year, month and day of birth). These simple demographic characteristics were chosen because they appear to be the most valuable to marketers, retailers and social and medical researchers. They are also the most commonly disclosed demographic characteristics in registration forms, surveys and medical and financial files.

To evaluate the privacy threat of disclosing these demographic characteristics, we need to estimate the number of individuals of a given gender and age in a given location. The more individuals, the lesser the privacy threat. When a group of k individuals share the same gender, location and age, each individual in the group is said to be k-anonymous [9, 11] after disclosing these characteristics.

Our source of demographic information is the 2000 Census data, which is available free of charge on the Census Bureau's website [4]. We looked at table PCT12 (Sex by Age), which lists the number of males and the number of females of a given age (from 0 to 99 years old) in a given geographic area. We retrieved this data for:

- All $3,219$ counties and county equivalents (boroughs, census tracts, parishes, independent cities and Municipios) in all 50 States, the District of Columbia and Puerto Rico.
- All $33,233$ ZIP Code Tabulation Areas (ZCTAs) in the 50 states, District of Columbia and Puerto Rico as of Census 2000. The list of ZCTAs can be downloaded from [13].

This data is almost all we need, but not quite. To consider the privacy implications of revealing one's full date of birth (year, month and day), we must estimate the number of individuals who live in a given location and were born on a given day, month and year. Table PCT12 gives the year of birth, but not the month or day (no other table in the census data gives such detailed information, precisely because it is a threat to privacy). We can however compute a precise estimate of the number of individuals born on a given day and month of the year as follows.

We assume that births are uniformly distributed across the days of the year. For our purpose, this is a reasonably close approximation ([8] shows that month-to-month variations in birth rates are negligible and intra-week variations are small). With this assumption, if n people are born in a given year, the expected number $f_k(n)$ of days on which k individuals are born is given by

$$f_k(n) = \binom{n}{k} (365)^{1-n} (364)^{n-k}.$$

This formula is proved in the appendix. We now have all the data needed to study the privacy threat of disclosing simple demographics (one's gender, location and age).

3. RESULTS

Following [10], we compute first the percentage of the U.S. population which is uniquely identifiable by {gender, location, date of birth}, where location is either a 5-digit ZIP code or a county, and date of birth is either the year of birth only, or the year and month of birth, or the full date of birth (year, month and day). Our results are summarized in Table 1.

	5-digit ZIP code	County
Year of birth	0.2%	0.0%
Year and month of birth	4.2%	0.2%
Year, month and day of birth	63.3%	14.8%

Table 1: Fraction of the U.S. population uniquely identifiable by {gender, location, date of birth}.

Our findings are close to [10] at the county level, but differ significantly at the level of ZIP codes. We show that in 2000, only 63% of the US population is uniquely identifiable by {gender, ZIP code, full date of birth}, whereas [10] found 87% uniquely identifiable by the same characteristics in 1990. Unfortunately, we lack detailed information about the methodology and data collection of [10], so we can offer no definite explanation for this discrepancy. We speculate however that the discrepancy might come in part from the fact that the 1990 census does not directly tabulate data by ZIP codes: Summary Tape File 1, which contains 100% of the 1990 census data, can not be queried by ZIP code. A smaller set of sample data from the 1990 census data, in Summary Tape File 3, can be queried by ZIP code, but gives only a coarser representation of the age distribution of individuals (ages are aggregated in 5 year intervals).

The overall number of individuals uniquely identifiable by {gender, location, full date of birth}, while dramatic, says nothing about the degree of anonymity enjoyed by the rest of the population, nor about the effect of age on anonymity. For example, we want to know if the threat to privacy of disclosing demographic data is confined mostly to older people in thinly populated ZIP codes, or if the threat is uniformly distributed across the population.

Anonymity, by age, given {Gender, Location, Full date of birth}. Figures 1 and 2 provide a first answer to this question. They give a more fine-grained view of the degree of anonymity of the US population, by age, given {gender, location, full date of birth}, where location is either a 5-digit ZIP code (Figure 1) or a county (Figure 2). These graphs show that the privacy threat of disclosing simple demographics is fairly uniform between the ages of 0 and 50, then rises rapidly after 50. The graphs also show that even individuals who are not uniquely identifiable enjoy very little anonymity. For example, disclosing {gender, county, full date of birth} leaves 14.8% of the population uniquely identifiable (1-anonymous) but also leaves 43.6% of the population 5-anonymous or less (i.e. hidden indistinguishably in a group of size 5 or less). The proportion of individuals who are 5-anonymous or less rises to 63% for people over the age of 60 (see Figure 2).

Anonymity, by age, given {Gender, Location, Year of birth}. We have shown that disclosing one's full date of birth, together with location information, clearly compromises privacy. In practice, however, surveys and registration forms often ask only for an individual's year of birth (or equivalently, age) rather than the full date of birth. Even when the day and month of birth are asked, one can of-

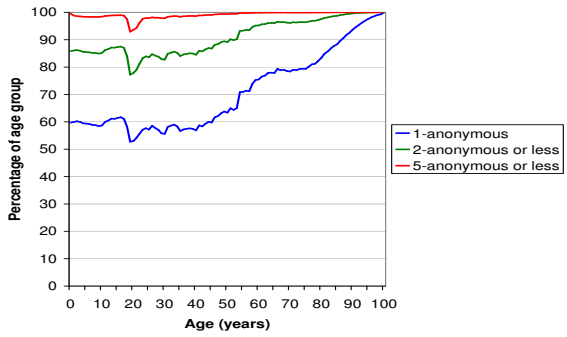

Figure 1: Anonymity of the U.S. population, by age, given {Gender, ZIP code, Full date of birth}.

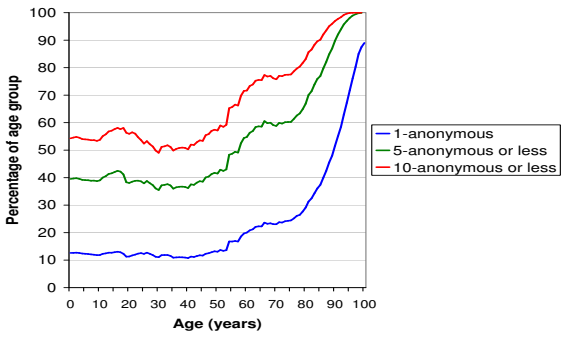

Figure 2: Anonymity of the U.S. population, by age, given {Gender, County, Full date of birth}.

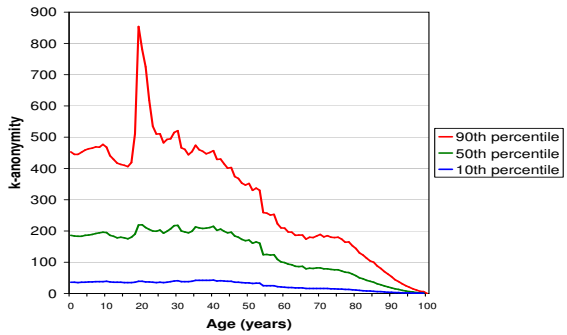

Figure 3: Anonymity of the U.S. population, by age, given {Gender, ZIP code, Year of birth}.

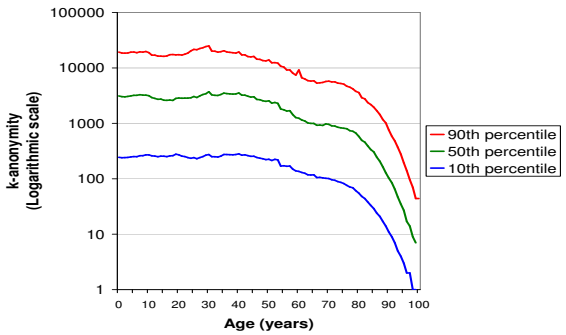

Figure 4: Anonymity of the U.S. population, by age, given {Gender, County, Year of birth}.

ten lie without negative consequence (lying about one's age is more problematic, since the lie is more easily detectable and may result in inadequate service). Table 1 shows that revealing one's {gender, location, year of birth} allows for unique identification of only 0.2% of individuals. Is it then safe for most people to disclose this information? In what follows, we analyze in more detail the privacy implications of disclosing one's {gender, location, year of birth}.

Figures 3 and 4 show the degree of anonymity of the US population, by age, given {gender, location, year of birth}, where location is either a 5-digit ZIP code (Figure 3) or a county (Figure 4). In both graphs, the green curve (the middle curve) shows the median degree of anonymity by age (the degree of anonymity of the 50-th percentile). We see for example that the median anonymity of the population under age 50 after disclosing {gender, ZIP code, year of birth} is 200-anonymity (the median is 3000-anonymity if the county is disclosed instead of the ZIP code).

The blue line shows the anonymity of the lower 10-th percentile. In other words, 10% of the population enjoys less anonymity than indicated by the blue line, and 90% of the population enjoys more anonymity.

Finally, the red line shows the anonymity of the 90-th percentile. We observe a sharp spike in the degree of anonymity of the 90-th percentile (the 10% of the population who are the most anonymous) around age 20. Analysis reveals that this spike comes from college and university towns with high concentrations of individuals between the ages of 18 and 22. For example, the 3 ZIP codes with the highest degree of anonymity in the range [18–22] are College Station, TX (77840); West Lafayette, IN (47906); and Austin, TX (78705). All three ZIP codes are home to large universities.

The conclusion we can draw from this data is that, for 90% of the population under the age of 50, disclosing {gender, ZIP code, year of birth} will result in 40-anonymity or more, while disclosing {gender, county, year of birth} will result in 250-anonymity or better.

Finally, those willing to sacrifice truthfulness for optimal anonymity should claim, when asked for their age and ZIP code, to be a 21-year-old male from Camp Pendleton, California (ZIP code 92054); or, if female, to be a 19-year-old from College Station, Texas (ZIP code 77840). They will share these characteristics with respectively 4,099 other males and 3,744 other females.

4. CONCLUSION

This short paper revisits the uniqueness of simple demographics in the US population based on the most recent census data (the 2000 census). We offer a detailed, comprehensive and up-to-date picture of the threat to privacy posed by the disclosure of simple demographic information. We hope this study will be a useful reference for privacy researchers who need simple estimates of the comparative threat of disclosing various demographic data.

5. REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Approximation Algorithms for k-Anonymity. Journal of Privacy Technology, 2005.

[2] A. Acquisti and J. Grossklags. Privacy and Rationality in Individual Decision Making. *IEEE*

Security and Privacy, IEEE Computer Society, Vol. 3, No. 1, January/February 2005, pp. 26-33.

[3] EPIC. The Census and Privacy. http://www.epic.org/privacy/census/

[4] U.S. Census Bureau FactFinder. http://factfinder.census.gov

[5] United States General Accounting Office. Record linkage and privacy. Issues in creating new federal research and statistical information. GAO-01-126SP.

[6] L. Lamberg. Confidentiality and Privacy of Electronic Medical Records. JAMA 2001, 285(24):3075-3076.

[7] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient Full-Domain K-Anonymity. In ACM SIGMOD International Conference on Management of Data, June 2005.

[8] National Center for Health Statistics. Births: Final Data for 2003. National Vital Statistics Reports Volume 54, Number 2. (116) pp. 2005-1120. http://www.cdc.gov/nchs/

[9] P. Samarati. Protecting Respondents' Identities in Microdata Release. IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001)

[10] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

[11] L. Sweeney. K-anonymity: a Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[12] L. Sweeney. Comments to the Department of Health and Human Services On "Standards of Privacy of Individually Identifiable Health Information". http://privacy.cs.cmu.edu/dataprivacy/HIPAA/HIPAAcomments.html

[13] 2000 Zip Code Tabulation Areas. http://www.census.gov/tiger/tms/gazetteer/zcta5.txt

APPENDIX

A. LEMMA

LEMMA A.1. *Assume that n individuals are distributed uniformly independently at random into N bins. Let $f_i(n)$ be the expected number of bins that contain i individuals. Then*

$$f_i(n) = \binom{n}{i} N^{1-n}(N-1)^{n-i}$$

PROOF. The value $f_i(n)$ is determined for all $i \geq 0$ and all $n \geq 0$ by the following formulas

$$\begin{aligned} f_0(0) &= N \\ f_0(n+1) &= f_0(n)(1-1/N) \\ f_i(n+1) &= f_i(n)(1-1/N) + f_{i-1}(n)/N \end{aligned}$$

Let us define $g_i(n) = f_i(n)N^{i-1}(1-1/N)^{i-n}$. The equations above become:

$$\begin{aligned} g_0(0) &= 1 \\ g_0(n+1) &= g_0(n) \\ g_i(n+1) &= g_i(n) + g_{i-1}(n) \end{aligned}$$

Therefore $g_i(n) = \binom{n}{i}$ and $f_i(n) = \binom{n}{i}N^{1-i}(1-1/N)^{n-i} = \binom{n}{i}N^{1-n}(N-1)^{n-i}$. □