Federated Clouds: A New Metric for Measuring the Quality of Data Anonymization

Youssoupha Gaye[1(✉)], Maissa Mbaye[1], Dame Diongue[1], Ousmane Dieng[2], Emmanuel Adetiba[3,4], and Joke A. Badejo[3]

[1] Laboratoire d'Analyse Numerique et Informatique/LANI, CEA-MITIC, Gaston Berger University, 234 Saint-Louis, PB, Senegal
{gaye.youssoupha,maissa.mbaye,dame.diongue}@ugb.edu.sn
[2] Power-management and Real-Time Systems Lab, University of Pittsburgh, Pittsburgh, PA 15260, USA
oud5@pitt.edu
[3] Covenant Applied Informatics and Communication African Center of Excellence, Covenant University, Ota, Ogun State, Nigeria
{emmanuel.adetiba,joke.badejo}@covenantuniversity.edu.ng
[4] HRA, Institute for Systems Science, Durban University of Technology, Durban 1334, South Africa

Abstract. Federated cloud has emerged as solution for cloud service providers to get scalability in serving the growing demand for cloud resources. In a federated cloud, a cloud member can provide service or request it from other cloud provider members in the federation. The federation enables its cloud provider members to be able to satisfy a service beyond the resources they owned by using the resources market in the federation. Data privacy is a major concern in federated clouds. As the privacy regulations and laws of the countries in the federation may vary, it is difficult to assess and confirm that they are in compliance. This makes protecting privacy even more challenging. Privacy management strategies primarily involve anonymization, cryptography, and data splitting. Anonymization is the traditional approach to preserving privacy, which aims at masking the link between the quasi-identifier and sensitive data. The most widely used anonymization techniques are k-anonymity, l-diversity and t-closeness. However, there is a lack of a formal metric to measure the quality of the anonymization process in terms of its ability to prevent re-identification. This paper examines the issue of assessing anonymization quality and introduces a new metric, $Mmaq$, for this purpose. It can be used to evaluate the anonymization of one or multiple attributes. The metric is a combination of the Shannon index, which measures diversity, and a stabilizer factor, which corrects the Shannon index for pathological cases. The initial results suggest that $Mmaq$ can be used to classify attributes as identifier, quasi-identifier, and anonymous. Furthermore, it can be employed as a Cloud Privacy Policy anonymization compliance checker.

Keywords: Data Anonymization Metric · Data Privacy · Federated Cloud Security · Federated Cloud · Cloud Computing

O. Habachi et al. (Eds.): UNet 2023, LNCS 14757, pp. 17–30, 2024.
https://doi.org/10.1007/978-3-031-62488-9_2

1 Introduction

Cloud computing offers Content Providers and Machine Learning Service Developers with highly accessible resources on the Internet. The next step in cloud computing is Federated Cloud Computing, which is designed to provide more efficient data access and sharing. This type of cloud computing is made up of resources from multiple clouds that are in different administrative domains. It is intended to address the increasing demand for computing and storage resources and to enable scalability for cloud providers to collaborate. In a federated cloud, members can both provide and request services such as computation, storage, and machine learning. This federation allows cloud providers to access resources beyond their own, enabling them to satisfy a service. Users of the federated cloud can be either simple users or other non-member clouds that require resources from the federation. Each cloud provider can offer data storage, machine learning, computation, and security services, while the federation provides access and authentication services, service discovery, negotiation, security, load balancing, and user interfaces (web, API, CLI). Federated cloud technology for medical applications can involve the sharing of personal and electronic medical records, diagnostic results, genomic data, and more. For example, Federated Genomics (FEDGEN) [2] is a federated cloud infrastructure in Nigeria for genomic data research on malaria and breast cancer in Africa. In a multi-national federated cloud, cloud providers' infrastructure can be located in different countries, and thus subject to different privacy laws and regulations. Providing federated cloud services presents three main privacy-related challenges: managing data authorizations, formalizing privacy protection legal rules, and developing privacy legal compliance verification services. There is a need to define a quantifiable metric to measure the quality of privacy protection measures in the federated cloud, particularly when transferring data between cloud members in different countries. Data anonymization is one approach used to ensure privacy, which involves modifying data [10] before sharing to prevent the identification of the data owner. However, anonymization should provide enough information to be useful while managing the risks of re-identification. The reconciliation of these two constraints leads to a risk of privacy-related attacks [16]. This paper proposes the use of the Shannon index and the stabilizer to assess the quality of anonymization of personal data in federated clouds. The Shannon index will be used to measure the diversity of the data, and the stabilizer factor will adjust the metric for any pathological cases of the Shannon index. The early results demonstrate that it is possible to classify attributes as identifier, quasi-identifier, and anonymous using $Mmaq$. This allows a Cloud Privacy Policy checker to use it to evaluate compliance in terms of anonymization. This paper is structured as follows. Section 2 provides the background and fundamental concepts of the research. Section 3 outlines the privacy protection strategies used in the federated cloud. Section 4 examines the proposed privacy protection approach and Sect. 5 summarizes the findings and potential future work.

2 Background

2.1 Federated Cloud Architecture and Services

A Federated Cloud is an aggregation of resources from different clouds that are in different administrative domains. These clouds provide services such as computation, storage, machine learning, and hosting that can be requested and managed at the federal level. Each cloud can provide services for other clouds in the federation and seek their resources for its customers. The architecture of federated clouds can be horizontal or vertical. In a horizontal federation, Provider-to-provider SLA (Service Level Agreement) enables collaboration between cloud providers in a peer-to-peer manner. In this architecture, each cloud must negotiate with other clouds to gain access to resources. The Vertical Federated Cloud architecture, on the other hand, includes a Federal/Federated Cloud Broker that manages internal and external interactions. This broker has interfaces with individual clouds via their local cloud brokers to discover and access resources and the user interface. All interactions in the federation are specified in a Federal Service Level Agreement (F-SLA). The F-SLA outlines interconnection rules and describes the responsibilities and authorized behavior of each member of the federation, as well as financial, administrative, or other sanctions in the event of non-compliance with the terms of the agreement. The F-SLA is managed by the federal broker, who is also responsible for negotiating with national brokers when federation users request data or resources. The Federal Broker hosts a variety of services (see Fig. 1):

- **Access Control**: This service provides Authentication and Access to members of the cloud federation;
- **Service Discovery**: This service will provide a map of the services and resources available within the federation;
- **Negotiation Service**: This service is for SLA establishment between clouds members for resources reservations, use and management;
- **Security service**: This service enforces the security policy at the federation level and prevents the propagation of security problems in the federation.
- **Load Balancing and Monitoring services**: Load balancing facilitates the allocation of tasks between cloud providers that are involved in a process. A monitoring service is used to collect data for the oversight of Federation resources.

Clients can access Federation clouds at the cloud level through an interface, such as a RESTful-API, Web application, or CLI, depending on the requested service platform. The Federal Cloud Broker is used to transparently obtain services. In recent years, the security and protection of privacy in the federated cloud has become a topic of increasing interest in the research community. Bernsmed et al. [4] identified four new security issues arising from the formation of federated cloud services: a longer chain of trust, limited auditability, the risk of malicious service components, and liability and legal issues. To address challenges related to security, interoperability, storage, processing, privacy, etc.,

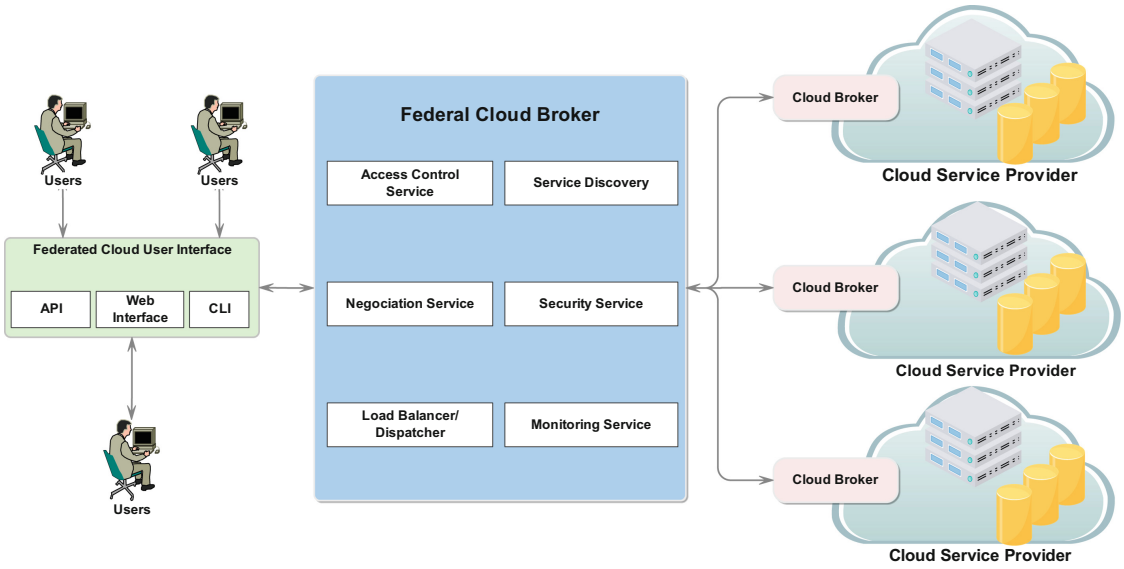

Fig. 1. Federated Cloud Vertical Architecture

several federated cloud architectures have been proposed in the literature. The RESERVOIR [13] project seeks to create an architecture to promote the adoption of open-federated cloud computing, address the scalability issues inherent in single-vendor cloud computing environments, and tackle the difficulties associated with interoperability between different cloud service providers. Additionally, the project aims to address the lack of integrated support for business service management in current cloud computing offerings. The ATMOSPHERE [6] project, a collaboration between Europe and Brazil, focuses on the development of a secure federated cloud platform to process sensitive data, particularly in the health field. The OPTIMIS [9] project proposes a specification and a toolkit for generating federated clouds, with security as one of the non-functional requirements that determine the final composition of a service. BioNimbuZ [14] is a federated cloud computing platform designed to meet the needs of bio-informatics applications, which often require high processing capacities and long run-times. This platform offers the ability to integrate and control multiple cloud infrastructure providers, each offering tailored bio-informatics tool chains as services.

2.2 Problem Statement

Federated clouds present a range of security issues, particularly in terms of privacy management. For instance, the services provided by the cloud federation often include storage and computational resources that are also used by other cloud services, such as Machine Learning. Storage services are widely used by data and content providers to enable low latency access and increased bandwidth. Machine Learning services may necessitate the transfer of data from one cloud data provider member to a geographically closer storage memory for improved efficiency. It is essential that sensitive or confidential information is kept safe in

national domain clouds, in accordance with the data protection laws and regulations of the country. Most countries' laws state that sensitive and private data must not be stored outside their borders until the privacy and access confidentiality requirements have been fulfilled. Verifying and assessing compliance with privacy laws is a major issue in federated cloud research. This is due to discrepancies in legal data privacy policies and the need for anonymization, which can reduce the value of the data. Figure 2 shows a common federated cloud architecture with two cloud members in different jurisdictions. Cloud brokers in each country are responsible for coordinating activities at the federal level. They enable communication between the cloud members and the federation, allocating cloud resources to the federation members in accordance with Service Level Agreements (SLAs). Additionally, they check for compliance with local laws on security and privacy before transferring data to other clouds. However, data transfer between clouds in the federation presents several security challenges, such as managing data authorizations, formalizing privacy protection legal rules, and developing privacy legal compliance verification services.

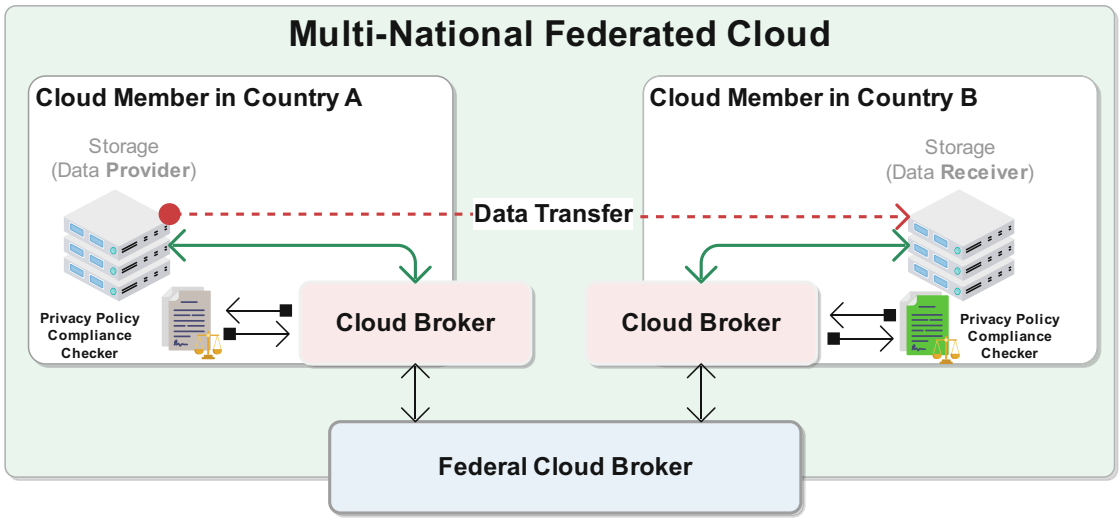

Fig. 2. Federated Cloud data transfer between cloud members in two countries

This paper presents an objective metric, $Mmaq$, to assess data privacy by evaluating the quality of data anonymization between clouds in the same nation or between two countries with different laws. The metric is based on the Shannon index to measure diversity and a stabilizer factor to address pathological cases. Initial results suggest that it can be used to classify attributes into identifiers, quasi-identifiers, and anonymizers, and thus could be employed by a privacy policy checker for automated evaluation.

3 Privacy Preserving Methods in Federated Cloud

Privacy protection in the federated cloud can be achieved through anonymization, cryptography, and data splitting [7]. Anonymization is a privacy-preserving

technique that seeks to obscure the connection between quasi-identifiers and sensitive data. The most commonly used anonymization methods are k-anonymity, l-diversity, and t-closeness. In terms of data privacy, data columns can be divided into four categories based on their level of anonymization:

- **Identifier:** This is an attribute that uniquely identifies an individual (e.g. national identification number) and can be used as a key to distinguish each record;
- **Quasi-identifier:** This is a set of attributes that, combined with external data, can potentially identify the owner of the data (e.g. age, gender). Quasi-identifier attributes must be anonymized to avoid re-identification.
- **Sensitive attribute:** This represents information that if published can be prejudicial to the individual, such as health status, salary, religion, etc. This metric is subjective.
- **Anonymous attribute:** This is an attribute that make it difficult re-identification. If an attribute is totally anonymous no re-identification is possible.
- **Equivalence classes:** These are sets of all records composed of the same quasi-identifier values.

Therefore, measuring the quality of anonymization means classifying data sets into one of these categories. The identifier has the worst anonymity level, as it reveals the identity of the owner. Total anonymous attribute can provide best privacy protection since the identity of the owner cannot be retrieved.

3.1 *K*-Anonymity

K-anonymity [15] is the most widely used privacy preservation method. This model guarantees that each tuple of quasi-identifier values appears at least k times in the table to be published. It uses two main techniques: deletion for identifiers and generalization for quasi-identifiers.

Definition 1. *Let $T(A_1, ..., A_n)$ be a table, and QI_T, the quasi-identifiers associated with it. T satisfies k-anonymity if, for each quasi-identifier $QI \in QI_T$, each sequence of values in $T(QI)$ appears at least with k occurrences in $T(QI)$.*

Table 1 presents an example of a non-anonymized dataset of fictitious patient health information. It contains five attributes and nine records. To implement k-anonymity for a given value of k, two common techniques are used: deletion and generalization.

- **Suppression:** An attribute or part of an attribute is deleted and replaced by special characters such as the asterisk (*);
- **Generalization:** Here, the process of anonymization involves substituting the values of an attribute with a more general but still semantically appropriate value.

Table 1. Example of k-anonymity, where k=2 and QI={Age, Gender}

NIN	Name	Age	Gender	Disease
25580917	Khady	28	F	Hepatitis C
60608183	Fatou	51	F	Malaria
47833705	Cheikh	34	M	Syphilis
42662361	Nogaye	63	F	Measles
86276474	Serigne	40	M	VIH
60275595	Samba	30	M	Tuberculose
75392358	Abdou	42	M	Gonorrhoea
60258041	Bintou	25	F	Cholera
36001413	Ramata	64	F	Meningitis

Original data

NIN	Name	Age	Gender	Disease
25*	*	20-30	F, M	Hepatitis C
60*	*	20-30	F, M	Cholera
60*	*	20-30	F, M	Tuberculose
47*	*	35-45	F, M	Syphilis
86*	*	35-45	F, M	VIH
75*	*	35-45	F, M	Gonorrhoea
60*	*	50-65	F, M	Malaria
42*	*	50-65	F, M	Measles
36*	*	50-65	F, M	Meningitis

Anonymous data

Table 1 has been anonymized to a degree of three with respect to the attributes `'Age'`, and `'Gender'`. Applying k-anonymity to a data set can create equivalence classes that do not provide enough variety for the values of the sensitive attribute, which can result in the disclosure of sensitive information. k-anonymized data can be exposed to attacks such as the Attribute Linking Attack, the Homogeneity Attack, and the Background Knowledge Attack [12].

L-diversity [12] and t-closeness [11] are extensions of the k-anonymity model that protect against attribute disclosure. An equivalence class respects the l-diversity constraint if it contains at least l "representative" values for a sensitive attribute. An equivalence class satisfies the constraint of t-closeness if the distance between the distribution of a sensitive attribute in that class and the distribution of the attribute in the complete table does not exceed a threshold.

3.2 Differential Privacy

Differential privacy, first introduced by [8], is a method of safeguarding individual records in a database by ensuring confidentiality.

Definition 2. *A randomized function k gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subset Range(K)$,*

$$Pr[K(D_1) \in S] \leq e^{\epsilon} * Pr[K(D_1) \in S] \quad (1)$$

where $Pr[K(D_1) \in S]$ means "the probability of observing S as a result of executing k on the database D_1".

This technique enables the extraction of useful information from databases that contain personal data of individuals while protecting the anonymity of their identities. This is usually accomplished by creating a system that adds random noise to the query results, making it impossible for an adversary to determine if a particular record is included in the database or not, regardless of any additional information they may have.

4 A New Metric to Measure Anonymization Quality

We introduce here our proposition, $Mmaq$ (Metric to Measure Anonymization Quality), which is a measure of anonymization quality. It is based on the Shannon

index for measuring diversity and a factor that corrects it for pathological cases where Shannon cannot differentiate the state of the data.

4.1 *Mmaq*: Definition and Properties

The metric is used to measure the quality of anonymization of data for one or more attributes of a dataset.

Definition 3. *The Mmaq is defined as follows:*

$$Mmaq = \begin{cases} P & \text{if } H = 1 \\ (\frac{1}{1-H}) \times P & \text{otherwize.} \end{cases} \tag{2}$$

Where H is the Shannon index and P is the correction factor.

When the $Mmaq = 1$, we have a **totally anonymous** attribute. If $Mmaq \longrightarrow 0$, then we have an **identifier**, whose metric value is $\frac{1}{N^N}$, where N is the number of lines. When $Mmaq \in]0, 1[$, the attribute is a **quasi-identifier**.

As S approaches N, the number of categories increases, leading to a decrease in anonymization quality as attributes become associated with an identifier.

4.2 Measuring Diversity of Data

The Shannon index [16], originally developed to measure the amount of information transmitted in communications, was later used in ecological research to quantify the diversity of a biological community. This index is a combination of two components: the number of species in the community (richness) and the relative frequency of these species (equitability). In this study, the index is used to measure the diversity of a column or a combination of columns in a dataset. The index H is calculated using the following formula:

$$H = -\sum_{n=1}^{s} p_i \log p_i \quad where\ p_i = \frac{n_i}{N}; \quad N = \sum_{i=1}^{s} n_i; \quad H_{max} = \log(S) \tag{3}$$

where p_i is the relative frequency of a category i in the attribute values and S is the number of categories, n_i is an absolute frequency of a Category, and N the total number of records. The maximum value of H is noted H_{max} and is reached when all categories are equally represented (same frequencies).

H increases as the number of categories increases ($S \longrightarrow N$) and the relative representation of each category becomes more balanced ($\forall i, j \in [1, S], p_i \approx p_j$). $H = 0$ when the column contains only one category (the same value in all rows). The higher H, the dataset is diversified. The maximum value of H_{max} will occur when all categories have the same frequency, in which case all p_i are equal ($\forall i, j \in [1, S], p_i = p_j$) and also when S = N.

To illustrate the properties of H, we used it in different synthesized datasets (like in Table 1) with seven attributes generally sensitive to privacy information: national identity number, surname, first name, gender, date of birth, place of

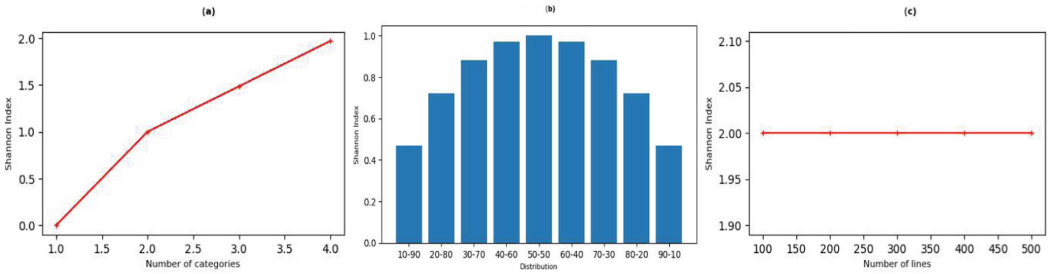

Fig. 3. Diversity index in relation to its variables

birth, disease, and hospital identifier. We worked mainly on an extract of 500 lines of data shows some of the attributes. The results are shown in Fig. 3.

In summary, for totally anonymized attributes $H = 0$. By increasing the number of categories, the anonymized attribute becomes a quasi-identifier, i.e. the number of categories tends towards the number of lines in the dataset, so the value of H increases. The column is an identifier when $S = N$ ($\forall\ i, j \in [1, N],\ if\ i \neq j \Rightarrow val_i \neq val_j$). By applying generalization, to an attribute in a dataset attribute, the number of categories decreases, making the quasi-identifier tend towards a totally anonymous attribute.

Figure 3 shows that if the frequency for each category is $\frac{N}{S}$, then the value of H is the same for 100, 300, or 500 lines (Fig. 3c). When categories have the same absolute frequencies, the $H = H_{max} = log(S)$. On the other hand, if there is only one category, $H = H_{min} = 0$.

So, H can be used to recognize an identifier and a totally anonymous attribute. However, there is a pathological case : $H = H_{max}$, when $\forall i, j \in [1, S],\ freq_i = freq_j$. It includes an identifier and an equally distributed attribute (Fig. 3b). So H gives the maximum for the worst and best cases. To correct this, we introduce another factor so the combination can be used to categorize attributes.

4.3 The Stabilizer Factor

The stabilized factor adjusts the value of H in order to distinguish Identifier and evenly distributed categories for a given attribute.

Definition 4. *Let p be the relative frequency that represents the probability that a value is in category $C(i)$. The stabilizer factor P is defined as follows:*

$$P = \prod_{i=1}^{s} p(x = C(i)) \quad where \quad p(x = C(i)) = \frac{n_i}{N} \quad and \quad \sum_{n=1}^{S} n_i = N \quad ; \qquad (4)$$

Where x is a value in the column, C is the set of categories, $C(i)$ is the category number $i \in [1, S]$, n_i the absolute frequency of $C(i)$ so $S \leq N$.

When $P = 1$, we have a **totally anonymous** attribute. If $P \longrightarrow 0$, then we have an **identifier**, whose stabilizer value is $\frac{1}{N^N}$. When $P \in]0, 1[$, the attribute

is a **quasi-identifier**. We describe the monotony of P considering the data distribution to ensure that it can correct H in pathological cases.

Case 1 - Equitable distribution among categories

– if $S = N$ then $n_1 = n_2 = n_3 = ... = n_S = 1$

$$P = \frac{n_1}{N} \times \frac{n_2}{N} \times \frac{n_3}{N} \times ... \times \frac{n_S}{N} = \frac{1}{N^N}$$

– if $S \in]1, N[$ then $n_1 = n_2 = n_3 = ... = n_S = a > 1$

$$P = \frac{n_1}{N} \times \frac{n_2}{N} \times \frac{n_3}{N} \times ... \times \frac{n_S}{N} = (\frac{a}{N})^S$$

Lemma 1. *Let D_1 and D_2 be two equitable distributions of the same attribute $C(i)$ with, respectively, S_1 and S_2 as a number of categories so that $S_1 < S_2$.*

$S_1, S_2, N \in \mathbb{N}^*$. $n_1 = n_2 = ... = n_{S_1} = a_1$; $m_1 = m_2 = ... = m_{S_2} = a_2$

$$\sum_{i=1}^{S_1} n_i = \sum_{i=1}^{S_2} m_i = N \ ; \ P_1 = \prod_{i=1}^{S_1} \frac{n_i}{N} = (\frac{a_1}{N})^{S_1} \ ; \ P_2 = \prod_{i=1}^{S_2} \frac{m_i}{N} = (\frac{a_2}{N})^{S_2}$$

$$S_1 < S_2 \Rightarrow P_1 > P_2 \tag{5}$$

Proof. Let $S_1, S_2, N \in \mathbb{N}^*$; $S_1 < S_2$

$$\frac{P_1}{P_2} = \frac{(\frac{a_1}{N})^{S_1}}{(\frac{a_2}{N})^{S_2}} \text{ however } N = a_1 S_1 = a_2 S_2, \text{then}$$
$$= \frac{(\frac{a_1}{a_1 S_1})^{S_1}}{(\frac{a_2}{a_2 S_2})^{S_2}} = \frac{(\frac{1}{S_1})^{S_1}}{(\frac{1}{S_2})^{S_2}} = \frac{{S_2}^{S_2}}{{S_1}^{S_1}} > 1$$

$$\frac{P_1}{P_2} > 1, \text{ then } P_1 > P_2$$

Conclusion: This mathematical proof demonstrates that when it comes to equitable distribution, the more categories there are, the lower the stabilizing factor's value becomes.

Case 2 - Non-equitable distribution among categories

This is the case when there are at least two categories that do not have the same frequency.

Lemma 2. *Let D_1 and D_2 be two non-equitable distributions of the same attribute $C(i)$ with, respectively, S_1 and S_2 as the number of categories, so that $S_1 < S_2$.*

$S_1, S_2, N \in \mathbb{N}^*$. $(n_{1,i})_{i=1}^{S_1}$ *and* $(n_{2,i})_{i=1}^{S_2}$ *with* $\begin{cases} n_{j,i} \in N^*; \ j \in \{1,2\}; \ i \in [1, S_j] \\ \sum_{i=1}^{S_j} n_{j,i} = N \end{cases}$

$$P_1 = \prod_{i=1}^{S_1} \frac{n_{1,i}}{N} \; ; \; P_2 = \prod_{i=1}^{S_2} \frac{n_{2,i}}{N}$$

$$S_1 < S_2 \Rightarrow P_1 > P_2 \tag{6}$$

Proof. The proof of lemma 2 is available on GitHub (https://urlz.fr/nnyU).

Conclusion: We have the same property.
Whether the distribution is fair or not, when $S_1 < S_2$ then $P_1 > P_2$.

4.4 Evaluation of $Mmaq$

$Mmaq$ have been calculated for a synthetic data set with seven relevant attributes with sensitive information about privacy: National Identity Number (NIN), surname, first name, sex, date of birth, place of birth, disease, and hospital identifier. The dataset we used was composed of 200 lines of data. We use the values of $Mmaq$ to classify attributes into relative identifiers, relative quasi-identifiers, and relative anonymous attributes. A relative identifier is associated with a specific database, while a global identifier is unique across the world and can be used to identify entities or individuals regardless of the database.

Table 2 shows the classification of the attributes in our dataset. Each row shows the name, category, metric value, and class of an attribute.

Table 2. Classifaction of attributes based on $Mmaq$

Name	Category	Mmaq	Class
NIN	200	8.758e-247	Identifier
FirstName	134	4.433e-171	Quasi-identifier
PlaceBirth	101	4.470e-157	Quasi-identifier
IdHospital	1	1.0	Anonymous
FirstName + Name	195	2.172e-228	Quasi-identifier
FirstName + Name + DateBirth	200	8.758e-247	Identifier

The number of categories is equivalent to the number of lines, and the $Mmaq$ value is nearly zero, thus it is an identifier. For instance, the attribute 'NIN' has two hundred different values out of two hundred lines, so it is classified as an identifier. When the attribute is completely anonymous, the metric value is 1. This is the case for the 'IdHospital' attribute. The $Mmaq$ decreases for the attributes 'PlaceBirth' and 'FirstName', which have 101 and 134 categories respectively, making them quasi-identifying attributes.

A drawback of our categorization technique is that it only permits us to perform local operations, meaning that the identifiers, quasi-identifiers, and anonymizers are specific to the dataset.

5 Related Works

Federate the clouds and you will also federate and scale the security problems. These latter are much more challenging in a federation of clouds due to two reasons: it inherits security issues from traditional clouds and new security problems will rise. In fact, even if the inherit security issues are well studied and most of them solved, they still remain challenging in a federation since a security issue is solved does not mean that it is implemented or properly implemented in a cloud provider of the federation. More importantly, some security problems are fundamentally a consequence of the federation of different clouds. These new issues are: longer chain of trust, limited auditability, the risk of malicious service components, liability and legal issues, and shared data privacy [4,18]. The most important and most challenging remain the privacy of shared user personal data.

Many approaches have been proposed to preserve user privacy when exchanging data in the context of clouds. However, traditional privacy management solutions cannot be applied as is in the context of cloud federation because - i) typical management of privacy and data-utility requirements must be extended to support multiple datasets and data owners [18]; ii) the lack of trust among cloud federation members on de-anonymization attacks protection [18]; iii) unclear user data privacy law to apply to what Cloud service due to the cross-border nature of federated Clouds [4]. Despite these limitations, promising methods are proposed to overcome the data privacy problem in a cloud federation.

Anonymization techniques are the key methods used to protect data privacy in clouds. Some well-known techniques such as k-anonymity [3,15], l-diversity [12], t-closeness [11], and differential privacy [8] are used for data anonymization in the federated cloud context. In addition to these well-known methods, specific solutions are proposed for a particular domain. For instance, to preserve privacy in a health data cloud, Abbasi and Mohammadi [1] proposed an anonymization approach that involves first removing the least frequent data that presents the greatest risk using a normal distribution and second, apply an improved k-anonymity through a k-means++ algorithm. Still in the medical domain, in a more scaled federated cloud - ATMOSPHERE, Blanquer et al. anonymize medical image data using in-memory and on-disk encryption in a secure environment following the following four steps: the deployment of the virtual infrastructure, the management of sensitive data through trusted execution environments, the distributed training of a classification model on GPUs attached to containers, and the production use of a trained classifier [5]. The most interesting aspect of their solution is that it reduces security risk even in untrusted cloud backends.

Even though anonymization is necessary to preserve privacy, after anonymization, the risk of re-identification is still a serious challenge and needs to be addressed. For this purpose, Taneja et al. [17] developed an approach to reduce the risk of identification on electronic medical data. They proposed a privacy model that combines k-anonymity, l-diversity, t-proximity, and δ-presence technique to identify the best values for the parameters to minimize risk and maximize the usefulness of the information.

Although these approaches focus on the anonymization of data while taking into account their usefulness and the risk of re-identification of already anonymized data. They give no information on the quality of the level of anonymization of research data. To our knowledge, this has not previously been developed in the literature. Our solution makes it possible to assess or judge the quality of data anonymization. This proposal is detailed in the next section, and in the conclusion, a more in-depth comparison with related work is presented.

6 Conclusion and Future Works

In this paper, we proposed a metric to measure the quality of data anonymization in a federated cloud, which takes into account the diversity of data to classify it into one of three categories: identifiers, quasi-identifiers, and anonymous.

The results of our study validated the proposed metric for the federated cloud, allowing data owners or cloud administrators to determine whether the level of data anonymization meets the minimum level of privacy protection. Our method to measure the quality of anonymization of personal data in the context of federated clouds can also be applied to other platforms, such as Big Data.

As part of our future work, we plan to develop a system that uses fuzzy logic to determine how close quasi-identifiers are to anonymous identifiers or attributes. Additionally, we intend to create a tool to assess the level of anonymization of personal data and the risk of re-identification of an individual in a set of anonymized data in a federated cloud.

References

1. Abbasi, A., Mohammadi, B.: A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. Concurrency Comput. Pract. Exper. **34**(1), e6487 (2022)
2. Adetiba, E., et al.: FEDGEN Testbed: a federated genomics private cloud infrastructure for precision medicine and artificial intelligence research. In: Misra, S., Oluranti, J., Damaševičius, R., Maskeliunas, R. (eds.) ICIIA 2021. CCIS, vol. 1547, pp. 78–91. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95630-1_6
3. Andrew, J., Karthikeyan, J., Jebastin, J.: Privacy preserving big data publication on cloud using Mondrian anonymization techniques and deep neural networks. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 722–727. IEEE (2019)
4. Bernsmed, K., et al.: Thunder in the clouds: security challenges and solutions for federated clouds. In: 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, pp. 113–120. IEEE (2012)
5. Blanquer, I., et al.: Federated and secure cloud services for building medical image classifiers on an intercontinental infrastructure. Futur. Gener. Comput. Syst. **110**, 119–134 (2020)
6. Brasileiro, F., Brito, A., Blanquer, I.: Atmosphere: Adaptive, trustworthy, manageable, orchestrated, secure, privacy-assuring, hybrid ecosystem for resilient cloud computing. In: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSNW), pp. 51–52. IEEE (2018)

7. Domingo-Ferrer, J., et al.: Privacy-preserving cloud computing on sensitive data: a survey of methods, products and challenges. Comput. Commun. **140**, 38–60 (2019)
8. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends® Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
9. Ferrer, A.J., et al.: OPTIMIS: a holistic approach to cloud service provisioning. Future Gener. Comput. Syst. **28**(1), 66–77 (2012)
10. George, R.S., Sabitha, S.: Data anonymization and integrity checking in cloud computing. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–5. IEEE (2013)
11. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond kanonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115. IEEE (2006)
12. Machanavajjhala, A., et al.: l-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Disc. Data (TKDD) **1**(1), 3-es (2007)
13. Rochwerger, B., et al.: The reservoir model and architecture for open federated cloud computing. IBM J. Res. Dev. **53**(4), 4 (2009)
14. Rosa, M., et al.: Bionimbuz: a federated cloud platform for bioinformatics applications. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 548–555. IEEE (2016)
15. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1998)
16. Silva, H., et al.: A re-identification risk-based anonymization framework for data analytics platforms. In: 2018 14th European Dependable Computing Conference (EDCC), pp. 101–106. IEEE (2018)
17. Taneja, H., Singh, A.K., et al.: Preserving privacy of patients based on re-identification risk. Procedia Comput. Sci. **70**, 448–454 (2015)
18. Yang, M., et al.: Differentially private data sharing in a cloud federation with blockchain. IEEE Cloud Comput. **5**(6), 69–79 (2018)