Participant ID: XXXX____

Exploring the Opportunity and Risk in Data Released to the Public

Spring 2022

Today we will develop your intuitions for privacy risk in public, shared data. First, we will ask why data that is released to the public matters: why would someone like the Census collect data at all, and why would they release it? Next we will think about the risks of releasing data. We will work with data "hands-on" to evaluate risk: can we reconstruct the identity of a survey participant using only statistics about the people living in their neighborhood block, even when those statistics are protected for privacy using standard methods? We will see! Finally, we will take stock of what we have learned, and think about how well it generalizes to other kinds of data, protected in other ways.

The following questions should be completed when instructed by the SEEK team.

Estimating risk

1. What percentage of the people surveyed would you estimate to be identifiable from table 1?

Looking at the data

1. Explain to the person next to you how complementary suppression works. Write a few notes from your discussion below.

2. Can you create a constraint even for suppressed data? Can you give an example?

3. Write the constraint representing that the age of A is between 0 and 125 years old.

   (a) Now write the constraint representing that the mean age of A and B is 30 years old.

4. Input the two constraints on the board into the Reidentification Explorer. How do you interpret the output you see?

Estimating Risk, again

1. How many of the people surveyed were identified?

2. Why does it matter that smaller groups bear greater risk from reconstruction attacks?

3. Is matching a unique row in a database the same as being identified? Consider sampling, where inferences about a broader population are made on the basis of a subset of the population. Can you think of an example to do with sampling when finding a unique assignment of values to a database row is not the same as identifying a person? Why might this matter?

		age		
statistic	group	count	median	mean
1A	total population	15	37	32.87
2A	female	8	31	29.88
2B	male	7	38	36.29
2C	black or AA	6	39	48
2D	white	9	16	22.78
3A	no STI	8	15	16.88
3B	STI	7	40	48.71
4A	black or AA male	(D)	(D)	(D)
4B	black or AA female	(D)	(D)	(D)
4C	white male	3	16	23.33
4D	white female	6	22.5	22.5
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	4	11	11
5C	persons 64 years or over	(D)	(D)	(D)

Table 1: Simplified statistical table over demographic groups, and a healthcare-related group of people diagnosed with a sexually-transmitted infection (STI).

		age		
statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or AA	3	40	48
2D	white	4	26.5	25.25
3A	no STI	(D)	(D)	(D)
3B	STI	4	38.5	45.25
4A	black or AA male	(D)	(D)	(D)
4B	black or AA female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

Table 2: A statistical table compiled from a subset of the individuals present intable 1.