CYBER 759 – Assignment 5
Gunnar Yonker

1.

ε=0.5

Maximum Bounds:

$$\frac{Pr[output\ o]|Eve\ is\ in\ the\ dataset]}{\Pr[output\ o]} \leq \frac{1}{0.7 + 0.3 * \frac{1}{e^{\in}}}$$

$$\leq \frac{0.7}{0.7 + 0.3 * \frac{1}{1.6487}} = 0.7936$$

Minimum Bounds:

$$\frac{Pr[output\ o]|Eve\ is\ in\ the\ dataset]}{\Pr[output\ o]} \geq \frac{1}{0.7 + 0.3 * e^{\in}}$$

$$\geq \frac{1}{0.7 + 0.3 * 1.6487} = 0.5859$$

The attacker's suspicion that Eve is in the dataset can lie between 0.5859 (minimum) and 0.7936 (maximum).

2.

First Scenario:

Given:

ε_A = 0.3 (for Query A)

ε_B = 0.4 (for Query B)

Calculate:

$\varepsilon_{Total} = \varepsilon_A + \varepsilon_B$

$\varepsilon_{Total} = 0.3 + 0.4 = 0.7$

ε_{Total} is 0.7 for the sequential queries on the same dataset.

Second Scenario:

Given:

ε_A = 0.3 (for Query A on D1)

CYBER 759 – Assignment 5
Gunnar Yonker

ε_B = 0.4 (for Query B on D2)

The overall privacy parameter, ε_{Total}, when both queries are executed is the maximum of the two, since the worst-case scenario determines the overall privacy level.

$\varepsilon_{Total} = \max(\varepsilon_A, \varepsilon_B)$

$\varepsilon_{Total} = \max(0.3, 0.4) = 0.4$

ε_{Total} is 0.4 when the queries are executed on non-overlapping subsets.

3.

K-anonymity:

Advantages:

Easy to Understand: It's relatively simple to understand and apply. If every person in the dataset can't be distinguished from at least k-1 other persons based on the quasi-identifiers, the dataset is said to have k-anonymity.

Generalization and Suppression: K-anonymity often uses generalization (replacing a value with a broader category) and suppression (removing a value) to anonymize data, which can be intuitive ways to mask data.

Direct Control: Data custodians can directly control the level of anonymity by choosing the value of 'k'.

Disadvantages:

L-Diversity Issue: Even if a dataset is k-anonymous, it may still leak information if the sensitive attributes are not well-represented (i.e., all k rows have the same sensitive value). This issue led to extensions like l-diversity.

Data Utility: To achieve k-anonymity, especially for high values of k, the data often needs to be overly generalized or many rows need to be suppressed, reducing the utility of the data.

Homogeneity Attack: If all k records have the same sensitive attribute value, then the sensitive value can be inferred even if the exact identity is hidden.

Differential Privacy:

Advantages:

Mathematical Guarantee: Offers a robust and rigorous definition of privacy. If a mechanism is differentially private, then it mathematically ensures privacy against a wide range of potential attacks.

Future-Proof: Protects against future re-identification attacks, even if additional external data becomes available.

Flexibility: Applicable to a variety of data analysis tasks and mechanisms, not just re-identification.

Disadvantages:

Complexity: Differential privacy concepts can be difficult to understand and implement correctly.

Noise Addition: To achieve differential privacy, noise is often added to the results of queries, which can reduce the utility of the data or the results, especially if the dataset is small or if strong privacy guarantees (low epsilon values) are desired.

Parameter Choice: The choice of the parameter epsilon which dictates the privacy level is not always intuitive and can be challenging for practitioners.

Comparison:

Purpose: K-anonymity is focused on de-identification, ensuring that records cannot be pinpointed to fewer than k individuals. Differential privacy focuses on ensuring the results of queries on a database remain private, often by adding noise.

Utility vs Privacy Trade-off: Both involve a trade-off between data utility and privacy. However, while k-anonymity achieves this through generalization and suppression, differential privacy typically achieves this through noise addition.

Robustness: Differential privacy provides stronger and more robust privacy guarantees than k-anonymity, especially against sophisticated attacks or when auxiliary information is available.

CYBER 759 – Assignment 5
Gunnar Yonker

4.

k = 2 – at least 2 records that are indistinguishable based on these identifiers.

Age: Group ages into ranges.

Role: Group roles by their general type or hierarchy.

Zip code: Only use the first 3 digits.

Patient ID	Age	Role	Diagnosis Code	Zip code
001	40–49	Managerial	D123	891**
002	70-79	Top Management	D456	573**
003	20-29	Marketing Related	D789	781**
004	40-49	Managerial	D123	661**
005	30-39	Managerial	D456	781**
006	60-69	Top Management	D123	414**
007	40-49	Marketing Related	D789	291**
008	70-79	Top Management	D123	273**
009	60-69	Top Management	D456	414**

k = 3 – further generalization to ensure that each combination has at least 3 records.

Age: Increase the range of ages.

Role: More general categories.

Zip code: Still use the first 3 digits.

Patient ID	Age	Role	Diagnosis Code	Zip code
001	20-41	Managerial	D123	891**
002	63-79	Top Management	D456	573**
003	20-41	Managerial	D789	781**
004	42-62	Managerial	D123	661**
005	20-41	Managerial	D456	781**
006	63-79	Top Management	D123	414**
007	42-62	Managerial	D789	291**
008	63-79	Top Management	D123	273**
009	42-69	Top Management	D456	414**

CYBER 759 – Assignment 5
Gunnar Yonker

K = 4 – more generalization is needed to ensure that each combination has at least 4 records.

Age: Further increase the age ranges.

Role: Broad Categories.

Zip code: Use only the first 2 digits.

Patient ID	Age	Role	Diagnosis Code	Zip code
001	20-49	Mid-Level Position	D123	89***
002	60-79	Top-Level Position	D456	57***
003	20-49	Mid-Level Position	D789	78***
004	20-49	Mid-Level Position	D123	66***
005	20-49	Mid-Level Position	D456	78***
006	60-79	Top-Level Position	D123	41***
007	20-49	Mid-Level Position	D789	29***
008	60-79	Top-Level Position	D123	27***
009	60-79	Top-Level Position	D456	41***