1 Privacy-preserving Data Publishing

1.1. One can identify 4 basic anonymization operations: generalization, suppression, anatomization and perturbation. Anatomization consists on de-associating QIDs and sensitive attributes. Explain what is the advantage and disadvantage of anatomization?

Anatomization does not modify the quasi-identifier or the sensitive attribute, but de-associates the relationship between the two. In other words, this method releases the data on QID and the data on the sensitive attribute in two separate tables: a quasi-identifier table (QIT) contains the QID attributes, a sensitive table (ST) contains the sensitive attributes, and both QIT and ST have one common attribute, $GroupID$. All records in the same group will have the same value on $GroupID$ in both tables and, therefore, are linked to the sensitive values in the group in the exact same way (Fig. 1).

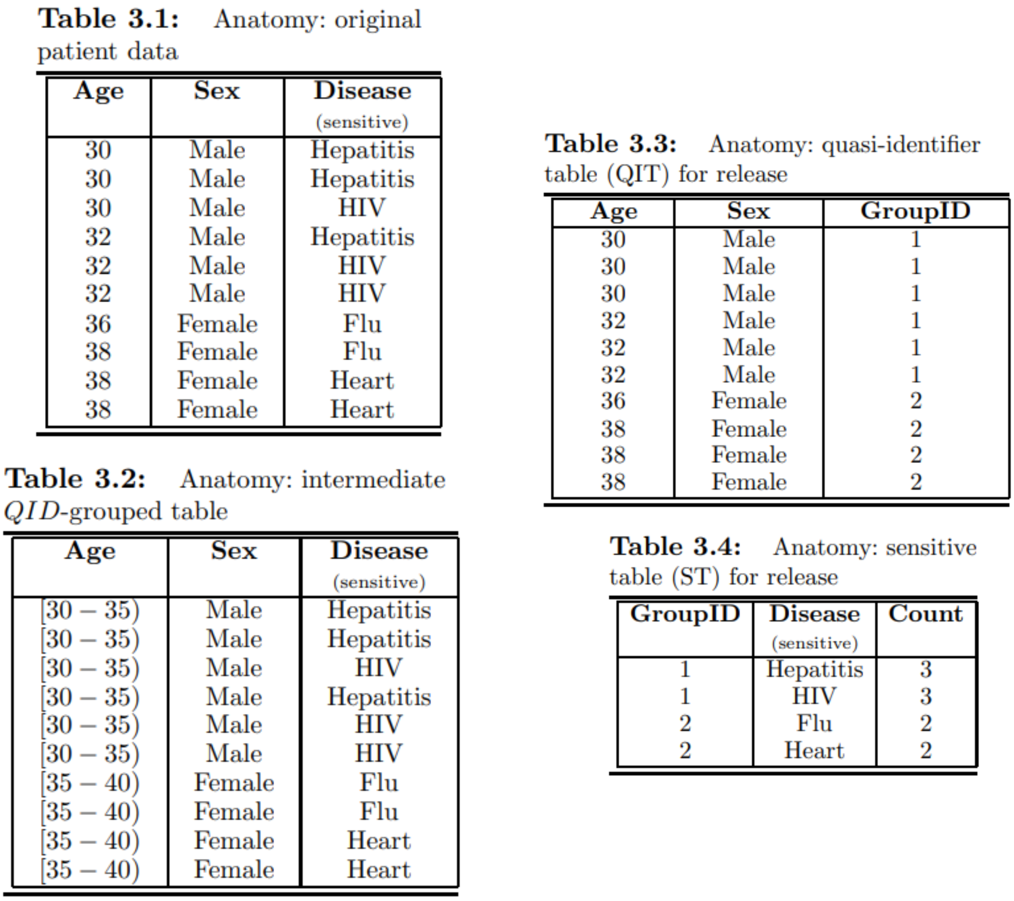

Table 3.1: Anatomy: original patient data

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

Table 3.2: Anatomy: intermediate QID-grouped table

Age	Sex	Disease (sensitive)
[30 – 35)	Male	Hepatitis
[30 – 35)	Male	Hepatitis
[30 – 35)	Male	HIV
[30 – 35)	Male	Hepatitis
[30 – 35)	Male	HIV
[30 – 35)	Male	HIV
[35 – 40)	Female	Flu
[35 – 40)	Female	Flu
[35 – 40)	Female	Heart
[35 – 40)	Female	Heart

Table 3.3: Anatomy: quasi-identifier table (QIT) for release

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

Table 3.4: Anatomy: sensitive table (ST) for release

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

Figure 1: Anatomization

Advantages: The data in both QIT and ST are unmodified (unlike what happens when we consider anonymization operations such as generalization or suppression). The anatomized tables can more accurately answer aggregate queries involving domain values of the QID and sensitive attributes than the generalization approach. The intuition is that in a generalized table domain values are lost and, without additional knowledge, it is more difficult to answer a query about domain values.

Disadvantages: With the data published in two tables, it is unclear how standard data mining tools (*e.g.* classification and clustering data mining tools) can be applied to the published data, and new tools and algorithms need to be designed. Also, anatomization is not suitable for incremental data publishing. The generalization approach does not suffer from the same problem because all attributes are released in the same table.

1.2. Differential privacy decreases the risk of an individual/register joining or leaving the database.

2 Secure Multiparty Computation and Privacy

2.1. In SMC two or more parties wish to jointly compute a function of their inputs while preserving certain security properties, such as privacy, correctness and independence of inputs. Considering the auction example, where users bid for a product, explain what privacy means in this context.

Protocols for Secure Multiparty Computation (SMC) enable a set of parties to compute a joint function of their private inputs while revealing nothing but the output. In other words, the aim of secure multiparty computation is to enable parties to carry out distributed computing tasks in a secure manner.

In this scenario, it is assumed that a protocol execution may be attacked by an external entity, or even by a subset of the participating parties. The aim of this attack may be to learn private information or cause the result of the computation to be incorrect. Thus, two important requirements on any secure computation protocol are **privacy** and **correctness**. The privacy requirement states that nothing should be learned beyond what is absolutely necessary, *i.e.*, parties should learn their output and nothing else. The correctness requirement states that each party should receive its correct output. Therefore, the adversary must not be able to cause the result of the computation to deviate from the function that the parties had set out to compute.

In the auction example, we consider a trading platform where parties provide offers and bids, and are matched whenever an offer is greater than a bid. In such a scenario, it can be beneficial to not reveal the parties' actual offers and bids since this information can be used by others in order to artificially raise prices.

In this scenario, the meaning of privacy and correctness is as follows:

- **Privacy**: No party should learn anything more than its prescribed output. In particular, the only information that should be learned about other parties' inputs is what can be derived from the output itself. In this case, the only bid revealed is that of the highest bidder, which makes it possible to infer that all other bids were lower than the winning bid. However, nothing else should be revealed about the losing bids.
- **Correctness**: Each party is guaranteed that the output that it receives is correct. In this case, this implies that the party with the highest bid is guaranteed to win, and no party including the auctioneer can influence this.

Other relevant properties of SMC include:

- **Independence of Inputs**: Corrupted parties must choose their inputs independently of the honest parties' inputs. This property is crucial in an auction, where bids are kept secret and parties must fix their bids independently of others. Note that independence of inputs is not implied by privacy. For example, it may be possible to generate a higher bid, without knowing the value of the original one. Such an attack can actually be carried out on some encryption schemes (*i.e.*, given an encryption of $100, it is possible to generate a valid encryption of $101, without knowing the original encrypted value).
- **Guaranteed Output Delivery**: Corrupted parties should not be able to prevent honest parties from receiving their output. In other words, the adversary should not be able to disrupt the computation.
- **Fairness**: Corrupted parties should receive their outputs if and only if the honest parties also receive their outputs. The scenario where a corrupted party obtains output and an honest party does not should not be allowed to occur. Note that guaranteed output delivery implies fairness, but the converse is not necessarily true.

3 Authentication Protocols and Anonymous Authentication

3.1. Encryption and signing/authentication should be done with two different public-key cryptography key pairs for each.