Privacy Worksheet

Solution Sheet

Zak Varty

The questions on this sheet are designed to let you test your own understanding of the course content on privacy. Some questions will test basic notions, while others will encourage you to think more deeply about some of the concepts introduced this week.

Question 1: Definitions of Privacy Terms

Define the following terms:

1. A k-anonymous database
2. Data supression
3. Data generalisation
4. Background-knowledge attack
5. Homogeneity attack

💡 Solution

1. A database containing information on a collection of individuals is said to be *k-anonymous* if the data for any one individual are indistinguishable from at least k - 1 other individuals in that database.

2. Data suppression is a method of anonymising data by fully or partially redacting certain data fields.

3. Data generalisation is a method for anonymising data by using grouping to reduce the granularity of a data set.

4. A background-knowledge attack uses additional data, external to a database, to break or reduce the level of anonymity within that database.

Example: Suppose the attacker knows that Jane Doe is 32 years old and that all women receiving HIV treatment in the area are over 40. Alone, this information tells the attacker that Jane does not have HIV. In combination with the hospital records from previous example it also reveals the diagnoses of all three women.

5. A homogeneity attack exploits k-anonymity to establish that an individual has a protected attribute that takes one of a small number of values.

   Example: A hospital records the gender, age and diagnosis of all patients on a ward. Three entries relate to women aged 35-45 on ward C. Two of these women are diagnosed with HIV while the other has hepatitis. An attacker can conclude the Jane Doe has one of these two diseases.

Question 2: A Teacher's Gradebook (K-anonymity)

A teacher keeps the following gradebook of her student's grades. Missing values are indicated be an asterisk (*).

student_id	dob	gender	score	grade
2166433	2000-05-18	F	76	A
2124771	1999-12-29	M	63	B
2197243	2000-04-06	M	48	F
2135974	2000-05-29	*	70	A
2176719	2000-04-14	M	65	B
2130069	1999-12-17	*	61	B
2199235	2000-05-20	F	63	B
2153174	2000-04-21	M	38	F
2199376	2000-04-23	F	54	C
2168752	1999-12-08	F	59	C

a) Why is this gradebook not 2-anonymous with respect to `student_id`, `dob`, `gender` and `grade`?

Solution (a)

There are several reasons why the gradebook is not 2-anonymous. The most obvious of these is that each student has a unique identification number `student_id` (by construction), date of birth `dob` and raw result `score` (by coincidence). There being even a single data attribute that is unique to each individual makes it 2-anonymity impossible, since each combination of attributes that appears in the database can not be repeated at least twice.

b) Suggest how suppression could be used to achieve 2-anonymity with respect to `student_id`, `dob`, `gender` and `grade` in this gradebook. Can you find multiple ways of doing this?

Solution (b)

The gradebook could be made 2-anonymous by suppressing `student_id`, `dob` and `score`.

student_id	dob	gender	score	grade
*	*	F	*	A
*	*	M	*	B
*	*	M	*	F
*	*	*	*	A
*	*	M	*	B
*	*	*	*	B
*	*	F	*	B
*	*	M	*	F
*	*	F	*	C
*	*	F	*	C

This could also be achieved with less information loss my masking only the last 5 digits of student id, so that we still know that all students are in the 2021 cohort.

student_id	dob	gender	score	grade
21*****	*	F	*	A
21*****	*	M	*	B
21*****	*	M	*	F
21*****	*	*	*	A
21*****	*	M	*	B
21*****	*	*	*	B
21*****	*	F	*	B
21*****	*	M	*	F
21*****	*	F	*	C
21*****	*	F	*	C

We can not retain the first digit of the raw score while maintaining 2-anonymity, becuase this carries the information needed to distiguish between the males who failed (rows 3 and 8).
Revealing the year of birth would also break 2-anonymity because (for example) in the database there is not a second male born in 2000 with a B grade to match the individual represented in the fifth row.

student_id	dob	gender	score	grade
21*****	2000 - * - *	F	*	A
21*****	1999 - * - *	M	*	B
21*****	2000 - * - *	M	*	F
21*****	2000 - * - *	*	*	A
21*****	2000 - * - *	M	*	B
21*****	1999 - * - *	*	*	B
21*****	2000 - * - *	F	*	B
21*****	2000 - * - *	M	*	F
21*****	2000 - * - *	F	*	C
21*****	1999 - * - *	F	*	C

c) Suggest how data generalisation could be used to achieve 2-anonymity for student grades in this gradebook.

Solution (c)

If we were to aggregate date of birth to year of birth and aggregate grades into groups of Pass and Fail marks then we can maintain 2-anonymity.

student_id	yob	gender	score	pass
21*****	2000	F	*	TRUE
21*****	1999	M	*	TRUE
21*****	2000	M	*	FALSE
21*****	2000	*	*	TRUE
21*****	2000	M	*	TRUE
21*****	1999	*	*	TRUE
21*****	2000	F	*	TRUE
21*****	2000	M	*	FALSE
21*****	2000	F	*	TRUE
21*****	1999	F	*	TRUE

d) How do the your assumptions about the missing values in the gender attribute impact the k-anonymity of this data set?

Solution (d)

We have assumed that the missing gender attributes are *missing at random*. This means that their missingness is unrelated to any of the measured attributes for those two individuals. If this is the case then the database is 2-anonymous.
If we have background knowledge that these values are missing because these two students identify as non-binary then our data base following suppression is no longer 2-anonymous as there is only one female with an A grade,

student_id	dob	gender	score	grade
21*****	*	F	*	A
21*****	*	M	*	B
21*****	*	M	*	F
21*****	*	NB	*	A
21*****	*	M	*	B
21*****	*	NB	*	B
21*****	*	F	*	B
21*****	*	M	*	F
21*****	*	F	*	C
21*****	*	F	*	C

and our data base following generalization is no longer 2-anonymous as there is only one male born in 1999 who passed.

student_id	yob	gender	score	pass
21*****	2000	F	*	TRUE
21*****	1999	M	*	TRUE
21*****	2000	M	*	FALSE
21*****	2000	NB	*	TRUE
21*****	2000	M	*	TRUE
21*****	1999	NB	*	TRUE
21*****	2000	F	*	TRUE
21*****	2000	M	*	FALSE
21*****	2000	F	*	TRUE
21*****	1999	F	*	TRUE

This gives a simple demonstration of how missingness can break k-anonymity when the fact that values are missing is informative about their value.

e) Why might the teacher wish to use each of suppression and generalisation to anonymise this database?

Solution (e)

Suppression is particularly useful for data attributes that take unique values for many individuals or when those attributes are not going to be used following publication of the database. Conversely, data generalisation is useful for attributes for which at least partial information is required following publication of the database.

For example, if the teacher were publishing this data so that the head teacher could compare pass rates of older and younger students. The gender attribute could be suppressed and the score attribute summarised to Pass/Fail, allowing greater granularity in the aggregation of dob:

student_id	dob	gender	score	pass
21*****	2000-05	*	*	TRUE
21*****	1999-12	*	*	TRUE
21*****	2000-04	*	*	FALSE
21*****	2000-05	*	*	TRUE
21*****	2000-04	*	*	TRUE
21*****	1999-12	*	*	TRUE
21*****	2000-05	*	*	TRUE

21*****	2000-04	*	*	FALSE
21*****	2000-04	*	*	TRUE
21*****	1999-12	*	*	TRUE

f) Give examples of suppressions and generalisations of this data set that lead to 3- 4- and 5-anonymity.

Solution (f)

There are likely multiple valid ways of achieving these goals. We list one example for each level of anonymity.

3-anonymous: suppress gender, score and grade; group student ID by first two digits and date of birth by year and month.

student_id	dob	gender	score	grade
21*****	2000-05-**	*	*	*
21*****	1999-12-**	*	*	*
21*****	2000-04-**	*	*	*
21*****	2000-05-**	*	*	*
21*****	2000-04-**	*	*	*
21*****	1999-12-**	*	*	*
21*****	2000-05-**	*	*	*
21*****	2000-04-**	*	*	*
21*****	2000-04-**	*	*	*
21*****	1999-12-**	*	*	*

4-anonymous: suppress date of birth, score and grade; group student ID by first two digits. (Note depending on how informative a missing value is this could be between 2- and 6-anonymous.)

student_id	dob	gender	score	grade
21*****	*	F	*	*
21*****	*	M	*	*
21*****	*	M	*	*
21*****	*	*	*	*
21*****	*	M	*	*
21*****	*	*	*	*
21*****	*	F	*	*
21*****	*	M	*	*
21*****	*	F	*	*
21*****	*	F	*	*

5-anonymous: depending on the interpretation of missing values in the gender attribute, the previous example may be considered 5-anonymous. An alternative 5-anonymisation is achieved by also suppressing gender. This results in a 10-anonymous database, which by definition is also 5-anonymous.

student_id	dob	gender	score	grade
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*
21*****	*	*	*	*

Question 3: Estimating prevalence of extra-marital affairs (Randomised response)

A sociology researcher is interested in estimating p, the population proportion of people in monogamous relationships who are currently or have in the past engaged in an affair. The researcher has established that her sampling frame is representative of her target population and is able to sample individuals uniformly at random from her sampling frame.

When conducting the survey the researcher has ten statement cards, which she asks the participants to shuffle and select one to secretly read. Eight of these cards are printed with statement (A) “I am currently having or have previously had an affair”, while the remaining two cards are printed with statement (B) “I have never had an affair”. The respondent then tells the researcher whether the statement that they read was “TRUE” or “FALSE”.

Let C be the event that an individual has cheated on their partner, A be the event of being shown card (A) and T be the event of responding “TRUE”. For each event E denote its compliment by E^C.

a) Draw a probability tree to describe this randomised response survey.

Solution (a)

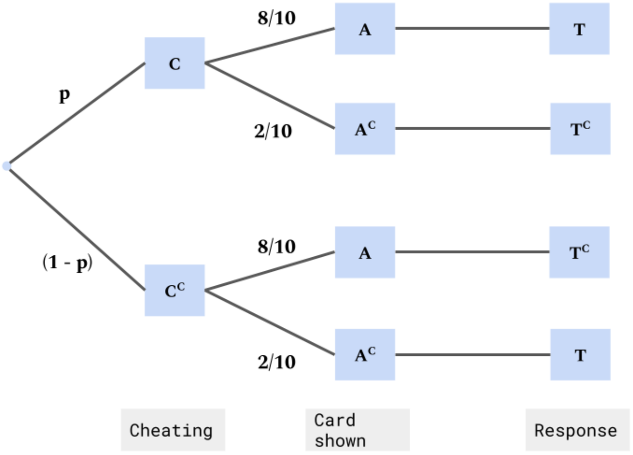

Figure 1: Probabiltiy Tree for estimating prevalence of extra-marital affairs.

b) Hence or otherwise calculate $\Pr(T)$ and $\Pr(T^C)$.

Solution (b)

From the probability tree we have that:

$$\begin{aligned}\Pr(T) &= \Pr\left((A \cap C) \cup (A^C \cap C^C)\right) \\ &= \Pr(A)\Pr(C) + \Pr(A^C)\Pr(C^C) \\ &= \frac{8}{10}p + \left(1 - \frac{8}{10}\right)(1-p) \\ &= \frac{2}{10} + \frac{6}{10}p.\end{aligned}$$

Hence,

$$\Pr(T^C) = 1 - \Pr(T) = \frac{8}{10} - \frac{6}{10}p.$$

c) Using your answer to part (b), suggest a method of moments estimator $\hat{P}$ for the population proportion of unfaithful partners, p.

Solution (c)

Rearranging our expression for $\Pr(T)$, we have that

$$\Pr(C) = p = \frac{\Pr(T) - 0.2}{0.6} = \frac{5}{3}\Pr(T) - \frac{1}{3}.$$

Let the survey responses $X_1, \dots, X_n$ take the value 1 if the individual responded "TRUE" and take the value 0 if they responded "FALSE". The sample proportion who responded "TRUE", $\hat{\Pr}(T) = n^{-1}\sum_{i=1}^n X_i$, can be used to estimate $\Pr(T)$, yielding the following estimator for p:

$$\hat{P} = \frac{5}{3}\hat{\Pr}(T) - \frac{1}{3} = \frac{5}{3n}\sum_{i=1}^{n} X_i - \frac{1}{3}.$$

d) Calculate the expectation and variance of the estimator $\hat{P}$ when using a survey with n responses. Use these to show that $\hat{P}$ is a consistent estimator of p.

Solution (d)

The responses $X_1, \dots, X_n$ can be considered as i.i.d. Bernoulli(θ) random variables where the probability of a "success" $\theta = \Pr(T) = 0.2 + 0.6p$.
It follows that $\mathbb{E}[X_i] = \theta$ and $\mathrm{Var}(X_i) = \theta(1-\theta)$ for $i = 1, \dots, n$. Therefore,

$$\begin{aligned}\mathbb{E}[\hat{P}] &= \mathbb{E}\left[\frac{5}{3n}\sum_{i=1}^{n} X_i - \frac{1}{3}\right] \\ &= \frac{5}{3n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] - \frac{1}{3} \\ &= \frac{5}{3}(\frac{1}{5} + \frac{3}{5}p) - \frac{1}{3} \\ &= p.\end{aligned}$$

So the expected value of our estimator $\hat{P}$ is the population proportion of unfaithful partners, and $\hat{P}$ is an *unbiased* estimator of p.
Similarly,

$$\begin{aligned}\mathrm{Var}(\hat{P}) &= \mathrm{Var}\left(\frac{5}{3n}\sum_{i=1}^{n} X_i - \frac{1}{3}\right) \\ &= \frac{25}{9n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= \frac{25}{9n^2}n(\frac{1}{5} + \frac{3}{5}p)(\frac{4}{5} - \frac{3}{5}p) \\ &= \frac{(1+3p)(4-3p)}{9n}.\end{aligned}$$

So the variance of the estimator $\hat{P}$ depends both on the population proportion of unfaithful partners and on the size of our sample. Since $\hat{P}$ is an unbiased estimator of p and $\mathrm{Var}(\hat{P}) \to 0$ as $n \to \infty$, it follows that $\hat{P}$ is also a *consistent* estimator of p.

e) Based on a set of 120 survey responses in which 34 respondents replied “TRUE”, calculate a point estimate $\hat{p}$ for p. Calculate the standard error of this estimate and give an approximate 95% confidence interval for p.

Solution (e)

The point estimate for p is $\hat{p} = \frac{5}{3}\frac{34}{120} - \frac{1}{3} = 13.89\%$.

The standard error of this estimate is $\mathrm{se}(\hat{p}) = \sqrt{\frac{(1+3\hat{p})(4-3\hat{p})}{9\times 120}} = 6.86\%$.
An approximate 95% confidence interval for p is therefore given by $\hat{p} \pm z_{0.975} \times \mathrm{se}(\hat{p}) = (0.45\%, 27.3\%)$. Based on the randomised response survey data the interval $(0.45\%, 27.3\%)$ covers the true proportion of unfaithful partners with 95% probability.

Question 4: The downside of randomised response

A survey company is deciding whether to use randomised or direct responses in a survey to estimate the proportion of voter who support a controversial news broadcaster.

In the suggested randomised response version of the survey, participants are asked to toss a fair coin before answering. If the coin lands heads up then they respond honestly. If the coin lands heads down then they toss the coin a second time. If on the second toss the coin lands heads up then they answer in support of the broadcaster and otherwise answer in opposition to the broadcaster.

a) Under this randomised response mechanism, derive an expression for the probability of a response against the broadcaster in terms of the true proportion of people p who are truly in favour the broadcaster.

Solution (4a)

Let A denote the event of responding against the broadcaster, while S denotes the event of truly supporting the broadcaster. Also let H be the event of the first toss landing heads up, TT the event of both tosses landing tails up and TH denote a tail on the first toss followed by a head on the second toss. Then:

$$\Pr(H) = \frac{1}{2} \quad \Pr(TT) = \Pr(TH) = \frac{1}{4},$$

and so

$$\begin{aligned}
\Pr(A) &= \Pr((S \cap TT) \cup (S^C \cap H) \cup (S^C \cap TT)) \\
&= \frac{1}{4}p + (1-p)\frac{1}{2} + (1-p)\frac{1}{4} \\
&= \frac{p}{4} + \frac{1}{2} - \frac{p}{2} + \frac{1}{4} - \frac{p}{4} \\
&= \frac{3}{4} - \frac{p}{2}.
\end{aligned}$$

b) Use your answer to the previous question to construct an estimator $\hat{P}$ for p under this randomised response scheme with n respondents. In your answer, let the response $Y_i = 1$ if the i^{th} respondent declares against the broadcaster and $Y_i = 0$ otherwise.

Solution (4b)

Denote by $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the proportion of responses in opposition to the broadcaster. Rearranging the expression for $\Pr(A)$ in the previous question we have that

$$p = \frac{3}{2} - 2\Pr(A).$$

Substituting the sample proportion against the broadcaster we obtain the method of moments estimator

$$\hat{P} = \frac{3}{2} - 2\bar{Y}.$$

c) State the method of moments estimator $\tilde{P}$ under a direct response survey with responses $Z_1, \dots, Z_n$ which take the value 1 when the respondent is against the broadcaster and the value 0 when the respondent is in support of the broadcaster.

Solution (4c)

Using a direct response survey the method of moments estimator for p is $\tilde{P} = 1 - \bar{Z}$, the sample proportion in favour of the broadcaster.

d) Calculate and compare the expectation and variance of the estimators $\tilde{P}$ and $\hat{P}$.

Solution (4d)

First we note that in in both surveys responses can be considered as independent Bernoulli trials. In the direct response survey a "success" is obtained when a respondent truly does not support the broadcaster and so $\mathbb{E}[Z_i] = (1-p)$ and $\mathrm{Var}(Z_i) = p(1-p)$. In the randomised response survey a "success" is obtained when a respondent declares that they do not support the broadcaster and so $\mathbb{E}[Y_i] = \Pr(A) = \frac{3}{4} - \frac{p}{2}$ and $\mathrm{Var}(Y_i) = (\frac{3}{4} - \frac{p}{2})(1 - \frac{3}{4} + \frac{p}{2}) = (\frac{3}{4} - \frac{p}{2})(\frac{1}{4} + \frac{p}{2})$.
It follows that for the direct response survey:

$$\mathbb{E}[\tilde{P}] = 1 - \mathbb{E}[\bar{Z}] = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Z_i] = 1 - (1-p) = p,$$

and

$$\mathrm{Var}(\tilde{P}) = \mathbb{E}[\bar{Z}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Z_i) = \frac{p(1-p)}{n}.$$

Similarly, for the randomised response survey:

$$\mathbb{E}[\hat{P}] = \frac{3}{2} - 2\mathbb{E}[\bar{Y}] = \frac{3}{2} - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{3}{2} - \frac{2}{n}\sum_{i=1}^{n}\left(\frac{3}{4} - \frac{p}{2}\right) = \frac{3}{2} - \frac{3}{2} + p = p,$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{P}) &= \mathrm{Var}\left(\frac{3}{2} - 2\bar{Y}\right) \\
&= \frac{4}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) \\
&= \frac{4}{n^2}\sum_{i=1}^{n} \left\{\left(\frac{3}{4} - \frac{p}{2}\right)\left(\frac{1}{4} + \frac{p}{2}\right)\right\} \\
&= \frac{(3-2p)(1+2p)}{4n}.
\end{aligned}$$

e) Describe why randomised response is used for sensitive questions but not in all survey responses. You should link your answer to your findings in part (d).

Solution (4e)

From part (d) we can see that both estimators are unbiased, as their expectations are equal to the estimand p. However, the variance of the randomised response estimator is always greater than that of the direct response survey, no matter the true value of p or the sample size n (See plots below). An estimator based on direct responses results in more precise inference than when using randomised response. This is because information is lost due to the randomisation.

When survey questions relate to sensitive properties, randomised response can be used to provide plausible deniability to the responses of individual participants but this comes at the cost of less precise estimation of the quantity of interest (as compared to if direct responses were collected). To make surveys as informative as possible, randomised responses should only be used for questions relating to sensitive properties.

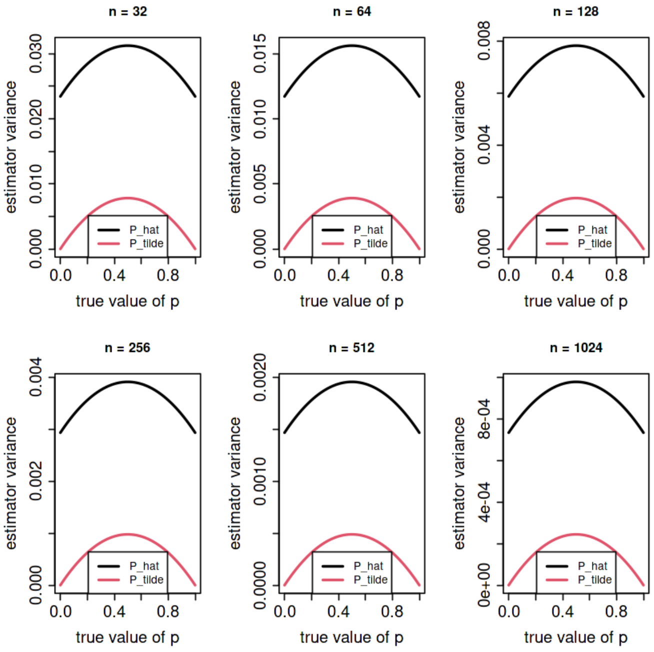

Figure 2: Plots of estimator variance as a function of p for increasing values of n. In each case the variance of P_hat is greater than that of P_tilde.