week 9 day 1

Bounds on cell entries
Algebraic & Geometric Methods in Statistics

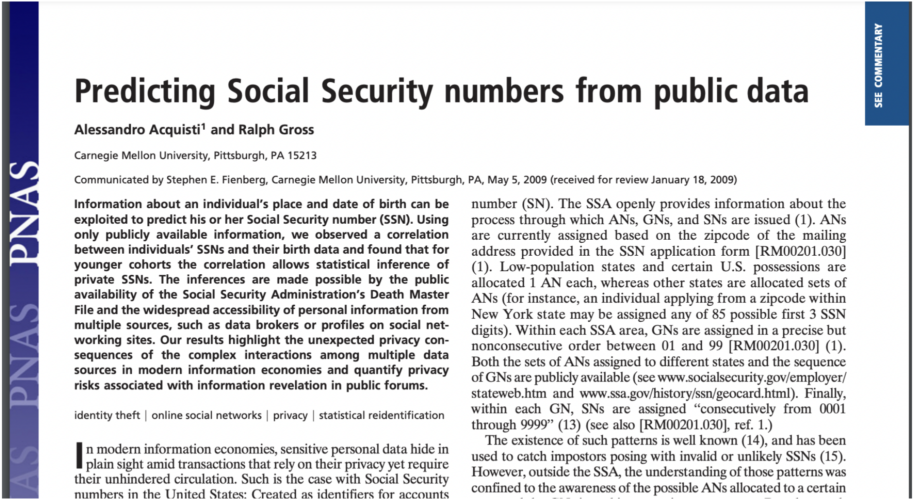
Sonja Petrović
Created for Math/Stat 561

Mar 1, 2023.

Material

- Chapter 10: Section 10.1 introduces the problem
- Statistical justifications for studying cell bounds: Disclosure limitation
    - Here are three sources of the very similar story that is easy to read:
        - Dobra&Fienberg 2000, PNAS
        - Dobra&Fienberg 2001, Statistical Journal of the United Nations
- Prof. A. Dobra has an algorithm implemented in C++ for computing the cell bounds. This is based on his 2002 PhD thesis.
- Section 10.5 has formulas for bounds on table entries
    - Specifically, theorem 10.5.6 gives tight cell bounds
    - This material is *advanced reading*. It relies on sections 10.2-10.4, with background in algebra, Gröbner bases, integer programming relaxation, and some polyhedral geometry.
    - This part of the chapter is left as supplementary reading.

What is *disclosure limitation*?



SEE COMMENTARY

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies and quantify privacy risks associated with information revelation in public forums.

identity theft | online social networks | privacy | statistical reidentification

In modern information economies, sensitive personal data hide in plain sight amid transactions that rely on their privacy yet require their unhindered circulation. Such is the case with Social Security numbers in the United States: Created as identifiers for accounts

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within New York state may be assigned any of 85 possible first 3 SSN digits). Within each SSA area, GNs are assigned in a precise but nonconsecutive order between 01 and 99 [RM00201.030] (1). Both the sets of ANs assigned to different states and the sequence of GNs are publicly available (see www.socialsecurity.gov/employer/stateweb.htm and www.ssa.gov/history/ssn/geocard.html). Finally, within each GN, SNs are assigned "consecutively from 0001 through 9999" (13) (see also [RM00201.030], ref. 1.)

The existence of such patterns is well known (14), and has been used to catch impostors posing with invalid or unlikely SSNs (15). However, outside the SSA, the understanding of those patterns was confined to the awareness of the possible ANs allocated to a certain

Figure 1: https://www.heinz.cmu.edu/~acquisti/papers/AcquistiGross-PNAS-2009.pdf

“Those freaked out by facial recognition technology have fresh fodder: a study from Carnegie Mellon University in which researchers were able to predict people’s social security numbers after taking a photo of them with a cheap webcam.” Forbes, 2001, ’How Facial Recognition Technology Can Be Used To Get Your Social Security Number’.

“In the second experiment, they used a $35 webcam to take photos of CMU students. They then asked the 93 participants to take a quick online survey. While they did that, the facial recognition software went to work figuring out who they were. Acquisti told me that 42% of those participants were linked to their Facebook profiles.

“For those participants who had date of birth and city publicly available on their account, the researchers could predict a social security number (based on the work from their 2009 study). The researchers sent a follow-up survey to their student participants asking them whether the first five digits of the social security number their algorithm predicted was correct.”

- The usual issue: Data privacy and confidentiality
- Trade-off with Statistical utility.

Disclosure limitation

How much of the data can one release to the public while preseving privacy & at the same time allowing for statistical utility?

Example: consider the context of a contingency table.

- If you have a model in mind, say a model of independence, then releasing *sufficient statistics* (marginals) is, well, sufficient for statistical analyses:
  - probabilities are determined;
  - you can compute p-values.
- Therefore, you are hiding the data completely, not releasing any sensitive information, and satisfying the statistical utility.

	M	F	T/Nb	totals
$\leq$ 135K	?	?	?	**13**
$>$ 135K	?	?	?	**13**
totals	**10**	**10**	**6**	**26**

. . . **are** you hiding the data completely?

From our book:

“For example, in Table 10.1.1” – it is a 5-dimensional table, but *very sparse* – ” the release of all 3-way marginals of the table does not mask the table details at all: in fact, it is possible to recover **all** table entries given all 3-way margins in this case! If we restrict to just 2-way marginals, then by computing linear programming upper and lower bounds, we are uniquely able to recover one of the table entries, namely the position marked by 1. This example shows that even releasing quite low-dimensional marginals on a 5-way table that is sparse can still reveal table entries that are sensitive.”

We will now consider a series of examples from Fréchet and Bonferroni Bounds for Multi-way Tables of Counts With Applications to Disclosure Limitation by Stephen E. Fienberg. * Similar example is in the book, 10.1 * "Computing bounds on cell entries in 2-way tables is especially easy. In general, it is difficult to find general formulas for the bounds on cell entries given marginal totals." * Summary of the story and its impact is in the two references listed on the first slide. * These tables may not be sparse but they showcase how combining different marginal information can help narrow down specific cell bounds. * They also show how this is not a straightforward problem to solve.

- Data from the 1990 U.S. decennial census public use sample for a local area,in the form of a $3 \times 2 \times 2$ table of counts
- “noteworthy features”:
  - it includes three counts of “1”, or sample uniques.
  - there are counts of “1” in two of the three two-way marginal totals.
  - Thus, if we think in terms of constraining the interior cells of the table given the margins, we can expect to get tight bounds for some of the cell entries.

Gender = Male

Income Level

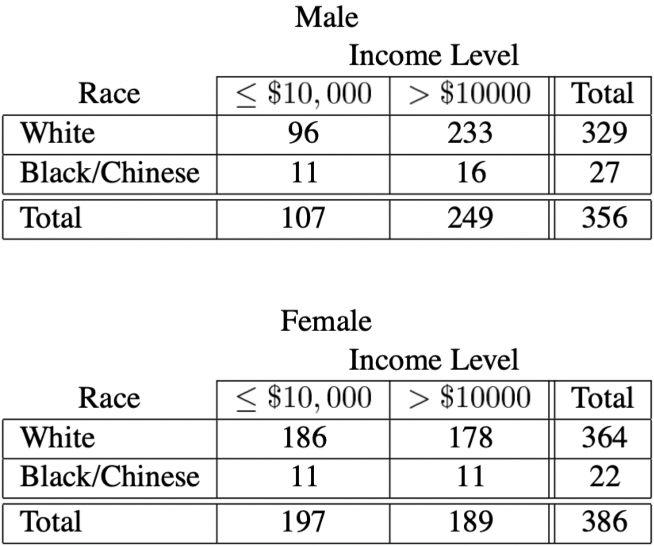
Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total
White	96	72	161	329
Black	10	7	6	23
Chinese	1	1	2	4
Total	107	80	169	356

Gender = Female

Income Level

Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total
White	186	127	51	364
Black	11	7	3	21
Chinese	0	1	0	1
Total	197	135	54	386

Table 1: Three-way cross-classification of Gender, Race, and Income for a selected U.S. census tract. (*Source*: 1990 Census Public Use Microdata Files)



Male

Race	Income Level		
	$\leq \$10,000$	$> \$10000$	Total
White	96	233	329
Black/Chinese	11	16	27
Total	107	249	356

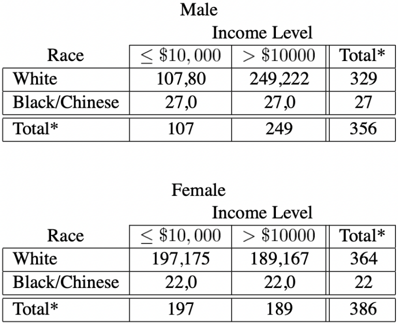
Female

Race	Income Level		
	$\leq \$10,000$	$> \$10000$	Total
White	186	178	364
Black/Chinese	11	11	22
Total	197	189	386

Table 2: Collapsed $2 \times 2 \times 2$ version of cell counts in Table 1.

In $2 \times 2 \times 2$ tables: consider layers 1 and 2 separately, then we have a pair of 2×2 tables.

- Simple bounds: $min\{u_{i+}, u_{+j}\} \geq u_{ij} \geq max\{u_{i+} + u_{+j} - n, 0\}$
- These bounds in effect fix the entries in two of the three 2-way margins of the full $2 \times 2 \times 2$ table:



Male

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	107,80	249,222	329
Black/Chinese	27,0	27,0	27
Total*	107	249	356

Female

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	197,175	189,167	364
Black/Chinese	22,0	22,0	22
Total*	197	189	386

Table 3: Fréchet bounds fixing the 1-way margins for each layer of Table 2.

Next we consider fixing all three 2-way margins.

- This problem has a simple generic form.
- In effect, we are given 7 values:
  - the sums for each of the (1,1) cells of the three 2-way margins,
  - the sums for the 1st entry in each of the three 1-way margins and the grand total.
- All of the other marginal values can be computed from these.
- Thus we need only one more quantity to determine the entries of the full table!

Let x be the true but unknown value of the count in the (1,1,1) cell.

We have:

$$
\begin{gathered}
u_{111} = x \\
u_{121} = u_{1+1} - x \\
u_{112} = u_{11+} - x \\
u_{211} = u_{+11} - x \\
u_{122} = u_{1++} - u_{1+1} - u_{11+} + x \\
u_{212} = u_{+1+} - u_{11+} - u_{+11} + x \\
u_{221} = u_{1++} - u_{+11} - u_{1+1} + x \\
u_{222} = n - u_{1++} - u_{+1+} - u_{++1} + u_{11+} + u_{1+1} + u_{+11} - x.
\end{gathered}
$$

Now if we add the non-negativity constraint for cell counts in a contingency table:

$$u_{ijk} \geq 0$$

- get 4 upper bounds and 4 lower bounds.
- Three of the 4 upper bounds components involve the 2-way marginal totals corresponding to the (1,1,1) cell
- the 4th one is:

$$n - u_{1++} - u_{+1+} - u + + + 1 + u_{11+} + u_{1+1} + u_{+11} = u_{111} + u_{222}.$$

Result? Cell bounds on x:

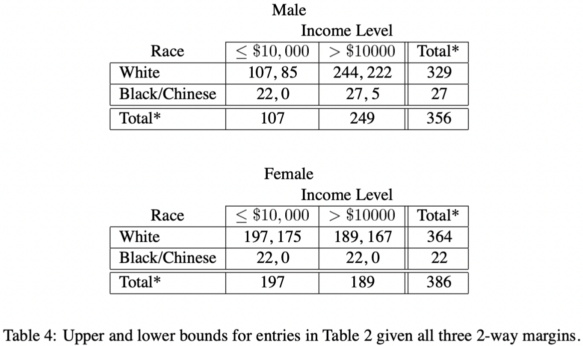
$$min\{u_{11+}, u_{1+1}, u_{+11}, u_{111} + u_{222}\} \geq x$$

$$x \geq max\{u_{1++} - u_{1+1} - u_{11+}, u_{+1+} - u_{11+} - u_{+11}, u_{1++} - u_{+11} - u_{1+1}, 0\}.$$

$$min\{u_{11+}, u_{1+1}, u_{+11}, u_{111} + u_{222}\} \geq x$$

$$x \geq max\{u_{1++} - u_{1+1} - u_{11+}, u_{+1+} - u_{11+} - u_{+11}, u_{1++} - u_{+11} - u_{1+1}, 0\}.$$

The result:



Male

	Income Level		
Race	$\leq \$10,000$	$> \$10000$	Total*
White	107, 85	244, 222	329
Black/Chinese	22, 0	27, 5	27
Total*	107	249	356

Female

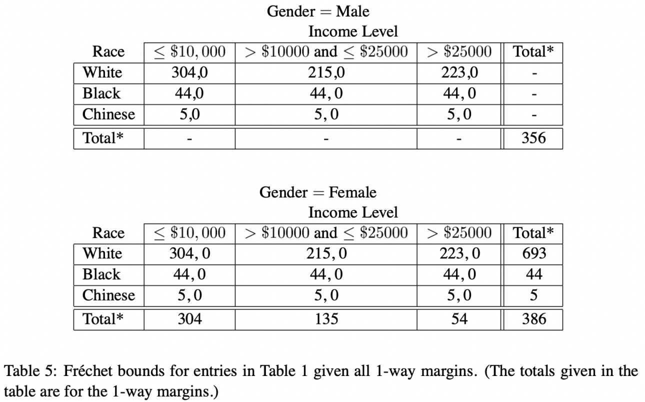
	Income Level		
Race	$\leq \$10,000$	$> \$10000$	Total*
White	197, 175	189, 167	364
Black/Chinese	22, 0	22, 0	22
Total*	197	189	386

Table 4: Upper and lower bounds for entries in Table 2 given all three 2-way margins.

Fienberg says:

“Despite the existence of explicit upper and lower bounds in the case of the $2 \times 2 \times 2$ contingency table with fixed 2-way margins various authors have suggested the need to resort to linear programming and other indirect methods to find the tightest possible bounds”

- Generalization to k-way tables! :)



Gender $=$ Male

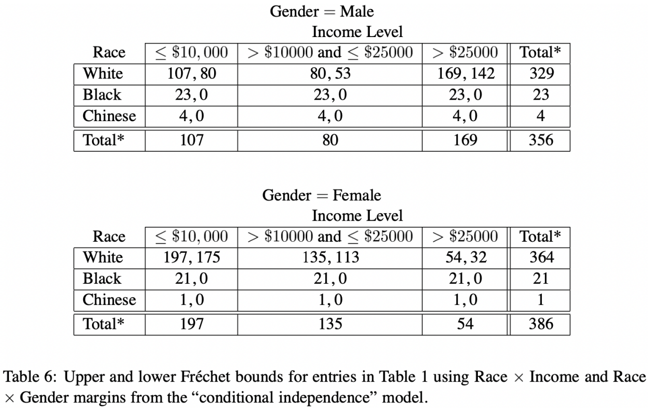
		Income Level		
Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	304,0	215,0	223,0	-
Black	44,0	44, 0	44, 0	-
Chinese	5,0	5, 0	5, 0	-
Total*	-	-	-	356

Gender $=$ Female

		Income Level		
Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	304, 0	215, 0	223, 0	693
Black	44, 0	44, 0	44, 0	44
Chinese	5, 0	5, 0	5, 0	5
Total*	304	135	54	386

Table 5: Fréchet bounds for entries in Table 1 given all 1-way margins. (The totals given in the table are for the 1-way margins.)

m-dimensional marginal bounds for k-way tables



Gender = Male

	Income Level			
Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	$107, 80$	$80, 53$	$169, 142$	329
Black	$23, 0$	$23, 0$	$23, 0$	23
Chinese	$4, 0$	$4, 0$	$4, 0$	4
Total*	107	80	169	356

Gender = Female

	Income Level			
Race	$\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	$197, 175$	$135, 113$	$54, 32$	364
Black	$21, 0$	$21, 0$	$21, 0$	21
Chinese	$1, 0$	$1, 0$	$1, 0$	1
Total*	197	135	54	386

Table 6: Upper and lower Fréchet bounds for entries in Table 1 using Race × Income and Race × Gender margins from the "conditional independence" model.

Applications & implications

- Many proposals for disclosure limitation deal with queries that arrive sequentially
  - Suppose that an agency has responded to a sequence of queries, by releasing g different but possibly overlapping sets of marginal totals, involving k variables having determined that the risk of disclosure is acceptable.
  - Now the agency receives a new query, for the $(g+1)$st set of marginal totals involving a different subset of the k variables (and possibly some additional ones).
  - To determine whether the new request is safe the agency need only compute the upper and lower bounds associated with holding the $(g+1)$ different margins fixed.
    - The bounds for each cell entry in a contingency table represent values associated with *extremal tables* that lie on the boundaries of a convex polytope and we typically get an upper bound occurring simultaneously with lower bounds for other cells, etc.
    - Shuttle algorithm; see Dobra's implementation.

The usual… license

This document is created for Math/Stat 561, Spring 2023.