Consider table T, depicted in Table 1. We will use table T throughout this problem set. The attribute Drinks/Day indicates the average number of alcoholic drinks consumed per day by the individual, counted only on days whenever she or he consumes alcohol. The attribute Hard Drugs indicates whether or not the individual has ever used cocaine, crack cocaine, heroin or methamphetamine in her or his life.

Sex	Age	Marital Status	Birth Country	Race	Drinks/Day	Hard Drugs
M	30	Married	US	Non-Hispanic White	2	No
M	30	Married	US	Non-Hispanic White	4	No
M	30	Married	Other	Mexican American	1	No
M	30	Married	Other	Non-Hispanic Asian	1	No
M	30	Widowed	US	Mexican American	2	No
M	30	Never Married	US	Non-Hispanic White	1	No
M	30	Never Married	US	Non-Hispanic White	2	No
M	30	Never Married	US	Non-Hispanic White	8	No
M	30	Never Married	US	Non-Hispanic Black	3	No
M	30	Living W/ Partner	Other	Mexican American	3	Yes
M	31	Married	US	Other Hispanic	2	Yes
M	31	Married	US	Non-Hispanic White	1	No
M	31	Married	US	Non-Hispanic White	2	Yes
M	31	Married	US	Non-Hispanic White	3	No
M	31	Married	US	Non-Hispanic White	14	No
M	31	Married	US	Non-Hispanic Black	4	No
M	31	Married	US	Non-Hispanic Asian	6	No
M	31	Married	Other	Other Hispanic	2	No
M	31	Married	Other	Non-Hispanic Asian	2	No
M	31	Divorced	US	Non-Hispanic White	2	No
M	31	Divorced	Other	Other Hispanic	8	No
M	31	Never Married	US	Non-Hispanic White	3	Yes
M	31	Never Married	US	Non-Hispanic White	4	No
M	31	Never Married	US	Non-Hispanic White	6	Yes
M	31	Never Married	Other	Non-Hispanic Black	1	No
M	31	Living W/ Partner	US	Other	3	No

M	31	Living W/ Partner	Other	Other Hispanic	8	Yes
M	32	Married	US	Other Hispanic	2	Yes
M	32	Married	US	Other Hispanic	3	No
M	32	Married	US	Non-Hispanic White	2	Yes
M	32	Married	Other	Mexican American	2	No
M	32	Married	Other	Non-Hispanic White	3	Yes
M	32	Never Married	US	Non-Hispanic White	1	No
M	32	Never Married	US	Non-Hispanic White	2	No
M	32	Never Married	Other	Non-Hispanic Asian	2	No
M	32	Living W/ Partner	US	Non-Hispanic White	2	No
M	32	Living W/ Partner	US	Non-Hispanic Black	2	No
F	30	Married	US	Mexican American	3	No
F	30	Married	US	Non-Hispanic White	1	No
F	30	Married	US	Non-Hispanic White	2	Yes
F	30	Married	US	Non-Hispanic White	2	No
F	30	Married	US	Non-Hispanic White	3	No
F	30	Married	Other	Other Hispanic	3	No
F	30	Married	Other	Non-Hispanic Asian	1	No
F	30	Married	Other	Non-Hispanic Asian	1	No
F	30	Married	Other	Non-Hispanic Asian	3	No
F	30	Divorced	US	Non-Hispanic Black	2	No
F	30	Separated	US	Other Hispanic	2	No
F	30	Never Married	US	Mexican American	4	No
F	30	Never Married	US	Non-Hispanic Black	1	No
F	30	Never Married	US	Non-Hispanic Black	2	No
F	30	Never Married	US	Non-Hispanic Black	4	No
F	30	Never Married	US	Other	3	No
F	30	Never Married	Other	Non-Hispanic Black	2	No
F	30	Never Married	Other	Other	2	Yes
F	30	Never Married	Other	Other	4	No
F	31	Married	US	Mexican American	1	No
F	31	Married	US	Mexican American	2	No
F	31	Married	US	Mexican American	7	No
F	31	Married	US	Non-Hispanic White	2	No

F	31	Married	US	Non-Hispanic White	2	No
F	31	Married	US	Non-Hispanic White	3	No
F	31	Married	US	Non-Hispanic Black	1	No
F	31	Divorced	US	Other Hispanic	1	No
F	31	Never Married	US	Non-Hispanic White	1	No
F	31	Never Married	US	Non-Hispanic Black	1	No
F	31	Living W/ Partner	US	Mexican American	2	No
F	32	Married	US	Mexican American	8	No
F	32	Married	US	Other Hispanic	1	No
F	32	Married	US	Non-Hispanic Black	1	No
F	32	Married	Other	Mexican American	1	No
F	32	Married	Other	Non-Hispanic White	1	No
F	32	Divorced	Other	Non-Hispanic Asian	1	No
F	32	Never Married	US	Other Hispanic	1	No
F	32	Never Married	US	Non-Hispanic White	2	No
F	32	Never Married	US	Non-Hispanic Black	2	Yes
F	32	Never Married	Other	Other Hispanic	4	No
F	32	Living W/ Partner	US	Non-Hispanic White	2	No
F	32	Living W/ Partner	US	Non-Hispanic Black	1	No
F	32	Living W/ Partner	Other	Mexican American	4	No

Table 1: Table T

Problem 1 (The Basics)

(a) Which of the following disclosure risks is k-Anonymity best designed to protect against, and which of the following disclosure risks is ℓ-Diversity best designed to protect against? (Select one for each privacy model)

- Identity disclosure;
- Attribute disclosure; or
- Membership disclosure.

Explain your answer.

(b) What is the set of attributes in table T?

(c) What is the largest k for which table T satisfies k-Anonymity with respect to the set of quasi-identifiers $\{\text{Sex}\}$? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that has k records.

(d) What is the largest k for which table T satisfies k-Anonymity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that has k records.

(e) What is the largest k for which table T satisfies k-Anonymity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Marital Status}\}$? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that has k records.

(f) What is the largest k for which table T satisfies k-Anonymity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Birth Country}\}$? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that has k records.

(g) What is the largest k for which table T satisfies k-Anonymity with respect to the set of quasi-identifiers $\{\text{Birth Country}, \text{Race}\}$? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that has k records.

Problem 2 (General Properties)

(a) (True/False)

Let P be a table and let P_1 be a generalized version of table P. If P_1 doesn't satisfies 3-Anonymity with respect to a set of quasi-identifiers Q and sensitive attribute S, then P doesn't satisfy 3-Anonymity with respect to quasi-identifiers Q and sensitive attribute S.

If you answer "true," justify why; otherwise, provide a counter example consisting of tables P and P_1 (clearly identify the quasi-identifiers).

(b) (True/False)

Let P be a table and let Q be a set of quasi-identifiers. Let Q_1 be a subset of Q. If P doesn't satisfy Entropy 4-Diversity with respect to the set of quasi-identifiers Q and some sensitive attribute S, then P doesn't satisfy Entropy 4-Diversity with respect to quasi-identifiers Q_1 and sensitive attribute S.

If you answer "true," justify why; otherwise, provide a counter example consisting of a table P, and sets of quasi identifiers Q and $Q_1 \subset Q$.

(c) Let c_1 and c_2 be real numbers such that $c_1 < c_2$. Let ℓ be an integer. Let T_1 and T_2 be generalizations of table T. You are given that T_1 satisfies recursive (c_1, ℓ)-Diversity (and c_1 is the best such guarantee for this ℓ) and T_2 satisfies recursive (c_2, ℓ)-Diversity (and c_2 is the best such guarantee for this ℓ).

Between T_1 and T_2, which table has a better guarantee in terms of (c, ℓ)-Diversity?

Problem 3 (Different Disclosures)

We will now investigate the different possible disclosures using table T.

(a) List all possible records that can correspond to Bob, a 32 year-old man who is non-Hispanic white (thereafter called Bob's demographics)?

(b) Based on the records in Bob's equivalence class identified in the previous part (with respect to quasi-identifiers $\{\text{Age}, \text{Sex}, \text{Race}\}$), what is the probability that you can correctly identify Bob's exact record (assuming we have no additional information about Bob)?

(c) Based on the records in Bob's equivalence class (with respect to quasi-identifiers $\{\text{Age}, \text{Sex}, \text{Race}\}$), what is the probability that Bob's record would indicate 'Hard Drugs' = 'Yes'?

(d) What is the probability that a random person in the dataset has ever used hard drugs in her or his life (i.e., 'Hard Drugs' = 'Yes')? Did knowing Bob's demographics improve our ability to infer this sensitive piece of information about him, compared to a random person from the dataset?

(e) Can you uniquely identify the record of Alex, a 32 year-old man who was born in the US?

If not, what is the probability that you uniquely identify Alex's record in the dataset?

Is this consistent with the k-Anonymity guarantee on the table T with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Birth Country}\}$ (from Problem 1 Part (f))?

(f) If you know, in addition to Alex's demographic information from the previous part, that Alex has used hard drugs in his life, can you uniquely infer how many alcoholic drinks Alex consumes per day?

(g) Is k-Anonymity equipped to deal with inferences made using auxiliary information such as the one presented in the previous part?

(h) Your friend Eve argues that k-Anonymity is not equipped to prevent attribute disclosure. She claims that even a 2-Anonymous table may be used to uniquely identify a sensitive attribute about individuals, only using their quasi-identifiers (i.e., without auxiliary information). Is Eve right or wrong?

If Eve is wrong, argue why.

If Eve is right, provide an example of an equivalence class from table T, with quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Birth Country}\}$, that has more than 1 record; but all of them agree on the value of at least one sensitive attribute (at least one of the attributes Drinks/Day or Hard Drugs).

Problem 4 (ℓ-Diversity)

(a) What is the largest ℓ_d for which table T satisfies distinct ℓ_d-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies distinct ℓ_d-Diversity but not any distinct $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_d$.

What is the largest ℓ_e for which table T satisfies entropy ℓ_e-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies entropy ℓ_e-Diversity but not any entropy $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_e$.

(b) What is the largest ℓ_d for which table T satisfies distinct ℓ_d-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies distinct ℓ_d-Diversity but not any distinct $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_d$.

What is the largest ℓ_e for which table T satisfies entropy ℓ_e-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies entropy ℓ_e-Diversity but not any entropy $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_e$.

(c) What is the largest ℓ_d for which table T satisfies distinct ℓ_d-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Marital Status}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies distinct ℓ_d-Diversity but not any distinct $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_d$.

What is the largest ℓ_e for which table T satisfies entropy ℓ_e-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Marital Status}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies entropy ℓ_e-Diversity but not any entropy $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_e$.

(d) What is the largest ℓ_d for which table T satisfies distinct ℓ_d-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Marital Status}, \text{Birth Country}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies distinct ℓ_d-Diversity but not any distinct $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_d$.

What is the largest ℓ_e for which table T satisfies entropy ℓ_e-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Marital Status}, \text{Birth Country}\}$ and sensitive attribute Drinks/Day? Give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies entropy ℓ_e-Diversity but not any entropy $\bar{\ell}$-Diversity with $\bar{\ell} > \ell_e$.

(e) What is the smallest c for which table T satisfies recursive $(c, 2)$-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day?

If no such value c exists, explain why!

Otherwise, give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies recursive $(c, 2)$-Diversity but not recursive $(\hat{c}, 2)$-Diversity for any $\hat{c} < c$.

(f) What is the smallest c for which table T satisfies recursive $(c, 3)$-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day?

If no such value c exists, explain why!

Otherwise, give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies recursive $(c, 3)$-Diversity but not recursive $(\hat{c}, 3)$-Diversity for any $\hat{c} < c$.

(g) What is the smallest c for which table T satisfies recursive $(c, 4)$-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day?

If no such value c exists, explain why!

Otherwise, give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies recursive $(c, 4)$-Diversity but not recursive $(\hat{c}, 4)$-Diversity for any $\hat{c} < c$.

(h) What is the largest ℓ for which table T satisfies recursive $(2.5, \ell)$-Diversity with respect to the set of quasi-identifiers $\{\text{Sex}, \text{Age}\}$ and sensitive attribute Drinks/Day?

If no such value ℓ exists, explain why!

Otherwise, give an example of an equivalence class (in terms of the values of its quasi-identifiers) that satisfies recursive $(2.5, \ell)$-Diversity but not recursive $(2.5, \bar{\ell})$-Diversity with any $\bar{\ell} > \ell$.

(i) In this problem, you are asked to generalize the table T so that you achieve the highest possible level of entropy ℓ-Diversity possible of any generalization for table T with respect to quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Marital Status}, \text{Birth Country}, \text{Race}\}$ sensitive attribute Drinks/Day.

In other words, if we denote the highest level of entropy ℓ-Diversity of your generalization by ℓ^*, then no other generalization of table T can satisfy any entropy ℓ-Diversity with $\ell > \ell^*$ (with respect to quasi-identifiers $\{\text{Sex}, \text{Age}, \text{Marital Status}, \text{Birth Country}, \text{Race}\}$ sensitive attribute Drinks/Day)

Describe your generalization and report the highest level of entropy ℓ-Diversity it yields (ℓ^*).

Finally, justify why your generalization yields the highest possible level of entropy ℓ-Diversity for table T.

Problem 5 (Anonymize (Optional))

In this problem, you will anonymize the table T with respect to the quasi-identifiers

$$\{\text{Sex}, \text{Age}, \text{Marital Status}, \text{Birth Country}, \text{Race}\}.$$

You are allowed to use any generalization (single-dimensional or multi-dimensional) you'd like as long as it is clearly defined. For example, you may generalize all values of the attribute Sex to "Person." As a second example, you may generalize all values of Marital Status that are not equal to "Married" to "Not Married."

The goal is to get the table to be 3-Anonymous but not 6-Anonymous. After you perform the generalization to achieve that, be sure to:

- Describe the generalization(s) performed;
- Explicitly write the resulting table after generalization (you may omit the values of the sensitive attributes in your write-up);
- Visually group the different equivalence classes after generalization; and
- Indicate the number of records in each one of the equivalence classes after generalization.

Problem 6 (Acknowledgments)

List all individuals and sources that you consulted with while working on this homework.