**ADSEE Project**

**Applied Data Science Educational Ecosystem**

**Privacy Case**

**Labour Market / Data Anonymization**

**Operalization of Results Whitepaper**

**2.2** **Application of data anonymization techniques**

**2.2.1** $k$**-Anonymity**

K-anonymity is a popular approach to address privacy concerns, and does so by leveraging generalization and suppression to diversify quasi-identifiers. Generalization involves the replacing of values with a less specific but semantically consistent value, and suppression is not releasing any value at all *sweeney2002*. One example of generalization would be recoding numerical values for a person’s age to categorical classes, e.g. 62 then becomes between 60-69. The new categorical value is less specific, but still semantically consistent in that it still approximately captures the same information about the person’s age. Typical suppression schemes include record suppression and value suppression, where the original value and/or record will be replaced with a special value (e.g. ’\*’) *xu2014*. *sweeney2002* proposed that data adheres to k-anonymity if for each record there are at least $\left(k-1\right)$ other records that have similar quasi-identifiers across a given range of attributes. For example, if we set $k=10$, and look at the sector a person works in, then a query for a particular sector will have to yield a minimum of $9$ data entries to avoid re-identification. In-short, *sweeney2002* states that k-anonymity provides privacy protection by guaranteeing that each record is indistinguishable from at least $\left(k-1\right)$ other records.

In our example [[Table2]](#Table2), we applied 4-anonymity on the employee’ dataset to provide at least $4-1=3$ distinguishable similar cases for each quasi-sensitive identifier. Since it is a relatively small company, and we collected 8 attributes on which we could potentially identify the individual, it is important to take into consideration the sparsity of some profiles. As mentioned before, there is a sparsity of female employees in the Marketing and Sales departments; as such, if we know an individual who is working as an employee for the marketing or sales department and that she is a female, then we can clearly identify that person’s record in the dataset. [[Table2]](#Table2) shows us that all age groups but the age group 30-39 is sensitive to re-identification. Let us look at the different equivalence classes that exist in our current data set. Upon examination of the dataset, we find the following partitioning $\{\{3,4,6,8\},\{1,2,5,7,9,10\}\}$, where the numbers resemble the row numbers. Note that there is quite some information loss, in that we have had to supress the ages of 6 individuals and are essentially left with but 2 age categories in this setting. The reason for this is the dimensionality of the data set. That is, providing privacy by suppression and/ or generalization using 4-anonymity on 8 attributes with only 10 rows really stretches the possibilities on providing useful data for further analysis.

Furthermore, the approach outlined in the above still leaves out of consideration the sensitive identifiers which might not be diverse enough to guarantee privacy *narayanan2008*. For example, if everyone would suffer from the same illness, then no matter how indistinguishable the records are at the attribute level, one can still glean the illness the person is suffering from the data that is provided. Furthermore, since k-anonymity takes into consideration the partial information of different dimension simultaneously, it is open to inference attacks *aggarwal2005*, potentially causing severe information loss even at $k=2$. This means that by increasing the number of attributes that you want to adhere to the $k$-anonymity, the larger the information loss becomes from a data mining point of view *aggarwal2005*. *aggarwal2005* states that this is caused by the exponential number of combinations of dimensions that can be used to make precise inference attacks.

*Present two vulnerabilities that can compromise the privacy protection provided by k-anonymity.*

The first method is the so-called homogeneity attack, which uses the homogeneity of the data to draw conclusions on values. For example, [[Table\_1]](#tab:table_1) shows that knowing a certain individual works as a "Laboratory Technician" is enough to figure out that the person must be working at the "R&D department", even though the quasi-sensitive identifiers are $\left(k-1\right)$ indistinguishable to other records. The second method is called a background knowledge attack. For example, knowing that all the highest paid employees are working in "R&D" might help inferring what other attributes a certain individual has.

To summarize, $k$-anonymity has the following advantages:

* It prevents potential adversaries from connecting a sensitive attribute with external data by providing $\left(k-1\right)$ equivalent cases for each dataset entry.
* It is intuitively easy to understand.

and the following disadvantages:

* It has relatively high loss of utility, meaning that the statistical relations between the variables become prone to information loss when there are many different attributes.
* There are multiple methods of attack to gain access to the sensitive attributes of certain individuals. In this white paper we further elaborate two, namely; the homogeneity attack and the background knowledge attack.



**2.2.2** $l$**-Diversity**

As mentioned earlier, k-anonymity is susceptible to homogeneity and background knowledge attacks *mach2007*. *mach2007* provides an adaptation to the k-anonymity algorithm, the so-called l-diversity, that ensures that the values of the sensitive attributes are "well represented" in each group. Let us start by defining an equivalence class as the set of records that are indistinguishable from each other with respect to certain "identifying" attributes *mach2007*. Intuitively, l-diversity looks at the equivalence classes. Specifically, a class is said to have l-diversity if there are at least $l$ "well-presented" values for the sensitive attribute. Consequently, the table is l-diverse if every equivalence class of the table has l-diversity *li2007*.

There are three ways in which the term "well represented" can be operationalized. However, in our text, we will stick to the simplest understanding of ensuring that there are at least $l$ distinct values for the sensitive attribute in each equivalence class, where $l$ is an integer bigger than zero. This means that, on top of the $k$-anonimity protection, we are also ensuring that there is at least $l$ distinct options for the sensitive attributes given to each equivalence class. This would still make individual records very susceptible to probabilistic inference attacks however, since an adversary may leverage the probability distribution of the values to re-identify particular individuals *li2007*. To combat this susceptibility, two stronger notions on l-diversity are; i) Entropy l-diversity and (ii) Recursive $\left(c,l\right)$-diversity. These approaches are beyond the scope of this white paper; thus, we refer the interested reader to the research by *li2007* for a more thorough exposition.

After applying the simplest "well represented" operationalization of 4-diversity to our employee data, we can see in [[Table3]](#Table3) that the sensitive identifiers are all changed to the marketing department. Note that $l$-diversity only changes the values on the sensitive attributes building on the generalizations and suppression devised by the 4-anonymity from [[Table2]](#Table2). Let us find out why it is changing all the department values to "Marketing". One reason for the failed convergence could be due to there not being four distinct values for the sensitive attribute department to begin with. Now let us try setting $l$-diversity to $l=2$. Again, the $l$-diversity does not converge, meaning that it is not capable of finding a suitable replacement for the sensitive attributes to satisfy the $l$ distinct values condition.



While the l-diversity adaption is valuable in that it provides additional protection against homogeneity and background knowledge attacks, it still has several vulnerabilities and/or shortcomings. *li2007* showed that $l$-diversity may be difficult and unnecessary to achieve and could be insufficient to prevent attribute disclosure. This means that the computational cost can sometimes far outweigh the benefits since $l$-diversity is not always adding towards the protection that $k$-anonymity provides. *li2007* demonstrates the insufficient protection due to attribute disclosure since it displays skewness and similarity attacks, which are two key examples where these vulnerabilities may reveal sensitive information pertaining to a particular individual. Let us summarize these given examples by reframing them to our case study.

Consider the situation where the equivalence classes are l-diverse, but the overall distribution is skewed: for example, the test results for a certain performance test for internal employees. The test results take on two values: pass or fail. Further suppose that 10000 people took the test and that only 1% (i.e. 100 people) passed it. Now, if we would use 2-diversity for this particular case, then we would get at most 100 equivalence classes and thus a large loss of information rendering its usage for any data mining task less useful, since we need to provide at least 2-diverse sensitive attributes for each equivalence class.

This makes a probabilistic inference attack very likely to succeed, since an adversary can easily conclude that an entity in the equivalence class is very likely to have failed the test even though the interest is much more likely to be in the person who did pass the test. This indicates that the privacy of that individual is still protected by mixing his/ her identity with 99 other individuals. Nevertheless, the combination of the loss of information and the leakage of the failure of almost all other participants limits the degree of privacy protection afforded.

There is also the case where the sensitive attributes in an equivalence class are semantically similar to *li2007*. For example, let us assume that we have an employee database that tracks the off-time of employees due to illness such as stress or burn-out. Regardless of whether the employee is gone due to a burn-out or stress, these two sensitive attributes are somewhat similar in the sense that they both refer to a similar construct even though the former is more severe than the latter. We can leverage this semantic similarity of the classes within the sensitive attribute with the so-called similarity attacks. These attacks potentially give an adversary important information regarding the individual, in our case, it reveals to the adversary that the person is overworked regardless of what class the individual is in. In short, it is important to note that distributions that have the same level of diversity need not have the same level of privacy *li2007*.

In sum, $l$-diversity has the following advantages:

* Protects against homogeneity and background knowledge attacks.
* Provides protection of both the quasi-sensitive identifiers and the sensitive identifiers.

and the following disadvantages:

* $l$-diversity is not always achievable and can be quite hard to compute.
* There are multiple methods of attack to gain access to the sensitive attributes of certain individuals. In this white paper we mentioned two: namely, the skewness and similarity attacks.

**2.2.3** $t$**-Closeness**

In order to consider these semantic relationships between attribute values, *li2007* proposes $t$-Closeness. An equivalence class is said to have $t$-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold $t$. A table is said to have $t$-closeness if all equivalence classes have $t$-closeness *li2007*.

The implications of $t$-closeness are probably best understood by providing an example. Let us assume there is a certain observer with a prior belief: let us denote it with $B\_{0}$, which indicates the sensitive attribute of a certain individual that is stored in a data set $D$. From the data set $D$, the observer gets a generalized version, meaning that it is generalized in such a way that all attributes in a quasi-identifier are either suppressed or generalized. After observing the distributions of the sensitive attributes in the whole table, denoted $Q$, the observer’s belief is changed to $B\_{1}$. Since the observer knows the quasi-identifier values of the individual, the observer is then able to identify the equivalence class that the individual’s record is in. In doing so, the observer learns about the distribution $P$ of the sensitive attributes in the class. The observer’s belief now changes to $B\_{2}$.

Intuitively, the $t$-closeness adaption is focused on limiting only the capabilities of the observer to learn new information about specific individuals *li2007*. As such, the focus is on minimizing the difference between $B\_{1}$ and $B\_{2}$, meaning that the focus of $t$-closeness is on $Q$ *li2007*. In contrast, the $l$-diversity approach applies certain constraints on $P$, so that it can only have a certain number of distinct classes and thus is focused on minimizing the difference between $B\_{0}$ and $B\_{2}$.

*li2007* justify their choice for focusing on $Q$ rather than on $P$ by stating that a large change from $B\_{0}$ to $B\_{1}$ means that the data contains a lot of new information. Since $t$-closeness focuses on the transition $Q$ that brings $B\_{1}$ to $B\_{2}$, it will never limit the new information gained from $B\_{0}$ to $B\_{1}$ *li2007*.

To conclude, *li2007* minimize the difference between $B\_{1}$ to $B\_{2}$ by limiting the distance between $P$ and $Q$. If the case arises where $P=Q$, then $B\_{1}=B\_{2}$ *li2007*. We can make similar statements about the distance between $B\_{1}$ and $B\_{2}$. An advantage of using $t$-closeness is that, in contrast to $k$-anonymity and $l$-diversity, it can hide sensitive attributes of specific records *li2007*. Furthermore, the number of records in the anonymized table is accurate, making it very suitable for applications *li2007*. Since this technique does not impact quasi-identifiers, it does not affect $k$-anonymity. Removing a value does decrease the diversity, meaning that it can negatively impact reaching $l$-diversity *li2007*.

In summary, $t$-closeness has the following advantages:

* $t$-closeness will never limit the new information gained from $B\_{0}$ to $B\_{1}$.
* It can hide sensitive attributes of specific records, making it very suitable for applications.
* $t$-closeness is capable of identifying the closeness of certain attributes, solving limitations that $l$-diversity has.

and the following disadvantages:

* It is hard to identify the distance between distributions.
* Necessitates that the sensitive attributes spread in the equivalence class need to be close to that in the overall dataset.

**3.** **Impact of our Solution**

The solution to the proposed problem can help drastically reduce the risk of the re-identification of the sensitive attributes of individuals. Using suppression and generalization, however, also destroys information that might have been relevant for the analysis that the external consultant in our case study would have liked to perform. Therefore, the role of the data provider cannot be understated, since it is their task to make the distinction between sensitive and quasi-sensitive attributes. As such, the analysis is basically as strong as the judgements of the data provider.

**4.** **Conclusion**

In this paper we have shown examples of how privacy preserving techniques can be used to protect sensitive attributes of the individual. Usage of the earlier defined methods can greatly help researchers, practitioners and organizations in releasing information to the public by keeping in mind the trade-off between statistical relevance of the data and the privacy of the individual. In general, the following recommendations are given:

* Ask different domain experts which attributes qualify as quasi-sensitive. Try making a list of attributes that are needed to link the attribute to the individual level.
* Carefully check whether you can use the mentioned attacks to derive to the sensitive attributes of particular individuals. Knowing how certain attacks work can help you devise methods of defense and/ or show where the vulnerabilities in your provided protection can still improve.
* etc.

Keep in mind that these are only very basic ways of protecting the privacy when releasing datasets to e.g. third parties. There exist more sophisticated approaches that are able to retain more statistical relevance than the methods demonstrated in this paper. We recommend looking at the work by *xiao2006* for an alternative method that can replace $t$-closeness. Furthermore, the work by *dwork2014* shows an overview of different methods and frameworks also showcasing perturbation and permutation methods.

**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.