**ADSEE Project**

**Applied Data Science Educational Ecosystem**

**Privacy Case**

**Labour Market / Data Anonymization**

**Approach Analysis Whitepaper**

# 1. Introduction

The labour market is the place where market demand meets the supply of labour. Furthermore, it is also the place where workers and (future) employees communicate with each other. This interaction yields a variety of data that is used in different fields of research (e.g. labour market intelligence (LMI), and People Analytics). The abundance of secondary data pertaining to individual students, job applicants, and/or employees is often used in analyses on disparate topics including learning analytics, recruitment and selection, personnel development and so forth. Although the interest in such research is often not on the person, but instead on examining relationships between variables, an unavoidable side effect of the individual as the focal unit of analysis is the potential (re)identifiability of those people whose data is under investigation. And although there is much to be said in favor of replicability, and the associated open science and open data movements, clearly they exacerbate the potentially severe (ethical) ramifications of privacy sensitive individual data.

Besides the research ethical obligations to protect individual privacy, governments and policy makers are increasingly seeking to protect individuals with rules and regulations, such as the General Data Protection Regulation (GDPR) in the E.U., which greatly relaxes many of the most restrictive regulations for data that are anonymized or pseudonimized. Consequently, researchers and practitioners alike need to rethink how to store and process data on local servers or cloud instances and how these data sources are accessed. Take for example, internal documents that contain information pertaining to sickness or the mental health of individuals within an organization. Even seemingly harmless information can be potentially harmful when linked with other information from external data sources. Seen from this perspective, it is worrying that *sweeney2000* demonstrated that knowing simple demographics such as someone’s birth date, zip code and gender is enough to uniquely identify $53\%$ of the U.S. population with the potential risk of linking their now known identity to what on the face of if appears to be anonymous data sets.

## 1.1 Basic Definitions

Before we start introducing data anonymization in greater detail, there are certain preliminary terms that need to be defined. The attributes, or characteristics, of certain individuals in a dataset can be classified as the following:

* **Quasi-Sensitive Identifiers:** Is the set of attributes in the dataset that can be joined with external information to re-identify individual records. Note that one needs domain-specific information of the dataset to judge whether an attribute is quasi-sensitive. An example is if we have information on the health status of a male between the age of 60-70 and we know where the person in question is working. Combining this information with HR data from the company, then, might lead to re-identification, since it might be possible there is only one male between the age of 60-70 working for that company.
* **Sensitive Identifiers:** Is the set of attributes in the dataset that can be used directly to identify individual records. For example, knowing a person’s name and address leads to direct identification of the individual.
* **Equivalence Class:** An equivalence class for the data set with respect to the column attributes is the set of all rows in the dataset containing identical values on the column attributes. For example, let us assume that we have a HR database with performance measurements for individuals over their years of employment where each row contains the measurements of a specific employee and each column contains a measurement indicator in a certain point in time. With this information, we can then refer to the equivalence class of the measurement indicator "A" as all the rows that contain the same value on that specific indicator.

## 1.2 How is data anonymization important to labour market related problems?

Employers collect personal data on job applicants and employees for a number of purposes; to carry out basic administrative tasks; to comply with laws and regulations; to assist in the selection for employment, training or promotion; to ensure personal safety, etc. In recent years, more and more HR departments have started using this collected data on workers in order to enhance their understanding of how to manage the employees. However, this also means that data containing information on individual employees working at a certain company is being used to facilitate analyses, with the potential risk of exposure to internal or external people who do not normally have the rights to access the information. For instance, HR data could constitute a gold mine for academics in the field of work and organizational psychology. Due to the fact that they are employed by universities granting them access to identifiable data, however, may constitute a violation of employees’ rights under the GDPR.

In starting to think about ways of protecting, and in doing so, reducing the risk of leakage, anonymization and pseudonymization are important. We define anonymization to be the process of removing sensitive identifiers from data sets to preserve anonymity for the individual. On the other hand, psuedonymization refers to the procedure by which we replace information sensitive to identification with artificial identifiers such as pseudonyms. For instance, sensitive identifiers could be replaced with random numbers, with only the data owner knowing which number maps on to which individual. It should be noted that, however, that despite anonymization or pseudonimization, records about specific individuals are problematic in that they tend to have a high dimensionality, which may comprise a unique profile across attributes (e.g. education, highest degree, etc.) that can be leveraged to re-identify particular individuals in the dataset. Combining high dimensionality with the fact that the average record does not have many similar records in the dataset yields a so-called ’fat tail’ phenomenon where individual characteristics tend to include statistically rare attributes, greatly simplifying the (re)identification of a particular individual *narayanan2008*. For instance, if one had knowledge of a particular employee calling in sick on a particular day, that single datum is likely to be enough to pin an entire vector (i.e. record) to that individual in a presumed to be fully anonymous dataset containing privacy sensitive data on (reasons for) absence from work.

To combat de-anonymization of the earlier mentioned sparse and highly dimensional individual records, we will explore within this whitepaper a variety of methods and their corresponding advantages and disadvantages. In particular we will look at methods of generalization, suppression, anatomization, perturbation, and permutation and their application to a labour market related data set containing census data. Furthermore, we will demonstrate privacy protection techniques on this data set. Finally, we will carry out several attacks to try to de-anonymize the sensitive attributes unveiling common pitfalls and rules of thumb in the process.

## 1.3 Learning Objectives

After completing this course, learners will know how and when to apply what basic data protection techniques that can help safeguard the sensitive attributes of individual records. Furthermore, the course shows common pitfalls and methods in which adversaries might attack the data to gain insight into the sensitive attributes of a certain individual. In addition, we provide learners with a case study to exhibit a possible workflow in which we demonstrate pre-processing of the datasets and how to recognize privacy issues.

# 2. A Case of Identity

Let us assume that a certain company hired an external consultant to analyze employee turnover. To do so, they ask us to prepare a dataset with all attributes necessary for conducting the analysis. Since we are dealing with an external consultant who has no access rights for any of the relevant data, but there is pressure from management to deliver on this project, we have to devise a way to retain as much information and latent statistical relationships between the variables as possible, while at the same time not safeguard the privacy of the individual employees working for the company.

Before we start our analysis, it is first necessary to provide some additional background information regarding the company for which we work and the data that is collected. It is a relatively small company that has three separate departments, namely "Sales", "Marketing" and "R&D". The sales department has 5 employees, of which only one is female. The research and development department has 4 employees, and again only one of these employees is female. For the last department, in the marketing department we have 10 employees, of which 5 are male, and 5 are female. For each employee, we have gathered the following information; the department for which the employee works, the age of the employee, gender, the role of the employee within the company and the employee’s income, performance, education and job satisfaction.

In the following sections we will provide tables to describe the influence of certain data privacy protection methods on information loss and privacy protection. In order to facilitate our discussion, we chose to present the fictional names of our "employees" in the tables. This sensitive attribute will not be taken into consideration when we run the different methods and can be deemed suppressed for the external consultant by you, the curator of the data.

## 2.1 Considerations

We now have all the information we need to start doing a preliminary analysis on what vulnerabilities we have to look out for. However, before we can start looking at the vulnerabilities, we have to use our "domain knowledge" to determine what attributes qualify as sensitive attributes, and what attributes are quasi-sensitive. As mentioned before, we have gathered quite some information on various employees. It is our task to distinguish between those attributes that can lead to direct re-identification and those attributes that cannot. Note that we have already determined that each of the departments is relatively small. Furthermore, we know that there are two departments with only a single female employee. With this in mind, we decided that knowing which department an individual belongs to can cause a high risk of re-identification. We therefore decided to classify "Department" as a sensitive attribute. All the other attributes do not directly lead to re-identification of the individual making them quasi-sensitive identifiers.

**References**

[1] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of largesparse datasets. In2008 IEEE Symposium on Security and Privacy (sp2008), pages 111–125. IEEE, 2008.

[2] Latanya Sweeney. k-anonymity: A model for protecting privacy.Inter-national Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,10(05):557–570, 2002.