Module 6: Homework 6

Harsh Manishbhai Siddhapura

Vaibhavi Nitin Honagekar

Arunava Lahiri

Thembelihle Shongwe

Sai Shashank Nagavaram

Anandha Krishnan Senthooraan

Group: Class96958 4

Ira A. Fulton School of Engineering, Arizona State University

IFT 520: Advanced Information Systems Security

Prof. Upakar Bhatta

September 25, 2023

Review Questions

1. **List and define the four types of attributes that may be present in a microdata file.**

   In the context of microdata files, which typically contain individual-level data, there are four types of attributes that may be present:

   - **Identifier Attributes:** Identifier attributes are unique identifiers assigned to each individual in the dataset. These identifiers are used to distinguish one individual from another and are typically alphanumeric values or numbers. They serve as the primary key for linking different pieces of information about the same individual across multiple datasets or tables. For example, a Social Security Number (SSN) or a unique customer ID in a marketing database can be considered identifier attributes.
   - **Categorical Attributes:** Categorical attributes, also known as nominal attributes, represent discrete categories or labels that describe characteristics of individuals. These attributes have no inherent order or numeric value associated with them. Examples of categorical attributes include gender (e.g., male, female, non-binary), marital status (e.g., single, married, divorced), or employment status (e.g., employed, unemployed, student).
   - **Ordinal Attributes:** Ordinal attributes represent ordered categories where there is a meaningful sequence or ranking among the values, but the differences between the values are not necessarily uniform or well-defined. These attributes convey relative information but do not specify precise measurement intervals. Examples include educational attainment (e.g., high school, bachelor's degree, master's degree) or customer satisfaction ratings (e.g., very satisfied, somewhat satisfied, neutral, somewhat dissatisfied, very dissatisfied).
   - **Numerical Attributes:** Numerical attributes, also referred to as continuous attributes, represent measurable quantities that can take on a wide range of numeric values. These attributes are used for quantitative analysis and can be subject to mathematical operations like addition, subtraction,

multiplication, and division. Examples of numerical attributes include age, income, height, and temperature.

It's important to recognize and differentiate these attribute types when working with microdata, as they influence the types of statistical analyses and data processing techniques that can be applied to the dataset. Properly understanding and handling these attributes is crucial for extracting meaningful insights from microdata files and ensuring the accuracy of analyses and modeling efforts.

2. **Explain the differences between microdata tables, frequency tables, and magnitude tables.**

Microdata tables, frequency tables, and magnitude tables are three distinct types of data tables used in statistics and data analysis. Each serves a specific purpose and provides different types of information:

- Microdata Tables:
    - Individual-Level Data: Microdata tables contain individual-level data, where each row represents a single individual or unit of analysis (e.g., a person, a household, a customer).
    - Detailed Information: Microdata tables provide detailed information about each individual, including various attributes or variables (e.g., age, income, education) for each case.
    - Granular Analysis: These tables are suitable for conducting granular and individual-level analyses, such as calculating means, medians, percentiles, or performing regression analysis.
- Frequency Tables:
    - Aggregated Counts: Frequency tables summarize data by showing the frequency or count of occurrences of each unique value or category within a single variable.
    - Categorical Data: They are commonly used for categorical or nominal data, where the focus is on understanding the distribution of categories.
    - Summary Statistics: Frequency tables provide insights into the distribution of data, enabling you to see which categories are most common or rare.

- Example: A frequency table for a survey question about favorite colors might show how many respondents chose each color option.

- Magnitude Tables:
  - Aggregated Summaries: Magnitude tables aggregate data to provide summaries that are not just counts but also include quantitative values (e.g., totals, sums, averages).
  - Numeric Data: They are typically used for numerical or continuous data, where the focus is on summarizing and analyzing the magnitudes of values.
  - Statistical Analysis: Magnitude tables are often used for aggregating data across different groups or categories to calculate sums, means, or other statistical measures.
  - Example: A magnitude table for sales data might show the total sales revenue for each product category.

In summary, microdata tables contain individual-level data with detailed information for each case, making them suitable for granular analysis. Frequency tables summarize the counts or frequencies of categories within a single variable and are used for categorical data. Magnitude tables provide aggregated summaries, including quantitative values, and are commonly used for numerical data. The choice of table type depends on the specific analysis and the nature of the data being examined.

3. What is a re-identification attack?

A re-identification attack, also known as a re-ID attack or anonymity attack, is a type of cybersecurity threat in which an attacker attempts to identify and associate an apparently anonymous or fake user about data and a particular person or business Many are associated with data privacy concerns, particularly when anonymous or partially encrypted to conceal people's information.

- **Data Anonymization:** Before sharing or analyzing data, organizations typically register by deleting or hiding identifiable data (PII) such as names, addresses, and social insurance numbers The goal is to create privacy, protect information and still allow data analysis. However, attackers may attempt to manipulate or "anonymize" the data in order to identify individuals.

- **Ancillary Information:** Attackers use ancillary information, whether public or from other sources, to communicate with anonymous sources. This supporting information may include information about demographics, spatial data, or other characteristics that may be associated with individuals.
- **Statistical Analysis:** Attackers often use machine learning and advanced statistical techniques to find patterns and relationships in anonymous data and supporting information By identifying specific patterns or combinations of traits they can make meaningful predictions about the individuals in anonymous data.
- **Anonymization:** If an attacker can successfully identify specific individuals in anonymized data, that individual has effectively been "anonymized," jeopardizing the privacy of all involved.

4. List and define three types of disclosure risk.

The risk of accidentally disclosing sensitive or confidential information, which may jeopardize privacy and security, is called exposure risk Here are three definitions of exposure risk:

- **Identity Risk:** Identity risk occurs when a person’s true identity is revealed or inferred from a data set, document, or conversation, despite anonymity efforts or falsifying names of data This threat is often associated with repeat attacks. anonymise the data.
- **Exposure Risk:** This risk refers to the possibility of specific traits or sensitive aspects of an individual and not their full personality. For example, disclosing a person's health condition or medical history without disclosing their name in a medical database can still be a serious invasion of privacy. Feature disclosure can lead to privacy. There have been many abuses, such as stigmatization or discrimination.
- **Theoretical Disclosure Risk:** An adversary faces the risk of theoretical disclosure when he can make accurate inferences about individuals or their activities based on the information provided. Even if specific attributes or identifying features are not revealed, the attacker can extrapolate important information from the data by examining patterns, associations, or statistical relationships.

Organizations and data controllers use a variety of privacy protection strategies to mitigate these disclosure risks, such as data anonymization, differential privacy, data reduction so and in addition, legal and regulatory frameworks often require companies to assess and mitigate the risks of disclosure while maintaining the confidentiality of sensitive data that individuals To protect and ensure compliance with data security regulations.

5. Explain the differences between pseudonymization, anonymization, and de-identification.

- Pseudonymization: Pseudonymization is a data privacy technique that involves the substitution or concealment of personally identifiable information (PII) with pseudonyms or codes. While the original data remains intact, it undergoes transformation into a format that doesn't readily reveal its identity. Typically, this process is reversible, allowing the original data to be re-identified through a designated key or mapping system. Pseudonymization is often employed when there's a need to associate data with individuals for specific purposes, such as medical research or customer analysis, while still preserving a level of privacy.
- Anonymization: Anonymization represents a stringent data privacy approach that aims to render data entirely unidentifiable. It entails the alteration or removal of information to such an extent that any linkage between the data and individuals becomes infeasible, even when supplemented with additional information or contextual cues. The anonymization process is typically irreversible, meaning that once data has been anonymized, its restoration to its original state is a complex task. This technique is employed when the utmost level of individual privacy protection is required, particularly in scenarios involving publicly disseminated datasets or situations where re-identification must be steadfastly prevented.
- De-identification: De-identification is a comprehensive data privacy concept encompassing both pseudonymization and anonymization. It involves modifying or eliminating personally identifiable information (PII) within a dataset with the aim of reducing the potential for identifying individuals. The extent of protection and whether it remains reversible or irreversible may vary when dealing with de-identified data. The choice of de-identification techniques

hinges on the specific prerequisites and usage scenarios. De-identification becomes instrumental when the objective is to safeguard privacy while still permitting specific and limited uses of the data, with the degree of de-identification tailored to meet the privacy and utility requisites of a particular application [5].

6. List and define approaches to privacy-preserving data publishing.

- Differential Privacy: Differential privacy serves as a robust privacy framework that introduces controlled randomness or noise into query responses. The magnitude of this noise is carefully adjusted to ensure that, even with access to the published data, it remains difficult for anyone to confidently ascertain the presence or absence of any individual's data. This approach provides robust privacy protection while still permitting meaningful statistical analyses.
- Data Masking or Redaction: Data masking or redaction involves modifying or omitting specific data elements, such as names, addresses, or social security numbers, which could directly lead to the identification of individuals. This process shields sensitive information by either concealing it or eliminating it from the dataset.
- Generalization and Suppression: Generalization encompasses the categorization of data into broader groups, like age brackets, to reduce its granularity. Suppression entails the removal of data points that might potentially expose individuals to re-identification. These techniques strike a balance between data utility and privacy by providing less detailed yet still valuable data.
- Data Perturbation: Data perturbation introduces controlled noise or randomness into individual data points, making it more challenging for potential adversaries to identify specific individuals. Through careful calibration of noise levels, privacy is upheld while preserving the dataset's analytical value [2].

7. Explain the differences between k-anonymity, [-diversity, and t-closeness.

K Anonymity	L - Diversity	T -Closeness
K-anonymity is a privacy model that seeks to ensure that every record in the dataset is distinguishable from at least k-1 additional records so as to prevent the re-identification of individuals in a dataset.	L-diversity is an adaptation of K-anonymity that aims at enhancing privacy through making sure that any combination of quasi-identifiers has at least k-1 indistinguishable records and that there is variation in important characteristics within those groups.	By requiring that the distribution of the vulnerable typical inside each group of indistinguishable records is nearly identical to the prevalence of the attribute across the whole dataset, T-closeness is an additional variant of K-anonymity that attempts to provide an improved privacy guarantee.
To put it simply, K-anonymity requires that every collection of data (like age, gender, or ZIP code) that may be utilized to identify an individual be shared by at least k entries in the dataset.	In an L-diverse dataset, there has to be at least l distinct values for the sensitive characteristic for each group of k-1 indistinguishable records.	T-closeness, in other words, ensures that the distribution for each group's sensitive attribute is statistically close to the distribution of that attribute all through the entire dataset, making it difficult for attackers to derive sensitive data from outliers or skewed distributions.
The main goal of K-anonymity is to give a level of privacy by making it difficult to locate specific individuals, however if the quasi-identifiers (attributes used for identifying) have few distinct characteristics, it might not offer sufficient safeguards against attribute disclosure.	By reducing the risk of attribute disclosure and making it harder for an attacker to infer sensitive information if an attribute that is sensitive has more variability within the groups, L-diversity conquers the drawback of K-anonymity.	In contrast to K-anonymity and L-diversity, T-closeness provides a more severe privacy guarantee as it focuses on the variety of personally identifiable attribute values along with their statistical distribution.

8. What types of attacks are possible on summary tables?

For context, summary tables contain data abstracted from a number of individuals for generating insights. The aggregation works better if the information is about a large number of people with different characteristics. This limits the potential of the little PII that remains in aggregated data being used to uniquely identify an individual.

Therefore, summary tables are subject to external attacks, which could be an outsider guessing the information by comparing two tables released in succession. Another attack on the same aggregated data could be internal, such as a data leak. The last type of attack is dominance, and this is more statistically inclined. It happens on magnitude tables.

9. List and define approaches to protecting privacy in frequency tables.

The privacy of frequency tables can be protected in the following ways:

- Restricting queries in large tables: done effectively on web facing DBs, restricting the search to exact keywords and/or multiple filters can conceal remaining PII. This, however, poses a computational burden.
- Perturbed responses: giving a mix of relevant and irrelevant feedback during searches, rounding number values and potentially restricting queries. This "throws-off" the search, and can help conceal remaining PII in microdata.

10. List and define approaches to protecting privacy in online queryable databases.

Many methods exist for securing privacy in online queryable databases. These strategies fall into two broad categories are data anonymization and differential privacy.

Data anonymization: Data anonymization entails changing the data in a way that makes it impossible to identify specific people while still enabling valuable queries to be run. Several popular methods for data anonymization include:

- Generalization: This entails swapping out particular values for more inclusive ones. For instance, you may keep a person's year of birth or age range rather than their precise date of birth.
- Suppression: This entails completely deleting sensitive material. You may eliminate a person's name or address from a database, for instance.
- Perturbation: This entails introducing noise to the data to make it more challenging to distinguish between different people.

Differential privacy: This is a more advanced method of protecting data privacy that offers solid assurances against re-identification, even in the presence of an attacker with extensive access to the past. In order to maintain the accuracy of the answers for aggregate searches while significantly increasing the difficulty of identifying specific people, differential privacy works by introducing a little amount of noise to the query results.

11. What is differential privacy?

A mathematical concept called differential privacy protects users' privacy in datasets. By enabling data analysis without disclosing sensitive information about any of the individuals in the dataset, it can offer a solid assurance of privacy.

Differential privacy makes it more difficult to identify any specific individual's data while maintaining the general accuracy of the findings by introducing noise to the data. This is accomplished by generating the result using a randomized process, which makes it such that the chance of any particular output is about the same for any two datasets that differ in only one entry.

In the era of big data, differential privacy is a potent weapon for preserving individual privacy. Several businesses, including Google, Apple, and Microsoft analyze user data while protecting users' privacy.

References

[1] *What are the Differences Between Anonymisation and Pseudonymisation - Blogpost*. (n.d.). https://www.privacycompany.eu/blogpost-en/what-are-the-differences-between-anonymisation-and-pseudonymisation

[2] Wang, J., Du, K., Luo, X., & Li, X. (2018, June 29). Two privacy-preserving approaches for data publishing with identity reservation. *Knowledge and Information Systems*, *60*(2), 1039–1080. https://doi.org/10.1007/s10115-018-1237-3