Data Gov Final Project Outline

**Assignment Instructions**

Consider the following scenario: Assume you work for a telecom company as a data manager and your bosses decide to release/sell some account characteristic data on good customers to another company with whom your firm has a partnership. Before you can release your dataset, however, it must first be scrubbed, or redacted, to remove all personally identifying information (PII) on the file. Failure to do so will potentially allow an accidental or deliberate breach of the data, resulting in an FCC violation or potential lawsuit by injured parties. In addition to possible data disclosures due to a data breach, this could prove to be very embarrassing to the company and could possibly harm its reputation. Your tasks before release include the following:

As a data manager, you will explore these questions by first data masking (de-identifying) the Customer Survey dataset through two principal identifiers: Gender and Age. Then you are tasked with re-identifying the data and calculating the risk of exposure of each of the 4 basic scenarios listed early on in the project description above.

In order to fully answer these questions, you will need to go beyond this dataset and scenarios alone and examine this problem from a larger perspective. You will need to research historical risk data for this kind of data breach/attack activity in this industry, and during the course of your research, you will need to justify your calculations and decision to release the data based on the following questions.

1. Question 1: Before I mask my company’s data, what legal, ethical, and security requirements do I have to fulfill in order to ensure a high degree of data confidentiality before these data can be prepared for sharing and released? Explain in detail and cite your work.
2. Question 2: What is the risk or likelihood of exposure of these data to the general population, or to a subset of this population, such as other employees others close to you?

Customer Identification Risk Certification Report

Executive Summary

* Risk of customer identification was evaluated for …
* Reviewed the risk of patient re-identification under four different scenarios using generally accepted probabilistic algorithms and scientific principles
* Determined that there is a … risk of customers being identified …
* Preliminary recommendations and a list of potential data issues in Appendix …

Data Description

* Telecommunications Survey Data …
  + … customers
  + … locations
  + Etc
  + Etc
* Timeframe
* Variables (columns):
* Each x has y
* Each age range has x
* Each gender has x

Customer Risk Identification Methodology

Step 1: Quality control of the data was performed to look for outliers and other extreme records (see Appendix 1).

Step 2: We selected the following customer indirect variables or quasi-identifiers: … We did not consider … in this report because …

Step 3: Equivalence classes (EC) of customers are created as a group of records with the same combination of values for any given number of quasi-identifiers

Step 4: We consider an acceptable threshold risk of x, as commonly used in risk evaluations of sensitive personal information. However, since there is no sensitive health information contained in this dataset, published literature indicates threshold risks in the range of x to x. We will consider all of the above threshold values in the calculation of overall risk

Step 5: The patient identification risk was measured under the following four possible attack scenarios:

Scenario 1: The intruder deliberately attempts to identify patient data

Scenario 2: The intruder inadvertently identifies patient data

Scenario 3: There is a data breach at the data recipient’s site and the data is exposed

Scenario 4: An adversary launches a demonstration attack on the data. This occurs when an adversary wants to make a point of showing that a dataset can be re-identified

Scenario 1: Deliberate Data Attack

* The main assumption in this scenario is the possibility that someone at the data recipient’s site will attempt to identify the data.
* For example, there may be a rogue staff member who wants to monetize the data for financial gain
* The calculation of risk is based on the product of the following two probabilistic measures:
  + Probability of attempt: There are about x employees of the data recipient with access to the data. Staying conservative, we will say that y of these employees may go rogue. This makes the probability of attempt y/x = …%
  + The probability of correctly identifying a patient given an attack was determined to be x, as computed directly from the dataset

Scenario 2: Inadvertent Data Attack

* The main assumption in this scenario is that someone at the data recipient’s site recognizes a patient in the dataset.
* For example, a sales representative working with the dataset may recognize an acquaintance, such as a relative or a neighbor, in the dataset through x and y
* The calculation of risk is based on the product of the following two probabilistic measures:
  + Probability of acquaintance: On average, people tend to have 150 friends (). Include information on the people in the dataset and the chances of knowing them??
  + The probability of correctly identifying a patient given that the adversary knows someone from the population covered by the dataset was determined to be x, as computed directly from the dataset

Scenario 3: Data Breach

* This scenario can take place if the data recipient loses the dataset, or a data breach occurs
* Based on the latest IBM Data Breach Report, x% of companies suffer from a reportable data breach per year ().
* The calculation of risk is based on the product of the following two probabilistic measures:
  + Probability of a breach: x%, as reported by IBM
  + The probability of correctly identifying a patient given a data breach situation was determined to be x, as computed directly from the dataset

Scenario 4: Demonstration Attack

* This scenario can take place when the data is disclosed publicly.
* We assume that there is an intruder who has background information that can be used to launch an attack on the data, and that the intruder will attempt a patient identification attack
* We are unaware of all of the people that will be utilizing the dataset, so we must consider this as a plausible scenario
* The calculation of this risk was computed directly from the dataset and determined to be x

Results

Insert table: Percentage of customers at each risk range for each scenario

Use range of <5%, <10%, <20%, <33%, <50%, >50% for S1-S4

Risk is considered tolerable for scenarios x and y (x and y) but unacceptable for scenarios z and a (z and a)

Insert table showing the max risk, median risk, and assessment for each scenario

Appendix 1: Data Quality Control

List all unknowns, multiples, nulls, outliers, etc.

Appendix 2: Risk Assessment Methodology

Use from El Emam book??

Appendix 3: Recommendations

* We recommend that the ongoing delivery of data is based on secure procedures.
* Full documentation of the data recipient’s data storage and accessibility protocols must be assessed to further determine risks.
* Recommendations for lowering the risk of the existing dataset:
  + Identify and correct errors
  + Identify and mask obvious outliers
  + Group x y z
  + Eliminate x y z

References:

El Emam book