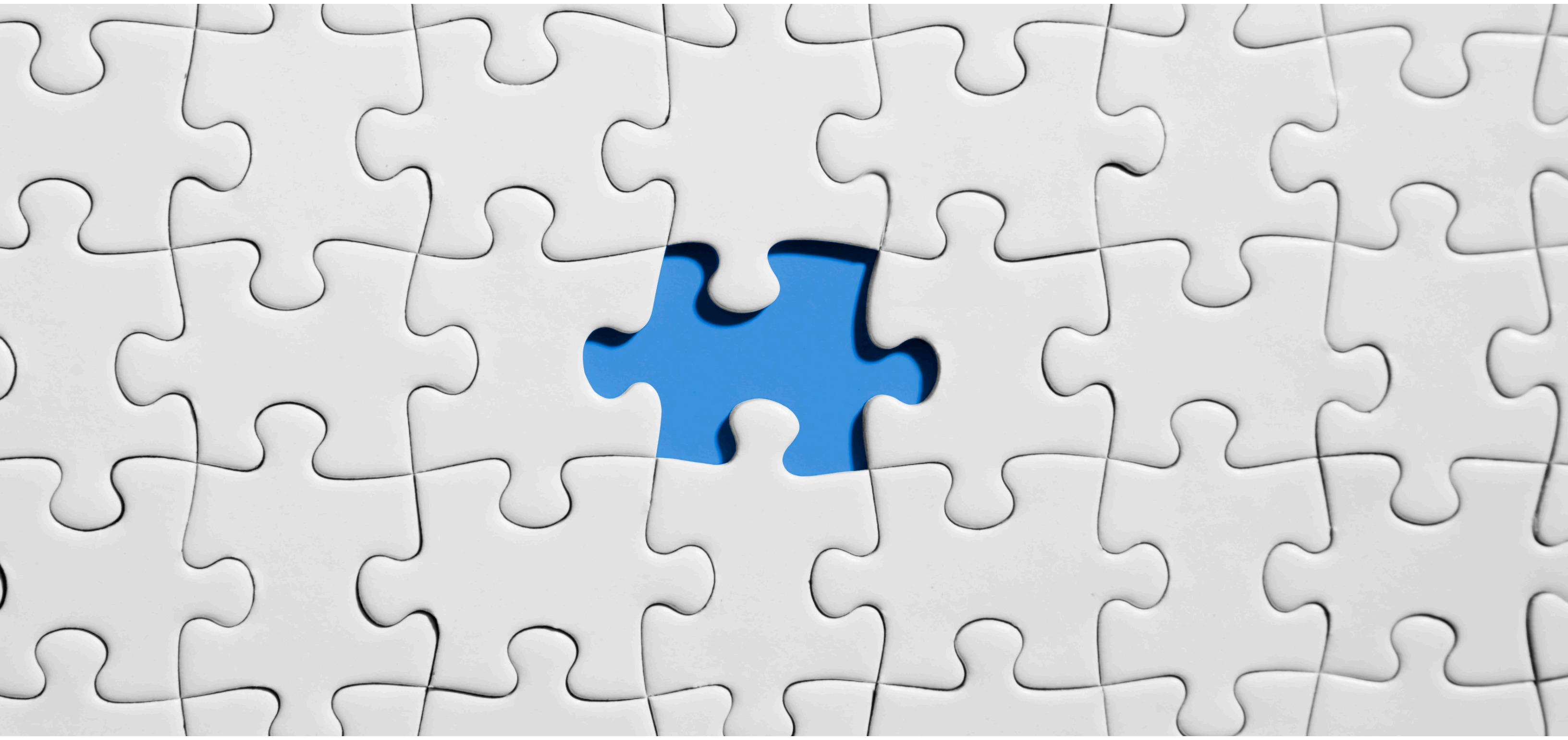
STUDY OF THE USE OF ANONYMITY MODELS



Carmen Marcano

Director, Business Process Management

Dell

Carmen.marcano@dell.com

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Abstract 4

Initial Concepts and Terminology 5

Anonymity Models 6

Anonymity: k-Anonymity 6

k-Anonymity Implementations 7

Problems with k-anonymity - Homogeneity Attack (Machanavajjhala, 3-4) 8

Problems with k-anonymity - Background Knowledge Attack (Machanavajjhala 3-4) 9

l-diversity Implementations 11

Problems with l-diversity - Skewness attack 11

Problems with l-diversity - Similarity Attack 13

Anonymity: t-closeness 13

t-closeness Implementations 15

Problems with t-closeness: Difficult to calculate if more than one sensitive attribute are present . 15

Problems with t-closeness: Limitations of the EMD 15

Summary of Observations 16

Conclusions and Future Studies 17

References 18

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Abstract

Uploading large amounts of data in big data repositories often ends up being shared and aggregated. Sometimes this data is either openly available or with restricted access sites for research of a private team's purposes. In either case, the data holder must ensure that the privacy of individuals whose personal information is included in the released data is not compromised.

This article analyzes that removal of directly identifying information is not enough to guarantee privacy of the subjects whose data is shared. We also analyze that in this removal process the data may lose its utility for research purposes. With this perspective we explore the most widely known anonymity models in the quest for answering the question: How can a data holder release the data without compromising the identity of the subjects while maintaining its utility? Specifically, we analyze k-anonymity, t-closeness and l-diversity as frameworks that can be implemented to reinforce the privacy levels of the Personal Identifiers.

Initial Concepts and Terminology

When collecting and storing large amounts of data into Big Data repositories and data lakes, the collection process may also include data that can be considered Personally Identifiable Information (PII). The PII is any type of data that can be used to identify a specific individual, such as social Security Number, Name, Last Name, IP addresses used for a connection, login IDs, email, social media publications or digital pictures.

Collecting large amount of data may be used for multiple purposes, for instance, for reporting and analytics, and this process may reveal PII tied to the data being analyzed. This creates a challenge for privacy compliance as the reporting data may be exposed if not handled properly. This privacy compliance is especially critical nowadays with the different regulations across the globe that aim to protect against disclosure of PII. In order to further analyze the anonymity algorithms as an alternative to conceal PII to its minimum factor that still maintains the usability of the data for the reporting and analytics, let's examine the concepts of identifiers and quasi-identifiers.

A quasi-identifier does not directly identify an individual but makes him/her more unique in a given population. Quasi-identifiers include age, gender, race, ethnicity, city of residence, zip code, educational level, dates (birth, death, admission, and discharge), etc. In contraposition the identifiers are directly related to the identity of an individual, such as name, address, SSN, phone number, etc.

Another important concept is the difference between de-identification and anonymization. De-identification is the removal or replacement of the Personal Identifiers so that it would be difficult to reestablish a link between the individual and his/her data; while anonymization refers to the irreversible removal of the link between the individual and his/her data to the degree that it would be virtually impossible to reestablish the link but the data remains useful for the research and utility of further research (Kushida 82).

Different heuristics methods are usually used for protecting the Personal Information using De-Identification. These de-identification heuristics mandate removal of the identifiers and quasi-identifiers before releasing any data. In the US, there are regulations that protect the privacy of individuals and limit the exposure of its sensitive and private data.

Some regulations, such as GDPR in Europe and the HIPAA Privacy Rule in US provides three standards for the disclosure of information without seeking authorization. For example, HIPAA (for medical providers) the Safe Harbor standard, the Limited Dataset, and the Statistical Standard (McGraw 29). However, this de-identification heuristics may provide insufficient protection for the complex datasets and may result in disclosure of Personal Identifiers or a re-identification (El Emam, *Heuristics for De-identifying Health Data* 59-61) (Mcgraw 29-31). An example can be found in Sweeney's description of the Re-identification by linking attack (*k-anonymity: a Model for Protecting Privacy* 2-3).

As the privacy concerns are particularly critical for Health Data exposure in US (HIPAA) or customers of European Union data (GDPR), a good practice is to start implementing some sort of anonymity to all PII data lakes in order to add an extra privacy protection layer to all PII from any potential disclosure, unintentional or not.

Anonymity Models

According to Sweeney (*Achieving k-anonymity privacy protection using generalization and suppression* 572), a vast majority of the US population can be uniquely identified based on zip code, gender and date of birth. The uniqueness of such combinations leads to a class of attacks where data records are being re-identified by joining multiple, often publicly available, datasets. To perform such *linking attacks*, the attacker needs two pieces of prior knowledge: the quasi-identifier of the victim and the victim's record in the published dataset.

To prevent privacy threats from linking attackers, the data publisher releases an anonymous version of the original dataset A. The resulting dataset A* is obtained by applying anonymization operations to the attributes in the Quasi-Identifier of the original records in A (Sweeney, *k-anonymity: a Model for Protecting Privacy* 4*).*

Anonymity: k-Anonymity

The *k*-anonymity proposed by Sweeney is a framework for constructing and evaluating algorithms and systems that release information such that released information limits what can be revealed about the properties of entities that are to be protected (*Achieving k-anonymity privacy protection using generalization and suppression* 572). With *k*-anonymity, an original dataset containing personal information can be transformed so that it is difficult for an intruder to determine the identity of the individuals in that dataset.

For example, a *k*-anonymized dataset has the property that each record is similar to at least another k-1 records on the potentially identifying variables. For example, if *k*=5 and the potentially identifying variables are age and gender, then a *k*-anonymized dataset has at least 5 records for each value combination of age and gender. The most common implementations of *k*-anonymity use transformation techniques such as generalization, global recoding, and suppression.

Formula for *k*-anonymity (Sweeney, *Achieving k-anonymity privacy protection using generalization and suppression* 573)0: Let RT(A1,...,An) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy *k*-anonymity only if each sequence of values in RT[QIRT] appears with at least k occurrences in RT[QIRT]. Table 1 exemplifies k-anonymity, where k=2 and QI={ Birth, Gender, ZIP}:

Table 1. *k*-anonymity example - Original Impatient Database

	Identifying Variable	Quasi-Identifiers				
ID	Name	Gender	Year of Birth	Zip	Company Name	Last purchase
1	Ben Parker	Male	1959	01243	ABC	laptop
2	Peter Parker	Male	1963	01355	XYC	Server
3	May Parker	Female	1955	01954	ABC	Monitor
4	Natasha Romanova	Female	1945	01986	WST	Monitor
5	Robert Banner	Male	1958	01255	ABC	laptop
6	Bruce Banner	Male	1970	01322	XYC	Server

Table 2. 2-Anonymous Impatient Database

	Quasi-Identifier			
ID	Gender	Decade of Birth	ZIP	Last purchase
1	Male	1950-1960	12XX	laptop
2	Male	1960-1970	13XX	Server
3	Female	1950-1960	19XX	Monitor
4	Female	1950-1960	19XX	Monitor
5	Male	1950-1960	12XX	laptop
6	Male	1960-1970	13XX	Server

As can be appreciated, for every combination of values of quasi identifiers in the 2-anonymous table, there are at least 2 records that share those values. As well, any record in a *k*-anonymized dataset has a maximum probability 1 / *k* of being re-identified. Thus, for the example, the probability is 0.5.

In practice, a data custodian would select a value of *k* commensurate with the re-identification probability they are willing to tolerate, called a threshold risk. Higher values of *k* imply a lower probability of re-identification, but also more distortion to the data, and hence greater information loss due to *k*-anonymization. In general, excessive anonymization can make the disclosed data less useful to recipients because some analysis becomes impossible or the analysis produces biased and incorrect results.

k-Anonymity Implementations

An example of a *k*-anonymity in the real world is the program Datafly developed in 1997 (Sweeney, *Guaranteeing anonymity when sharing data, the Datafly system)* and the Application Incognito developed in 2005 (Lefevre 7) which code in a high level is as follows:

Table 3. *k*-anonymity Implementation of Incognito

Input: A table *T* to be k-anonymized, a set *Q* of *n* quasi-identifier attributes, and a set of dimension tables (one for each quasi-identifier in *Q*)
Output: The set of k-anonymous full-domain generalizations of *T*
Variables: C1 = {Nodes in the domain generalization hierarchies of attributes in Q} E1 = {Edges in the domain generalization hierarchies of attributes in Q} queue = an empty queue

for i = 1 to n **do**
//Ci and Ei define a graph of generalizations
Si = copy of Ci
{roots} = {all nodes ϵ Ci with no edge ϵ Ei directed to them}
Insert {roots} into queue, keeping queue sorted by “last purchase”
while queue is not empty **do**

```
node = Remove first item from queue
if node is not marked then
    if node is a root then
             frequencySet = Compute frequency set of T with respect to attributes of
    node using T.
        else
        frequencySet = Compute frequency set of T with respect to attributes of
    node using parent's frequency set.
       end if
    Use frequencySet to check k-anonymitywith respect to attributes of node
    if T is k-anonymous with respect to attributes of node then
      Mark all direct generalizations of node
    else
     Delete node from Si
     Insert direct generalizations of node into queue, keeping queue ordered by "last
purchase"
    end if
end if  end while
Ci+1;Ei+1 = GraphGeneration(Si, Ei)
end for return Projection of attributes of Sn onto T and dimension tables
```

Problems with k-anonymity - Homogeneity Attack (Machanavajjhala, 3-4)

Situation: Alice and Bob are neighbors. One day Bob received a large package from XYX.com with a large sign XYX on the front of the package. Alice saw UPS delivery leave the package on the curbside and she wanted to find out what Bob just bought. She visits the XYX.com web page and discovers the 4-anonymous table published by the company. One of the records in this table contains Bob's. She knows Bob must be over 30 years and knows his zip code. Alice deduced that record 12 was probably related to Bob because the other groups were not diverse enough and hence, she discovers that Bob just bought a new Power XXX system. See tables:

Table 4. Original Customer Data

	Identifying Variable	Quasi-Identifiers			
ID	Name	Nationality	Age	Zip	Last purchase
1	Boris	Slovenian	27	01754	Robot
2	John	American	25	01752	Robot
3	Akira	Japanese	22	01752	Game Server
4	Peter	American	26	01752	Game Server
5	Guru	Indian	52	01592	Power XXX
6	Ivan	Russian	66	01592	Robot

7	Matt	American	42	01590	Game Server
8	Mark	American	44	01590	Game Server
9	Mike	American	36	01758	Power XXX
10	Mahatma	Indian	38	01750	Power XXX
11	Yamato	Japanese	37	01754	Power XXX
12	Bob	American	34	01754	Power XXX

Table 5. 4-Anonymous Published Customer Data

	Identifying Variable	Quasi-Identifiers			
ID	Name	Nationality	Age	Zip	Last purchase
1	*	*	<30	017*	Game Server
2	*	*	<30	017*	Game Server
3	*	*	<30	017*	Robot
4	*	*	<30	017*	Robot
5	*	*	>40	017*	Power XXX
6	*	*	>40	015*	Game Server
7	*	*	>40	015*	Robot
8	*	*	>40	015*	Robot
9	*	*	3*	017*	Power XXX
10	*	*	3*	017*	Power XXX
11	*	*	3*	017*	Power XXX
12	*	*	3*	017*	Power XXX

Based on the observation*,* *k*-anonymity can create groups that leak information due to lack of diversity in the sensitive attributes.

Problems with k-anonymity - Background Knowledge Attack (Machanavajjhala 3-4)

Situation: Alice's friend Akira, who also lives in the neighborhood and also bought and received a large package from XYX.com. Alice knows that Akira is a 22-year old Japanese female who currently lives in zip code 01752. Based on this information, Alice learns that Akira's information is contained in record number 1, 2, 3, or 4 of Table 5. Without additional information, Alice is not sure whether Akira bought a Mega server or a Robot. However, Alice may assume, as Japan hosts the largest Robots exhibitions every year, that Akira may be buying a type of robot (the argument is that information of general public knowledge may affect privacy in the k-anonymity). Also, she knows that Akira works a lot and probably does not have time for computer games, and she does not have children. Therefore, Alice concludes with near certainty that her friend bought a new Robot.

In this example, the adversary can correctly identify the value of a sensitive attribute with high probability by elimination or negative disclosure. Thus, Akira cannot have bought probably a new Game Server then she must has bought the Robot. Based on this observation, *k*-anonymity does not protect against attacks based on background knowledge.

1.1 **Anonymity: *l*-diversity**

The l-diversity principle was defined by *Machanavajjhala et al* as an improvement to the k-anonymity with the objective to make the anonymization not susceptible to homogeneity and background Knowledge Attacks. It is based on a principle defined as Bayes-Optimal Privacy and involves modeling background knowledge as a probability distribution over the attributes and uses Bayesian inference techniques (7-9).

l-Diversity in summary reduces significantly the granularity of the data representation by using generalization and suppression. Returning to our example, consider the inpatient records shown in Table 2 and the 3-diverse version of it:

Table 6. 3-diverse Anonymous Data

ID	Name	Nationality	Age	Zip	Last purchase
1	*	*	⩽40	0175*	Game Server
2	*	*	⩽40	0175*	Game Server
3	*	*	⩽40	0175*	Robot
4	*	*	⩽40	0175*	Robot
5	*	*	>40	0175*	Power XXX
6	*	*	>40	0159*	Game Server
7	*	*	>40	0159*	Robot
8	*	*	>40	0159*	Robot
9	*	*	⩽40	0175*	Power XXX
10	*	*	⩽40	0175*	Power XXX
11	*	*	⩽40	0175*	Power XXX
12	*	*	⩽40	0175*	Power XXX

Following the example, Alice cannot be sure if Bob, an American with about 34-36 years old living in the zip code 01754 that bough a Power XXX, a Game Server or a Robot. As well she cannot be sure that Akira, a 22-year-old Japanese living in the zip code 01752 has bought a Robot or a Power XXX, whether she can eliminate the Game Server by the previous knowledge she may have about her friend.

l-diversity Implementations

An example of a l-diversity Implementation is found in the Clustering Based l-diversity Anonymity Model named CLDPP (Malaisamy 59). A high level overview of its internal implementation is as follows:

Table 7. Clustering based ℓ-diversity Algorithm used in CLDPP

```
Input: a dataset D and a diverse anonymity threshold value l.
Output: the anonymized dataset D∗
Begin Step 1: create a collection of buckets for different sensitive attribute values and the perform sorting based on their sizes resulting in B =b1,b2….bn
Step 2: Return if the number of buckets is < l
Step 3: Let result =∅
Step 4: While (the number of non-empty buckets is ≥ l)
  Step 4:1 Randomly select a record reci from the maximal non-empty bucket b and create it as a cluster c={reci}
  Step 4:2 b=b−{reci}
  Step 4:3 While c <l
    Step 4:3.1 select a record recj from the smaller non-empty bucket so that Information Loss(c∪{recj}) is minimal
    Step 4:3.2 b=b−{recj}
    Step 4:3.3 c=c∪{recj}
  Step 4:4 result=result ∪c
Step 5: While (the number of non-empty buckets is > 0)
  Step 5:1 Randomly select a record reck from the non-empty bucket b
  Step 5:2 b=b−{reck}
  Step 5:3 select a cluster c so that Information Loss(c∪{reck}) is minimal
  Step 5:4 c=c∪{reck}
Step 6: generate anonymous similarity group by using local recoding techniques on each Cluster
Step 7: Return a anonymized dataset D∗
```

Problems with l-diversity - Skewness attack

When the overall distribution is skewed, satisfying *l*-diversity does not prevent attribute disclosure or l-diversity efforts may become non-practical. One of the requirements of *l*-diversity is that each equivalence class has an entropy of uniformly distributed L distinct sensitive values to function correctly. It means that each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough. When some values are very common, the entropy of the entire table may be very low. This leads to the less conservative notion of l-diversity that may not guarantee attribute disclosure (Li 3).

Let’s illustrate with an example: Considering that Privacy is measured by the information gain of an observer. The gain is the difference between the prior belief and the posterior belief. Each belief is denoted by Bn where n is the number of the belief.

B0: Alice knows that Bob may have some issues with his new Power XXX system because he has been acting very stressed and an IT support van came to his house the other day. She finds the anonymized database by searching in the XYX.com website of the results of technical issues with the Power XXX with a value of negative or positive, with positive being the value if a customer has complained about the quality of the equipment or experienced issues with it.

B1: Alice knows that the global distribution of manufacturer problems with the Power XXX in her state is about 1% for positive and 99% for negative results, two values with very different degrees of sensitivity.

B2: Alice looks at the table and finds that Bob is in equivalence class 3 because he is 32 years old. She learns P, the distribution of the sensitive attribute values in this class just by analyzing the table and compares that is considerablely higher than the 1%. Based on P she decides that it is quite likely that Bob is having issues with his newly acquired system.

Table 8. Anonymized data vulnerable to Skewness attack

ID	ZIP	Age	Average income in USD	Manufacturer Malfunction
1	476**	2*	30k	negative
2	476**	2*	40k	negative
3	476**	2*	50k	negative
4	476**	2*	60k	negative
5	4790*	>=40	30k	negative
6	4790*	>=40	80k	positive
7	4790*	>=40	30k	negative
8	4790*	>=40	100k	positive
9	476**	3*	30k	positive
10	476**	3*	30k	positive
11	476**	3*	130k	positive
12	476**	3*	30k	negative
13	4770*	4*	150k	negative
...	...	...	...	...
10,000	488**	>=60	30k	negative

In this case, 2-diversity is unnecessary for an equivalence class that contains only records that are negative, rendering the l-diversity not effective.

Problems with l-diversity - Similarity Attack

l-Diversity is not effective when the equivalence class created after applying the *l*-diversity model renders the data into semantically similar groups (Li 3).

Table 9. 3-diverse Anonymous Data Susceptible to Similarity Attack

	Identifying Variable	Quasi-Identifiers				
ID	Name	Nationality	Age	Zip	Salary	Last purchase
1	*	*	⩽40	0175*	$20K	laptop
2	*	*	⩽40	0175*	$30K	mouse
3	*	*	⩽40	0175*	$20K	Laptop case
4	*	*	⩽40	0175*	$35K	laptop
5	*	*	>40	0175*	$80K	Server
6	*	*	>40	0159*	$100K	Cabling
7	*	*	>40	0159*	$30K	Gaming
8	*	*	>40	0159*	$22K	Gaming accessories
9	*	*	⩽40	0175*	$32K	laptop
10	*	*	⩽40	0175*	$28K	mouse
11	*	*	⩽40	0175*	$35K	Laptop case
12	*	*	⩽40	0175*	$32K	laptop

Using the same example of the neighbors Alice and Bob, Alice discovers the summary report with the 3-diverse data and immediately realized that Bob bought a laptop and he must have a salary between $20K and $35K.

In this example, sensitive information leaks may occur because, while *l*-diversity requirement ensures "diversity" of sensitive values in each group, it does not recognize that values may be semantically close.

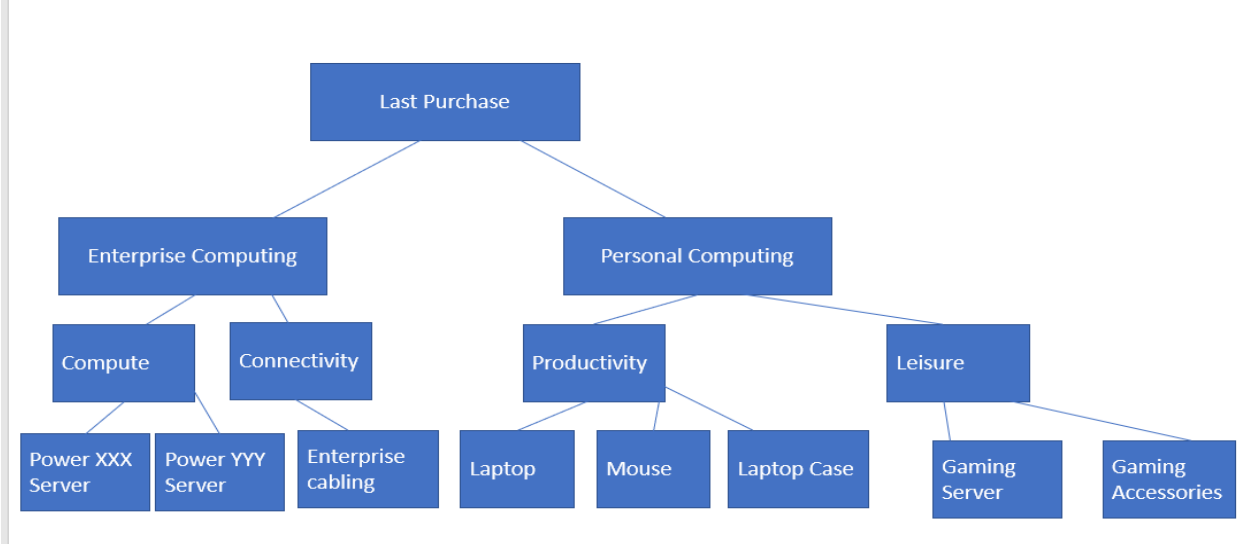
Anonymity: t-closeness

Li et al. (3-4) observed that the l-diversity principle is still insufficient to protect sensitive information disclosure against the skewness and similarity attack described above. To offer a more robust anonymization they introduced the *t*-closeness principle, which requires that the sensitive value distribution in any group differs from the overall sensitive value distribution by at most a threshold *t.*

There is a metric space defined on the set of possible sensitive values, in which the maximum distance of two points (i.e. sensitive values) in the space normalized to 1. For calculating this metric, the method uses the Earth-Mover Distance (EMD) (Liang 3-4), widely used in many areas of computer science. Intuitively, the EMD measures the minimum amount of work needed to transform one probability distribution to another by means of moving distribution mass between points in the probability space. The EMD between two distributions in the (normalized) space is always between 0 and 1.

Using a simplified example for the last purchased product attribute, we use the hierarchy in Fig. 1 to define the ground distances. For example, the distance between “Power XXX” and “Power YYY” is 1/3, the distance between “Power XXX “and “Enterprise Cabling” is 2/3, and the distance between “Power XXX” and “Laptop Case” is 3/3 = 1. Based on the EMD calculation described in (Li 7) the distance between the distribution {laptop, mouse, laptop case} and the overall distribution is 0.5 while the distance between the distribution {laptop, laptop case, enterprise cabling} is 0.278 (the objective is to illustrate that EMD is derived from a distance mathematical equation).

Fig. 1. Example of Hierarchy for “Last Purchase” to Calculate EMD (Li 8)



The key for *t*-closeness is the usage of the EMD measure to take into consideration the semantic closeness of attribute values when building the anonymization instead of only using generalization of quasi-identifier and suppression of records. So, instead of suppressing a whole record, one can hide some sensitive attributes of the record; one advantage is that the number of records in the anonymized table is accurate, which may be useful in some applications. Following the example, Table 10 now reflect the result of executing *t*-closeness into the dataset of reference. The table now shows 0.167-closeness for Salary and 0.278-closeness for Last Purchase. For detailed calculations refer to Li (5-7).

Table 10. 0.167-closeness for Salary and 0.278-closeness for Last Purchase

Identifying Variable	Quasi-Identifiers				
Name	Nationality	Age	Zip	Approximate Salary	Last Purchase
*	*	⩽40	0175*	$36K	Productivity
*	*	⩽40	0175*	$108K	Enterprise
*	*	⩽40	0175*	$60K	Leisure
*	*	>40	0175*	$72K	Laptop
*	*	>40	0159*	$133K	Enterprise
*	*	>40	0159*	$96K	Enterprise
*	*	⩽40	0175*	$48K	Laptop
*	*	⩽40	0175*	$84K	Enterprise
*	*	⩽40	0175*	$120K	Leisure

t-closeness Implementations

Examples of *t*-closeness algorithms to explicitly control non-discrimination in databases is shown in the dMondrian Anonymize Algorithm (Ruggieri 110-112) which is an implementation of t-closeness with the objective of removing social discrimination hidden in data and the subsequent dSabre Anonymize Algorithm (Ruggieri 114).

Problems with t-closeness: Difficult to calculate if more than one sensitive attribute are present

According to the *t*-closeness author the method is not effective when Multiple Sensitive Attributes are present. For example, the presence of two sensitive attributes X and Y in the data. We can consider the two attributes separately, i.e. an equivalence class E has t-closeness if E has *t*-closeness with respect to both X and Y. Another approach is to consider the joint distribution of the two attributes. To use this approach, we have to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between *t* and the level of privacy becomes more complicated (Li 7-8)0.

Problems with t-closeness: Limitations of the EMD

The *t*-closeness uses EMC as the distance measures but has shown flaws in various scenarios. For example, EMD between the two distributions (0.01, 0.99) and (0.11, 0.89) is 0.1, and the EMD between (0.4, 0.6) and (0.5, 0.5) is also 0.1. However, the change between the first pair is much more significant than that between the second pair (Li 10).

Summary of Observations

Table 11 summarize the key observations.

Table 11. Anonymity Models Compared

Model	Technique	Advantages	Vulnerabilities
k-anonymity	Generalization and Suppression	Effective against identity disclosure by removing the links to a dataset with less than 'k' values.	Susceptible to Homogeneity and/or Background knowledge attack
l-diversity	Generalization and suppression based on three parameters: the value that appears more recurrently, the entropy and the recursive diversity (the sensitive values in each class do not occur either too frequently or too rarely)	Provides a greater distribution of sensitive attributes within the group. Is an enhancement of k-anonymity.	It can be redundant and laborious to achieve. This technique may be too prohibitive in the case of low entropy of entire table when only a few values are the same. It is susceptible to skewness attack and similarity attack as it is inadequate to avoid attribute exposure due to the semantic relationship between the sensitive attributes.
t-closeness	Involve the reduction in the correlation between the quasi-identifier and the sensitive attributes based on the calculation of the distance between the distributions using EMD.	Protects against homogeneity and background knowledge attacks. It identifies the semantic closeness of attributes.	Using EMD measure in t-closeness is hard to achieve, especially when there is more than one sensitivity value. Requires that the sensitive attribute spread in the equivalence class to be close to the attributes in the table.

Conclusions and Future Studies

The anonymization techniques summarized in Table 11 are not applicable to non-numerical and non-categorical records thus alternative techniques should be used. What was observed during this analysis is that most of the work on anonymity to preserve personal data privacy is focused on numerical or categorical data. However, there exists specific data domains such as strings, text, or collection of good/services associated with the medical record. The anonymity on this type of data was not explored in this article and will probably imply different techniques.

The anonymization techniques summarized in the Table 11 are only applicable for Relational Data. This research used examples of relational data types; table datasets that consist of records with a fixed number of attributes. However, many real-world applications do not use relational data, i.e. images, XML data, text coming from social media, text coming from doctors and labs analysis, images, etc. Per the literature, the discussed methods are not effective when dealing with this type of non-relational data types (Neamatullah 1).

t-Closeness solves the vulnerabilities of *k*-anonymity and *l*-diversity but is not perfect. As a conclusion of the comparison between the anonymity models; the t-closeness principle has been accepted as an enhanced principle that fixes the main drawbacks of k-anonymity and l-diversity which are the vulnerabilities against homogeneity, background knowledge, skewness and similarity attacks. Recent works have revealed that t-closeness implementations, as well as the k-anonymity and l-diversity, may be vulnerable to new type of attacks that involve analysis of the implementation of different anonymity models to the same dataset. Those are the Minimality and Composition attacks. In the minimality attack (Wong 543-544), the attacker exploits the principle that should be approached for any anonymization mechanism; the need to define some notion of minimality or a limit beyond that the anonymization model cannot generalize, distort or suppress data. Then the attacker can use the knowledge of the privacy model and the minimality principle to infer sensitive information for some equivalence classes. In the Composition Attack (Ganta 1-2) an adversary may gain access to the different releases of anonymized datasets and attempt to join them in order to breach personal privacy of individuals by performing composition between the anonymized tables.

Alternatives methods for future studies are Differential Privacy, Data Transformation and Synthetic Data Generations. To address those new discovered vulnerabilities, researchers are putting effort toward the notion of the use of randomization of privacy mechanisms. Other methods for future study are Differential Privacy (Ganta 1) (Dwork 1) and m-confidentiality model (Xiao 1-2). Another effort toward preservation of Privacy when doing Distributed Data Mining is described in (Clifton 1) which may offer guidance for aggregation of anonymized datasets to avoid their vulnerability to attacks such as minimality and composition. In this respect other studies reveal the possible effectiveness of executing data transformations in the datasets to make it free of sensitive inferences (Wang 1)0. Another recommendation would be to explore the notion of Synthetic Data Generation. Instead of using the real data of the population, a mapping program is executed to propose synthetic data that mimics the original patterns of the population while providing privacy guarantees. The synthetic data will work as a surrogate of the original data but will preserve the original patterns maintaining its original utility.

References

Note: The bibliography used is in great measure related to the Health Industry in US as the privacy of Anonymizing Personal Identifiers in electronic health records for this industry is critical and reinforced strictly with regulations.

Anderson, Ross. *A security policy model for clinical information systems*. In Proc. of the 1996 IEEE Symposium on Security and Privacy, pages 30-43, Oakland, CA, May 1996.

Clifton, Chris; Murat Kantarcioglu; Jaideep Vaidya. *Tools for Privacy Preserving Distributed Data Mining*. ACM SIGKDD Explorations Newsletter, Volume 4 Issue 2, December 2002,Pages 28-34

Dwork, C. *Differential Privacy: A Survey of Results*. In Proceedings of the In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP). 2006, pages 1-12.

El Emam, Khaled. *Heuristics for De-identifying Health Data*. IEEE Security & Privacy, Volume: 6 , Issue: 4 , July-Aug. 2008.

El Emam, Khaled. *Risk-Based De-Identification of Health Data*. IEEE Security & Privacy, Issue No. 03 - May/June 2010 vol. 8

El Emam, Khaled. *Methods for the de-identification of electronic health records for genomic research*. Genome Medicine 20113:25.

El Emam, Khaled; Fida Kamal Dankar. *Protecting Privacy Using k-anonymity*. J Am Med Inform Assoc. 2008 Sep-Oct; 15(5): pages 627–637.

Ganta, S. R., S.; Prasad; A. Smith. *Composition Attacks and Auxiliary Information in Data Privacy.* In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.

Gardner, James; Li Xiong. *Hide: An Integrated System for Health Information De-identification*. 2008 21st IEEE International Symposium on Computer-Based Medical Systems.

Gervais, Thomas J.; Robert M. Siragusa; Prasanna V. Sundaram; Joan L. Knighton. *Systems and methods for de-identification of personal data.* United States Patent No US 8,069,053 B2.

Kushida, Clete A.; Deborah A. Nichols; Rik Jadrnicek; Ric Miller; James K. Walsh; Kara Griffin. *Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies.* Medical Care Vol. 50, No. 7.

Landi, William A, R. Bharat Rao. *Systems and methods for encryption-based de-identification of protected health information*. US Patent No US7519591B2.

Lefevre, Kristen; David J. Dewitt; Raghu Ramakrishnan. *Incognito: Efficient full-domain k-anonymity.* University of Wisconsin Madison.In: SIGMOD, pp. 49–60 (2005)

Li, Ninghui; Tiancheng Li ; Suresh Venkatasubramanian. *T-Closeness: Privacy Beyond k-anonymity and l-diversity.* 2007 IEEE 23rd International Conference on Data Engineering

Liang, Hongyu; Hao Yuan. *On the Complexity of t-Closeness Anonymization and Related Problems.* 18th International Conference, DASFAA 2013, Wuhan, China, April 2013, Proceedings, Part I

Machanavajjhala, Ashwin; Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam. *L-diversity: Privacy Beyond k-anonymity. Cornell University.* ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, Publication date: March 2007.

Malin, Bradley; Latanya Sweeney. *How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems.* Journal of Biomedical Informatics, Volume 37, Issue 3, June 2004, Pages 179-192.

Malaisamy, A; G. M. Kadhar Nawaz. *Clustering Based l-diversity Anonymity Model for Privacy Preservation of Data Publishing.* International Journal of Enhanced Research in Science, Technology & Engineering. ISSN: 2319-7463, Vol. 5 Issue 11, November-2016

Mcgraw, Deven. *Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data*. Journal of the American Medical Informatics Association, Volume 20, Issue 1, 1 January 2013.

Neamatullah, Ishna; Margaret M Douglass; Li-Wei H Lehman; Andrew Reisner; Mauricio Villarroel; William J Long; Peter Szolovits; George B Moody; Roger G Mark; Gari D Clifford. *Automated de-identification of free-text medical records*. BMC Medical Informatics and Decision Making.

Rajendran, Keerthana; Manoj Jayabalan; Muhammad Ehsan Rana. *A Study on k-anonymity, l-diversity, and t-closeness.* Techniques focusing Medical Data. IJCSNS International Journal of Computer Science and Network Security, 172 VOL.17 No.12, December 2017

Ruggieri, Salvatore. *Using t-closeness anonymity to control for non-discrimination*. Transactions On Data Privacy 7 (2014) 99–129.

Sweeney, Latanya. *Achieving k-anonymity privacy protection using Generalization and suppression.* International Journal of Uncertainty, Puzziness and Knowledge-Based Systems Vol. 10, No. 5 (2002) 571-588 © World Scientific Publishing Company

Sweeney, Latanya. *Guaranteeing anonymity when sharing medical data, the Datafly system.* Proceedings, Journal of the American Medical Informatics Association. (AMIA). Washington, DC: Hanley & Belfus, Inc., 1997.

Sweeney, Latanya*. k*-anonymity: a Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Wang, Ke; Benjamin C. M. Fung; Philip S. Yu. *Handicapping Attacker's Confidence: An Alternative to k-Anonymization.* School of Computer Science, Simon Fraser University, BC, Canada, V5A 1S6;2IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA

Weitzman, Elissa R; Liljana Kaci; Kenneth D Mandl. *Sharing Medical Data for Health Research: The Early Personal Health Record Experience*. J Med Internet Res. 2010 Apr-Jun; 12(2): e14.

Wong, Raymond Chi-Wing; Ada Wai-Chee Fu; Ke Wang. *Minimality attack in Privacy Preserving Data Publishing.* In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), pages 543-554, Vienna, Austria, 2007.

Xiao, X.; K. Yi; Y. Tao. *The Hardness and Approximation Algorithms for L-Diversity*. In Proceedings of the 13th International Conference on Extending Database Technology (EDBT), Lausanne, Switzerland, 2010.

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.