Disclosure control for a dataset with uncommon characteristics: A case study of the Census of Fatal Occupational Injuries (CFOI)

Danny Friel

Office of Compensation and Working Conditions

FCSM Research & Policy Conference
October 23, 2024

Co-authors

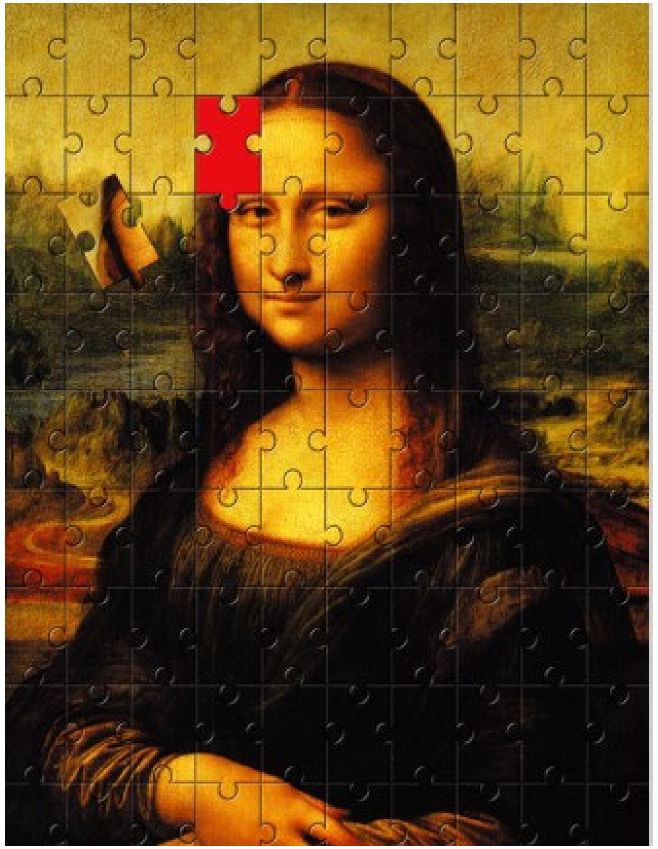
- Alyssa Gillen
- Julie Krautter
- Yvan Saastamoinen

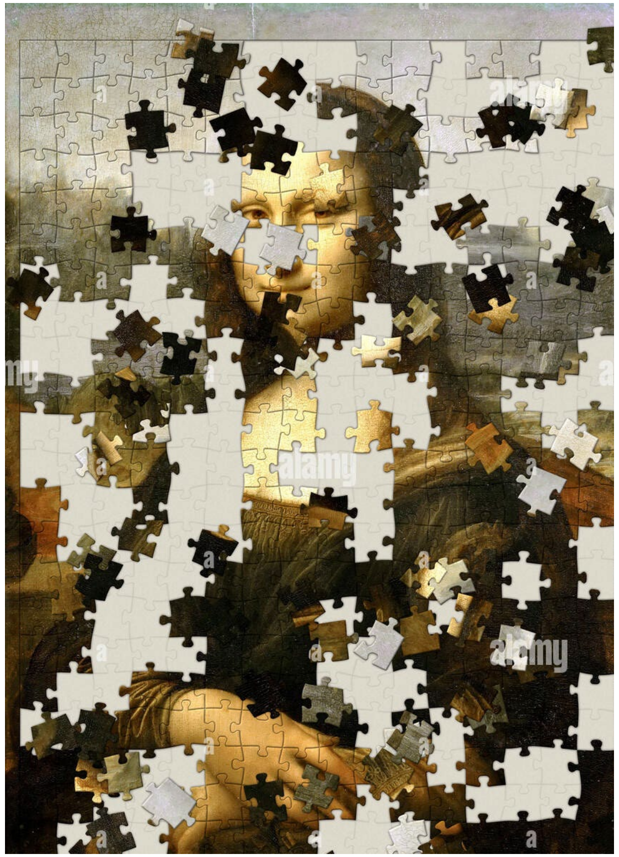
The views expressed here are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics or any other agency of the U.S. Department of Labor.

Outline

- What is disclosure control?
- Disclosure control for CFOI
- Defining utility
- Refining the hypercube
- Discussion

Inference in the face of uncertainty





Census of Fatal Occupational Injuries (CFOI)

- Publishes a complete count of fatal injuries each year
- Protecting CFOI data is challenging
  - No sampling
  - Fatal injuries are rare events
  - Exact counts are important
  - Cases are classified into 16 categorical variables (industry, occupation, gender, nature of injury, …)

Primary vs. secondary suppression

Primary suppression only

The count for occupation 3 doesn't meet publishability criteria

Occupation	Number of fatal injuries
All occupations	**100**
Occupation 1	80
Occupation 2	18
Occupation 3	--

Even though this cell is suppressed, we have enough information to compute its value: 100 – 80 – 18 = 2

Primary and secondary suppressions

The count for occupation 2 is suppressed as well

Occupation	Number of fatal injuries
All occupations	**100**
Occupation 1	80
Occupation 2	--
Occupation 3	--

With two cells suppressed, we don't have enough information to compute either value.
Possible values include 20 and 0, 19 and 1, 10 and 10, 15 and 5...

Table differencing

Occupation 2	Number of fatal injuries
Full-time	12
Part-time	6

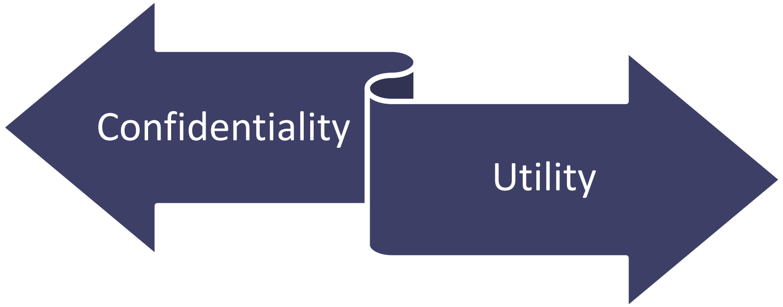
Table differencing

Occupation	Number of fatal injuries
1	80
2	--
3	--
Total	100

Occupation 2	Number of fatal injuries
Full-time	12
Part-time	6

Practical considerations

- Current method: custom Hypercube approach
- We need to effectively manage disclosure risk with limited computational resources
  - 1.06 octillion (10^{27} possible cells)
    - 117 billion are part of the publication subset
- Utility function is complex



Defining utility

Industry A	
	Unprotected data
	50
Violence	2
Transportation	8
Fires	5*
Falls	15
Harmful substances	1
Contact w/equipment	3
Exhaustion	6
Unknown	10

Contact w/equipment	
	Fatal injuries
Contact w/equipment	**30**
Industry A	--
Industry B	10
Industry C	10
Industry D	--
Industry E	5

Changing the order of operations

- Post-processing steps
  - If the high-value cells aren’t at the top of a hierarchy, we can screen them first and then aggregate up to the higher levels

Screen using ownership hierarchy

	Unprotected	Protected
All ownerships	**10**	**10**
Private	2	--
Federal	3	3
State	3	3
Local	1*	--

OR

Screen children first

	Unprotected
Private	2
Federal	3
State	3
Local	--

Screen children first

	Protected
All ownerships	--
Private	2
Federal	3
State	3
Local	--

Leveraging empty cells

	All Events	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6
Industry A	**21**	**4**	**--**	**8**	**--**	**4**	**--**
Industry A-1	**7**	--	--	--	--	1	--
Industry A-2	**12**	--	--	3	--	3	--
Industry A-3	**2**	--	--	--	--	--	--

Leveraging empty cells

	All Events	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6
Industry A	**21**	**4**	**--**	**8**	**0**	**4**	**--**
Industry A-1	**7**	--	--	--	0	1	--
Industry A-2	**12**	--	--	3	0	3	--
Industry A-3	**2**	--	0	--	0	0	--

CFOI tabulation and review

Table shell

	Fatal injuries
Industry A	0
Industry B	0
Industry C	0
Industry D	0
Industry E	0

Tabulate data

	Fatal injuries
Industry A	7
Industry B	29
Industry C	2*
Industry D	0
Industry E	0

Screen for disclosure

	Fatal injuries
Industry A	??
Industry B	??
Industry C	??
Industry D	??
Industry E	??

CFOI tabulation and review

Tabulate case data	
Total	**40**
Industry A	10
Industry B	28
Industry C	2*

Generate empty cells	
Total	**40**
Industry A	10
Industry B	28
Industry C	2*

Screen for disclosure	
Total	**40**
Industry A	--
Industry B	28
Industry C	--

CFOI tabulation and review

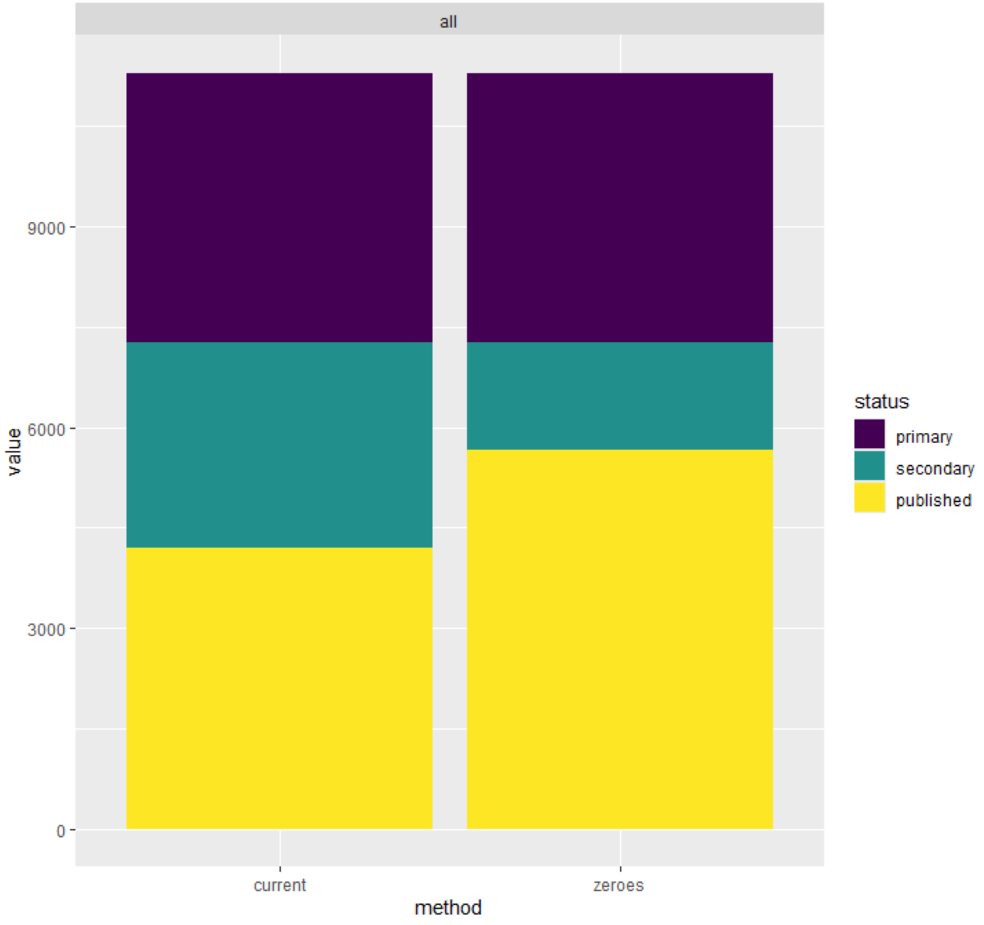
Tabulate case data	
Total	**40**
Industry A	10
Industry B	28
Industry C	2*

Generate empty cells	
Total	**40**
Industry A	10
Industry B	28
Industry C	2*
Industry D	0
Industry E	0

Screen for disclosure	
Total	**40**
Industry A	10
Industry B	28
Industry C	--
Industry D	--
Industry E	0

Results: Leveraging empty cells

- Sharp increase in processing time
- 48% decrease in nonzero secondary suppressions
- Among zeros, 89% published
  - 90,000 zeroes added to dataset compared to ~20,000 nonzero cells



Summary and future work

- Cell suppression algorithms are flexible
    - But, tweaks must be carefully evaluated
- Leveraging zeroes during disclosure screening greatly reduces secondary suppressions
    - But, generating zeroes sharply increases the size of the dataset
- Many ways to optimize for utility
    - But, utility remains ill-defined concept

Contact Information

Danny Friel
Office of Compensation and Working Conditions
Friel.Daniel@bls.gov