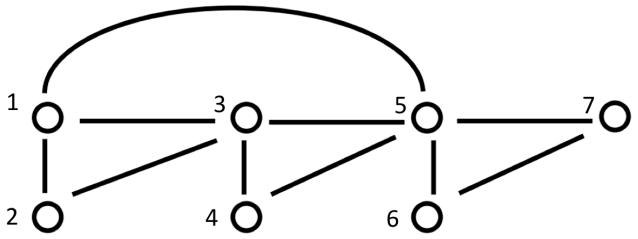
In-class Exercise 1: Privacy Attacks Practicum

Instruction: Divide into groups of 3.

Problem 1: An email service provider released an email communication graph of its customers. Each node represents an individual and each edge between two individuals indicates that they have exchanged emails. All the identifiers of the customers are removed from this graph. The following graph is a sub-graph for 7 individuals (Alice, Bob, Cathy, Diane, Ed, Fred, and Grace).



Unfortunately for the email service providers, Alice and Cathy are investigative journalists are able to gather the following pieces of information:
(i) Alice has sent emails to Bob, Cathy, and Ed;
(ii) Cathy has sent emails to everyone, except Ed.
Can Alice and Cathy map nodes in the released graph above to identities of people? Besides the identify of the individuals in the graph, what other information can they learn?

Problem 2: The same email service provider released the ages of all its users. But instead of releasing the true ages, the service provider released perturbed ages by using a linear function, $\alpha x + \beta$, where **x** is the true age. The email service provider keeps parameters α, β secret (and thus calls its release privacy preserving). The following table was released.

Node ID	Age (perturbed)
1	40
2	34
3	52
4	28
5	48
6	22
7	92

Alice, age 25, and Cathy, age 29, are at it again. Can they deduce Bob and Ed's ages?

Problem 3: Two hospitals released their patient records collected in one month for healthcare study and research. They anonymize their records by generalize the sensitive information of the demographic information of the patients. Hospital A ensures at least 4 patient records have the same demographic information (4-anonymous) while Hospital B ensures at least 6 patient records have the same demographic information (6-anonymous). Below are subsets of tables released by Hospital A and B respectively.

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

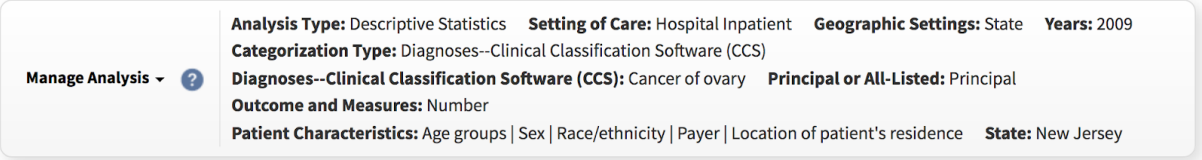
(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

Alice's employer knows that she is 28 and she had taken off to visit both hospitals. Can Alice's employer deduce her medical condition?

Problem 4: Visit https://hcupnet.ahrq.gov/ Create a new analysis with the following setting:



(Click on 'Impatient' → 'Descriptive Statistics' → 'Year 2009' → 'Yes, Statistics about a specific condition/disease' → 'CCS' → '27. Cancer of Ovary' → 'Create Analysis'; Select 'NJ' for state and all the patient characteristics on the left bar)

An aggregate table will be shown to you. Some counts are labeled as '*' because they are smaller than k=10 and are suppressed to achieve k-anonymity. Can you tell what their values are?

How about this table? Can you tell what values these * are?

Age	#discharges	White	Black	Hispanic	Asian/Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	*	19	22
1-17	*	*	*	*	*	*	*	*
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*