Towards trustworthy cloud service selection: monitoring and assessing data privacy

Tania Basso, Hebert Silva, Leonardo Motecchi, Breno de França, Regina Moraes

[1]University of Campinas - UNICAMP - Campinas, Brazil

taniabasso@ft.unicamp.br, hebert.oliveiras@gmail.com

leonardo@ic.unicamp.br, breno@ic.unicamp.br, regina@ft.unicamp.br

Abstract. *Cloud services consumers deal with a major challenge in selecting services from several providers. Facilitating these choices has become critical, and an important factor is the service trustworthiness. To be trusted by users, cloud providers should explicitly communicate their capabilities to ensure important functional and non-functional requirements (such as security, privacy, dependability, fairness, among others). Thus, models and mechanisms are required to provide indicators that can be used to support clients on choosing high quality services. This paper presents a solution to support privacy measurement and analysis, which can help the computation of trustworthiness scores. The solution is composed of a reference model for trustworthiness, a privacy model instance, and a privacy monitoring and assessment component. Finally, we provide an implementation capable of monitoring privacy-related information and performing analysis based on privacy scores for eight different datasets.*

Resumo. *Usuários de serviços na nuvem lidam com um grande desafio para selecionar serviços de diferentes provedores. Auxiliar nessas escolhas tornou-se crítico e um fator importante é o grau de confiança que o usuário pode depositar no serviço. Para isso, os provedores de nuvem devem ser capazes de deixar claro quais recursos estão disponíveis para garantir requisitos funcionais e não funcionais (tais como segurança, privacidade, confiabilidade, justiça, entre outros). Sendo assim, modelos e mecanismos são necessários para fornecer indicadores que possam ser usados para apoiar os clientes na escolha de serviços de alta qualidade. Este artigo apresenta uma solução para apoiar a medição e a análise da privacidade, o que pode auxiliar no cálculo de uma medida de confiança. A solução é composta por um modelo de referência para confiança, uma instância desse modelo para a privacidade de dados e um componente de monitoramento e avaliação de privacidade. Por fim, uma implementação capaz de monitorar informações relacionadas à privacidade foi utilizada para realizar análises com base em medição de privacidade para oito conjuntos diferentes de dados.*

1. Introduction

Cloud computing is an established computing paradigm which allows the sharing of massive, heterogeneous, elastic resources among users. Most organizations

have been using cloud computing to better serve their customers around the world [Douglas Miller 2017], and services are increasingly published in clouds.

Despite all the hype surrounding cloud computing, customers are reluctant to deploy their business in the cloud, mainly due to security and privacy concerns, especially for the possibility that sensitive information is exposed to unwanted parties in case the cloud servers storing such information are compromised [Ahmed and Hossain 2014].

Due to these reasons, it is necessary that cloud users are vigilant while selecting the services and their providers in the cloud [Bedi et al. 2012]. To address this problem, individual users and enterprises should be able to assess cloud services and to select the most trustworthy ones.

Trust is defined differently in distinct areas [Artz and Gil 2007, Cho et al. 2015] and, inspired by the existing definitions, we can define it in cloud computing as *the reliance of a client on a service, that it will exhibit some expected behaviour*. To increase trust, cloud providers should explicitly communicate their capabilities to ensure important functional and non-functional requirements, such as security, privacy, dependability, fairness, transparency, among others. Then, trustworthiness can be defined as the level in which a cloud service meets a set of those requirements, i.e., the worthiness of cloud services for being trusted.

Identifying trustworthy services in cloud environments is a challenge due to several factors, such as the complex and dynamic nature of the cloud, the existence of several types of services (e.g., non-critical or business-critical), the large number of characteristics involved in trustworthiness (e.g., security and interoperability), and the subjective notion of trust. Despite several attempts in the literature to address this issue [Habib et al. 2011, Kuehnhausen et al. 2012], a mechanism that accurately measures cloud service trustworthiness is still missing. In this context, the ATMOSPHERE (Adaptive, Trustworthy, Manageable, Orchestrated, Secure Privacy-assuring Hybrid, Ecosystem for REsilient Cloud Computing) project [1] was conceived. It is an Europe-Brazil collaborative project whose main goal is to provide a solution to assess trustworthiness of cloud applications dealing with data, and to support the development of more trustworthy cloud applications.

In the context of the ATMOSPHERE project, we propose a solution that represents a first step to assess trustworthiness of cloud services. We focus on privacy, which is one of the properties (or attribute) of trustworthiness. The goal is to define a quality model for quality attributes of privacy, which may be used to calculate privacy metrics and compose trustworthiness scores.

The solution is composed of (i) a reference model for trustworthiness attributes and a privacy model instance, and (ii) a privacy monitoring and assessment component, which calculates a privacy score and increases the privacy protection according to established needs. The data privacy protection is provided by anonymization techniques and models (more specifically by the k-anonymity model) and the privacy score is based on the re-identification risk (i.e., the probability of an individual to be identified in the anonymized dataset) [Silva et al. 2018]. The monitoring and assessment component is based on the MAPE-K (Monitor-Analyze-Plan-Execute over a shared Knowledge)

[1] https://www.atmosphere-eubrazil.eu/

reference architecture [IBM Corporation 2006], which allows increasing data anonymity level dynamically. The solution is currently implemented and deployed, showing its feasibility and how it contributes to the composition of a trustworthiness score.

The paper is organized as follows: Sections 2 and 3 present, respectively, relevant concepts and related work that guide our study. Section 4 describes the proposed solution for monitoring and assessing privacy in cloud platforms. Section 5 shows the results of experiments where the proposed solution was applied in real datasets. Finally, Section 6 presents the conclusions and future work.

2. Background

2.1. Privacy and Anonymity

Organizations' interest in privacy protection occurs for two main reasons: to comply with privacy laws and regulations (companies and organizations that hold private data must comply with them), and to address business interests (the more a company protects the privacy of its customers and business partners, the more credibility it gets)[Basso et al. 2016].

Regarding privacy laws, several regulations for the protection of personal information has been established, e.g., PIPEDA (The Personal Information Protection and Electronic Documents Act) in Canada [Office of the Privacy Commissioner of Canada 2018] and HIPAA (Health Insurance Portability and Accountability Act) in USA [U.S. Department of Health & Human Services 2017]. In April 2016, the EU General Data Protection Regulation (GDPR) was approved to harmonize data privacy laws across Europe. Its enforcement occurred in May 2018 [GDPR.ORG 2017]. Similar to GDPR, the *Lei Geral de Proteção de Dados* (LGPD) [Planalto.gov.br 2018] was approved in Brazil and it is supposed to be enforced in December 2020.

Data privacy protection is strongly connected with the idea of preventing information disclosure. Data anonymization, also known as de-identification, consists of techniques to prevent the recovery or leakage of individual information while data is shared and disclosed to the public. The anonymization process is carried out to alter the data before it is disclosed, in a way that prevents the identification of key information [Sedayao 2012].

There are several anonymization techniques that can be applied on data in order to protect the privacy of individual. Some of these existing and most used techniques are generalization (attribute values are generalized to a range in order to reduce the granularity of representation) and suppression (the key attributes or the quasi-identifiers are removed completely to form the anonymized table). Also, anonymization models can be applied to avoid re-identification. The κ-*anonymity* model uses the generalization and suppression techniques to anonymize data in a way that any combination of *quasi*-identifier (i.e.,attributes that can be combined with external information to expose some individuals) appears at least in κ-records in the anonymized dataset. The κ must be a positive integer value and its minimum value is 2. A high value of κ indicates that the anonymized table has a low risk of disclosure. We selected κ-anonymity due to its mention in privacy regulations as LGPD and GDPR.

There is a trade-off between anonymization and data utility as *the higher the anonymity level, the lower the data utility is* [Dwork 2008]. Perfect privacy can be achieved by publishing nothing at all, but this has no utility; perfect utility can be obtained by publishing the data exactly as received from the individuals, but this offers no privacy [Brickell and Shmatikov 2008, Alvim et al. 2011]. The challenge regarding this trade-off is to maximize data utility while satisfying a required level of protection. In this scenario, two measures can be used to derive privacy scores: the *re-identification risk* and the *information loss*. The re-identification risk measures the probability of matching anonymized data using publicly available information or auxiliary data to discover the individual to which the data belongs to. The information loss measures the amount of information that can be obtained about the original values of variables in the input dataset. In this work, both measures are calculated using the ARX tool functionalities [Prasser and Kohlmayer 2015].

2.2. Reference and Quality Models

As trustworthiness can be understood as a multi-dimensional construct combining several properties or characteristics, to assess the trustworthiness of a system it is possible to include security, privacy, coherence, isolation, stability, fairness, transparency, dependability, among others. All of them have other subattributes that expand a lot the possibilities to be addressed. Since several conflicting properties may be involved in the analysis, a technique based on multi-criteria decision-making (MCDM) is needed.

Many MCDM techniques consider a system in terms of properties and their respective relations, which are processed and aggregated in a single score [Martinez et al. 2014]. In the area of dependability, the following two scoring techniques have been successfully used: Analytic Hierarchy Process (AHP) [Saaty 1988] and Logic Score of Preferences (LSP) [Dujmovic and Elnicki 1982]. In this work, LSP was chosen for its capability to assess and compare complex systems and also due to its simplicity when compared with AHP.

A set of quantifiable **attributes** is chosen to characterize the system, e.g., memory usage, throughput. When these attributes represent input measures they must be normalized applying adequate functions. For that, the definition of **thresholds** is necessary once they specify the maximum and minimum values for the inputs of the leaf-level components of the quality model.

The values for each component are influenced by an adjustable element **weight**, which specifies a preference over one or more characteristics of the system, according to established requirements (e.g., in certain contexts memory usage might be more important than throughput). The final score is computed using the aggregation of the weighted values of the attributes, starting from the leaf-level to the root attributes, using **operators** that describe the relation between them.

A possible solution to use the LSP technique is to define a reference model, which in turn can be further instantiated for each required attribute toward a wider system characteristic. Thus, the model can be used to compose the different attributes, weights, thresholds and operators to make it possible to calculate a global characteristic score, allowing to guide the choice between similar systems. This is done by walking through the model tree, from the leaves to its root, aggregating child scores. The instance of the

model for a given attribute represents its quality model (as proposed by ISO/IEC 25000), fulfilling the necessary requirement for the use of the LSP technique.

3. Related Work

To the best of our knowledge, there are few works defining trustworthiness reference models or meta-models for software systems. The majority of the meta-models are focused on a specific characteristic that are related to trustworthiness (and not in full). Also, those works do not define trustworthiness measures.

Zarrabi et al. [Zarrabi et al. 2012] presented a meta-model that combines legal and trust related concepts to enable developers to model and reason about the trustworthiness of a system in terms of its law compliance. Bernardi et al. [Bernardi et al. 2011] proposed a UML profile for quantitative dependability analysis of software systems modeled with UML, with particular focus on reliability, availability, maintainability and safety sub-attributes. Similarly, Biggs et al. [Biggs et al. 2016] propose a SysML profile for modelling the safety-related concerns of a system.

Regarding quality models for privacy, there are some works targeting domain-specific solutions, e.g., mobile applications, location-based services. Xiao et al. [Xiao et al. 2007] present a quality-aware anonymity model for protecting location privacy, where a mobile user can specify the minimum anonymity level requirement. Kim et al. [Kim et al. 2017] define a quality of private information (QoPI) model, which represents common and personalized privacy control in mobile applications. The model considers the user's privacy preferences and the context of using the application.

To calculate privacy scores, we adapted the component presented in [Silva et al. 2018], which evaluates the re-identification risk of an anonymized dataset and, based on predefined risk thresholds, iteratively increases the anonymity level to reduce this risk. Also, it provides the calculation of information loss to inform the dataset utility. The component functionality was adapted to the MAPE-K architecture [IBM Corporation 2006] to analyze the results and decide whether to continue the anonymization process or not.

4. Privacy Monitoring and Assessment

It is important to mention that privacy is only one of several trustworthiness properties that composes trustworthiness scores (other properties can be, for instance, security, fairness, SLA (service Level Agreement), performance, among others). The focus of our work is privacy. Figure 1 presents the solution proposed for monitoring and assessing privacy in cloud platforms, which is based on the MAPE-K reference architecture [IBM Corporation 2006].

In Figure 1, the users or client applications submit queries to the database through the *TDMS (Trustworthy Data Management Services)* (Step 1). TDMS is similar to a database service in cloud systems dealing with protocols and mechanisms for data storage, indexing, distribution, replication, access and management, but this component also considers trustworthiness properties such as privacy. When data stored in this component is retrieved (i.e., a raw dataset as a JDBC ResultSet or CSV file/stream) they must be anonymized by the *Anonymizer* component (Step 2).

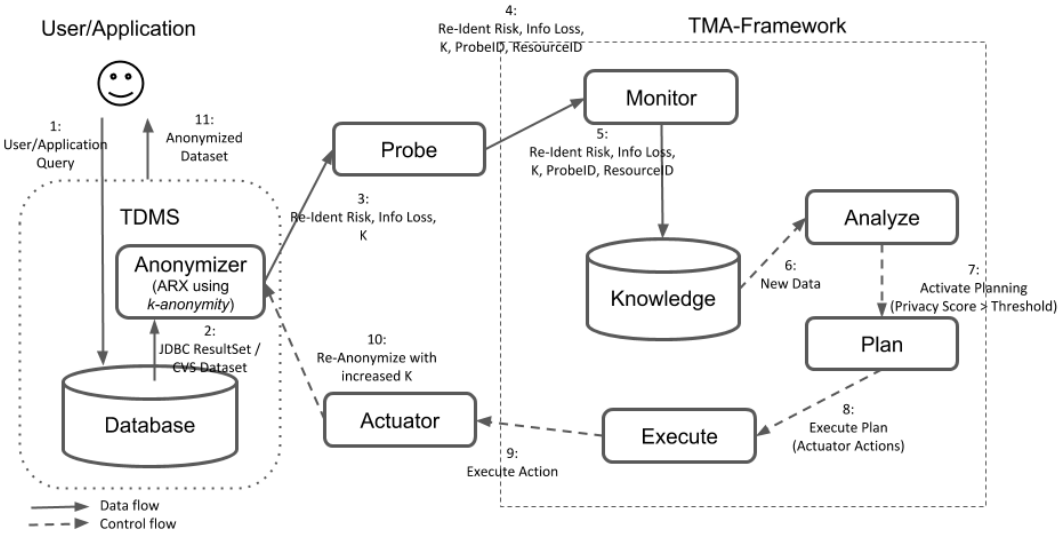

Figura 1. trustworthiness and privacy monitoring and assessment platform

The *Anonymizer* component applies the κ-anonymity model in the input data. The first application of κ-anonymity is done with κ=2, which is the minimum value for κ. Then, the re-identification risk and the information loss are calculated for the anonymized dataset. These calculations are provided using ARX [2] tool features.

Probes are responsible to collect information from the managed system (cloud resources and applications) under the MAPE-K architecture and send it to the *TMA (Trustworthiness Monitoring & Assessment) Framework*, specifically to the *Monitor* component. Thus, after the calculation, the probe creates a message containing the measures for the re-identification risk and information loss, as well as the κ value (Step 3). Each probe is identified by the *ProbeID* property and the resource containing the data (TDMS, in this case) is identified by the *ResourceID* property. This way, the cloud system (ATMOSPHERE platform) is aware of which of its components is under privacy risks.

The *Monitor* is a component that receives the information sent by probes, through a RESTful API (Step 4), and store them at the *Knowledge* repository (Step 5). The *Knowledge* repository stores all the trustworthiness-related information along with definitions of quality models, including their weights and thresholds. This threshold represents the highest risk the input data can assume in the platform. It is usually defined by data source owners and privacy analysts and is used to decide whether the re-identification risk of the input data is acceptable or not. Some works (e.g., [ElEmam et al. 2011]) suggest thresholds from 1% to 5% as acceptable to be used for research data.

The *Analyze* component obtain the most recent measures stored at the *Knowledge* repository and re-calculate the privacy scores (Step 6) based on the privacy quality model (Section 4.1). For every update on the scores, the *Plan* component is activated and evaluates if the privacy score is higher than the established threshold (Step 7). The *Plan*, in turn, calculates the ideal value of κ to anonymize the input data according to the threshold.

[2]https://arx.deidentifier.org/

So, the simplest plan is to increase the value of κ to increase also the anonymity level and, consequently, reduce the risk. These actions are sent to the *Execute* component (Step 8) that, in turn, call the privacy *Actuator* (Step 9) through a RESTful interface. Finally, another round of anonymization is performed (Step 10). This is performed successively, until the re-identification risk is equal to or lower than the threshold (the higher the κ, the lower the risk), when the anonymized data set is delivered (Step 11).

4.1. Privacy Quality Model

The main new contribution of this work is the reference quality model (Figure 2 and its respective instance for privacy (Figure 3).

The *Analyze* component (Figure 1) adopts the LSP technique to calculate trustworthiness scores. It means that a quality model is represented by a weighted tree of attributes. A quality model instance calculates scores by walking through the tree, from the leaves to its root, aggregating child scores using selected operators. Figure 2 presents the reference model for the trustworthiness quality models.

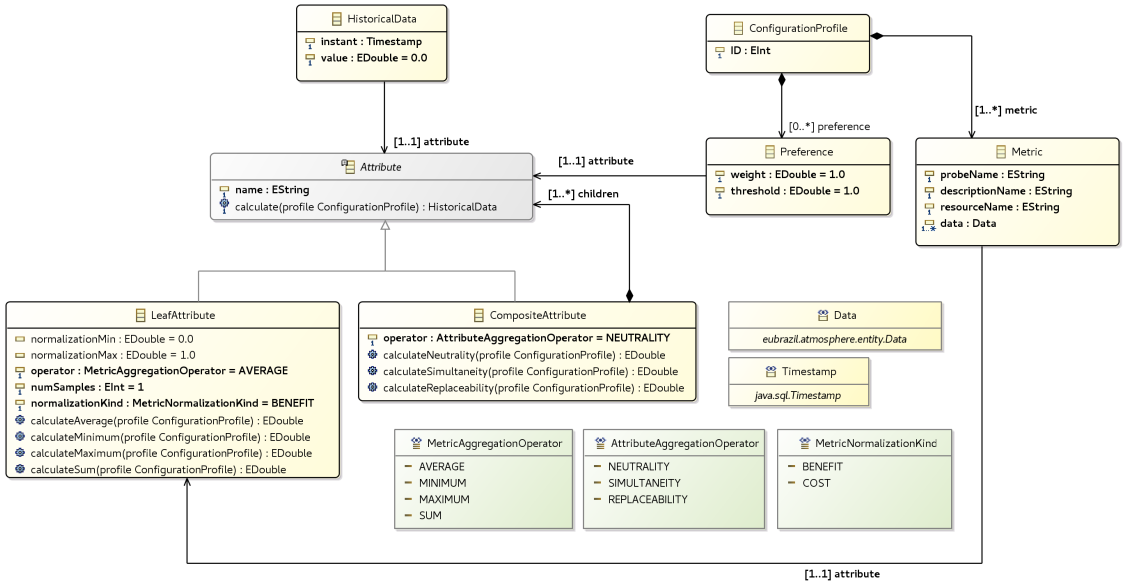

Figura 2. Reference Quality Model

LeafAttribute represents metric definitions and their associated scores are based on the actual measures collected by monitoring probes. They must be normalized to be evaluated against **thresholds** and to ensure operators work at the same scale. This normalization process uses the X_{max} (normalizationMax) and X_{min} (normalizationMin) values along with the measures of leaf attributes. These measures may be aggregated to calculate scores based on simple operators (type *MetricAggregationOperator*) such as **sum, average, minimum**, or **maximum**.

Benefit and **Cost** (type *MetricNormalizationKind*) guide the interpretation of leaf attributes for the normalization process. The first says higher values means better assessment, the latter says higher values worse assessment. For benefit normalization, all values below X_{min} and above X_{max} will always be equal to 0 or 100 respectively, while for cost these notions of maximum and minimum are interpreted in the reverse way.

The central element is the *Attribute* class representing the trustworthiness properties / attributes. The method *calculate* provides the score and uses the *ConfigurationProfile* element and its *Preference* to set weights and compare with thresholds. Every calculate result should be stored as a *HistoricalData*.

The *Attribute* class implements the Composite design pattern so that it can be a *LeafAttribute* or a *CompositeAttribute*, where the former represents metrics available and the latter is a composed or aggregated value towards the global (trustworthiness) score (root of the tree).

The *Analyze* component performs the aggregation of composite attributes scores based on specific operators (type *AttributeAggregationOperator*), including:

- *Neutrality:* refers to the weighted mean and represents the combination of simultaneous satisfaction of requirements;
- *Simultaneity:* refers to a conjunction (i.e., the logical operator *and*), in which all requirements must be satisfied;
- *Replaceability:* refers to a disjunction (i.e., logical operator *or*), in which one of the requirements of the system has a higher priority replacing the remaining requirements.

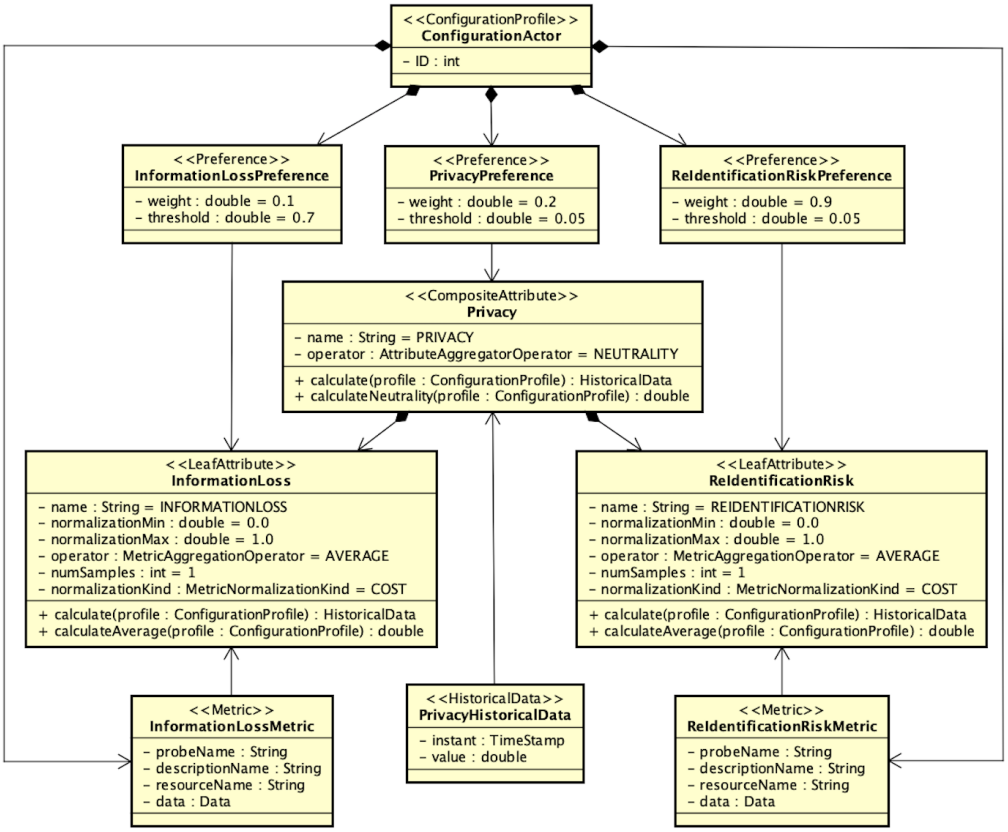

Figura 3. Quality Model for Privacy

The quality model for privacy is an instance of the reference model (Figure 2). Therefore, the composite attribute *Privacy* is the main attribute of this instance and the calculation of the privacy score aggregating the measures for *InformationLoss*

and *ReIdentificationRisk* using the simultaneity operator. Both metrics are obtained by average values and each of them has its own **threshold** (for now, based on the literature) and its own **weight** being *ReIdentificationRisk* responsible by 90% in the composition of privacy score. As in this instance the *normalizationKind* of the *InformationLoss* and the *ReIdentificationRisk* are COST and the operators are AVERAGE, the calculation of the privacy score is given by:

$$PrivacyScore = ((1 - ReIdentificationRisk) * weight + ((1 - InformationLoss) * weight) \quad (1)$$

where weight belongs to the respective <<Preference >>steriotyped classes (ReIdentificationRiskPreference and InformationLossPreference).

It is important to highlight that in the privacy instance the classes steriotyped by <<ConfigurationProfile>>is instantiated as *ConfigurationActor* because in the context of the project the configuration of thresholds can be done by a privacy analyst or automatically by the system when the self-adaptive procedures are running.

4.2. Probes and Actuators

In order to calculate the re-identification risk and information loss, we implemented a wrapper (called Anonymizer) for the ARX anonymization tool. This tool performs the anonymization over an input dataset using the κ-anonymity algorithm and implements statistical risk models to determine these measures. Re-identification risk can be calculated using three risk models representing possible attack scenarios:

- *Prosecutor*: the attacker aims at identifying an specific individual, which is already known to be in the dataset. This is the risk model used in ATMOSPHERE because it represents the role that holds more information about the dataset;
- *Journalist*: the attacker also aims at identifying an specific individual, however, there is no information regarding the individual information in the dataset;
- *Marketer*: the attacker aims at identifying a collection of individuals. The attack succeeds only if a larger fraction of the individuals could be re-identified.

For the information loss, the ARX tool also identifies the global optimum from the solution space of the κ-anonymity algorithm. Based on this global optimum, we calculate a ratio from the lowest and highest scores for information loss. The *Anonymizer* exposes different alternative interfaces to interact with the TDMS services:

- *File anonymization*: an input CSV (Comma Separated Value) file containing the dataset to be anonymized is passed to the *Anonymize*r and an anonymized dataset is returned, along with calculated risks and information loss;
- *JDBC ResultSet anonymization*: a JDBC ResultSet is received as input. Then, the records are anonymized and returned using other in-memory data structure, along with calculated risks and information loss. Such data structure is required as JDBC ResultSet maintains a connection with the database and any modifications would be propagated to the raw data;
- *In-memory data structures*: the same data structure used to return anonymized data from the JDBC ResultSet can be used as an input. This is then anonymized and returned with processed dataset, along with calculated risks and information loss.

After anonymizing the input datasets, the re-identification risk and information loss rate are calculated following the Equations (2) and (3) respectively. The *calculate* method in each respective class (Figure 3) is responsible to implement the calculation of each measure.

$$Risk = \frac{\sum_{i=1}^{\infty} I(F_i=1)}{N} \tag{2}$$

where I is the indicator function and the expression measures the proportion of records in the known dataset list that are unique.

$$C_{DM}(g,k) = \sum_{\vee Es.t|E|\geq k}(|minority(E)|) + \sum_{\vee Es.t|E|\geq k}(|E|) \tag{3}$$

refers to the equivalence classes of tuples in *D* induced by the anonymization g. The first sum computes penalties for each non-suppressed tuple, the second for suppressed tuples.

As already mentioned, in this case, the obtained results are into ranges from zero to one [0,1] and do not need additional normalization processes. Thus, these measures are collected by a probe and sent to the trustworthiness evaluation platform via its monitoring interface.

5. Evaluating data anonymization and privacy scores results

As a feasibility study, we performed some experiments where the proposed solution was applied in eight real datasets. The goal was to obtain the privacy score in different levels of anonymization and respective information loss in order to identify the score that deals better with the trade-off between identification risk and data utility.

We used some datasets provided by UCI [3] and Figure Eight [4] platforms, both with datasets for the machine learning community. From UCI we selected three databases regarding social data: *Adults* data sample extracted from the 1994 US Census database (32561 tuples), *Internet* contains general demographic information on US internet users collected from October through November,1997 (10104 tuples) and *Contraceptive Method Choice (CMC)* subset of the 1987 National Indonesia Contraceptive Prevalence Survey (1473 tuples). From Figure Eight we selected: *Indian Terrorism Death (Terrorism)* sentences from the South Asia Terrorism Portal were selected and the deaths mentioned in it were counted (27233 tuples); *Airline Twitter Sentiment (Airlines)* collected the sentiment expressed in the Twitter, related to the problems in the major US airlines from February 2015 (16000 tuples); *New England Patriots Deflategate sentiment (Deflategate)* collected the sentiment expressed in the Twitter, related to deflated footballs and whether the Patriots cheated, before 2015 Super Bowl (11814 tuples); *Police-involved fatalities (Deaths)* police-involved shootings over a two-year span since May 2013 (2355 tuples); *Medicine Sales (Medicine)* data referring to 1 month of medicines sales (22586 tuples).

In Figure 4, the results for the *Adults* dataset are presented. The κ *value* column represents the κ values of the κ-anonymity algorithm, whose minimum is 2. The *Re-identification risk* and *Information loss* columns represent the correspondent values

[3] https://archive.ics.uci.edu/ml/datasets.html

[4] https://www.figure-eight.com/

calculated for each κ. The *Privacy Score* was calculated according to the Equation (1). The anonymization process was performed for different steps, where the (*Acceptable risk*) is 100%, 50%, 10%, 1%, 0.5% and 0.1% respectively.

Adults				
Acceptable risk	Re-identification risk	*k* value	Information loss	PRIVACY SCORE
100%	0.5000000	2	0.0081	0.549190
50%	0.0600000	4	0.0240	0.943600
10%	0.0600000	4	0.0240	0.943600
5%	0.0082000	48	0.1200	0.980620
0.5%	0.0006470	123	0.2500	0.974418
0.1%	0.0000307	1546	1.0000	0.899972

Figura 4. Results for the *Adults* dataset

In this figure, as the κ value increases (which means that the anonymity level of the dataset increases), the re-identification risk values decrease and the information loss values increase. The highest privacy score (0.980620) refers to κ=48. It is important to observe that at this point the re-identification risk (0.0082000) reached a value lower than the established threshold (0.05) and also the information loss (0.1200) is lower than its threshold (0.7). These results give an indication that the highest privacy score represents the scenario where the thresholds are respected and the best result is obtained considering the balance between data privacy and data utility. In addition, the risk value at this point is very close to 1% corroborating with the threshold value suggested in the literature. Although we have experienced several risk rates, the best scores, in the majority of the datasets (Airlines is the unique exception) were observed for the risk rates up to 5%.

Particularly for this dataset there is one more possibility that still meets both thresholds and presents a score less than 1% lower. For κ=123, although the usefulness of the data is reduced by half, the risk rate is decreased giving an option for the users give up a bit the usefulness of the data to have an extra drop in risk rate.

There are some steps that, even defining different acceptable risk, present the same results (e.g., for 50% and 10% acceptable risk). It happens because the k-anonymity needs to provide a value of k whose respective risk is lower than the acceptable one. In some cases, the value of k satisfies this condition for more than one acceptable risk.

A similar behavior is observed in Figure 5: the highest privacy score (0.926600) refers to κ=5 and indicates the best scenario for the trade-off while respecting both thresholds (re-identification risk = 0.0250000 and information loss = 0.5090)

Medicine				
Acceptable risk	Re-Identification risk	*k* value	Information loss	PRIVACY SCORE
100%	0.33333000	2	0.3355	0.666453
50%	0.33333000	2	0.3355	0.666453
10%	0.02500000	5	0.5090	0.926600
5%	0.00063131	62	0.8587	0.913562
0.5%	0.00063131	62	0.8587	0.913562
0.1%	0.00003070	62	0.8587	0.914102

Figura 5. Results for the *Medicine* dataset

However, in this case (κ=5), this is the only possibility to meet both thresholds with advantages of more than 1% in the privacy score.

Due to space restrictions, the tables of all datasets is not presented. Figure 6 presents, for each dataset, the highest privacy score and the correspondent κ, re-identification risk and information loss. Although the datasets are quite diverse in terms of quantity of records, context and data composition (semi identifiers and data sensitivity), we obtained, for the experiments in this work, similar results.

Scores Results					
Dataset	#Tuples	Re-identification Risk	K value	Information Loss	Privacy Score
Adults	32561	0.0082000	48	0.1200	0.980620
Internet	10104	0.0028000	11	0.2500	0.972480
CMC	1473	0.0100000	13	0.2000	0.971000
Terrorism	27233	0.0434000	5	0.4142	0.919520
Medicine	22586	0.0250000	5	0.5090	0.926600
Airlines	16000	0.0830000	6	0.6309	0.861940
Deflategate	11814	0.0007758	17	0.5874	0.940562
Deaths	2355	0.0104000	5	0.7377	0.916870

Figura 6. Score Results

6. Conclusion

This work proposes a solution to assess the privacy score towards to the trustworthiness of cloud services. The approach defines a quality model formalization and presents an instance for quality attributes of privacy, which may be used to calculate privacy measurement and compose trustworthiness scores.

Eight datasets of two different repositories, quite diverse in terms of quantity of records, context and data sensitivity is used in the experiments and the results are quite promising as a similar behavior was observed for all datasets. The approach shows the usefulness of the quality model to express a score that respects the configuration of the measurement based on the sub-attributes. Also, the usefulness of the reference model to guide the attributes configuration in terms of normalization ranges, thresholds and weights is validated when privacy instance was instantiated and used in the experiments.

As future work, we intended to build instances of the other attributes of trustworthiness and to perform experiments using larger number of datasets. In addition, our intention is to get the trustworthiness score allowing the benchmark of cloud systems.

Acknowledgment

This work has been partially supported by the project **ATMOSPHERE** (https://www.atmosphere-eubrazil.eu/ - Horizon 2020 grant agreement No 777154 - MCTIC/RNP).

Referências

[Ahmed and Hossain 2014] Ahmed, M. and Hossain, M. A. (2014). Cloud computing and security issues in the cloud. *International Journal of Network Security & Its Applications*, 6(1):25.

[Alvim et al. 2011] Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. (2011). Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer.

[Artz and Gil 2007] Artz, D. and Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2):58 – 71. Software Engineering and the Semantic Web.

[Basso et al. 2016] Basso, T., Matsunaga, R., Moraes, R., and Antunes, N. (2016). Challenges on anonymity, privacy, and big data. In *Dependable Computing (LADC), 2016 Seventh Latin-American Symposium on*, pages 164–171. IEEE.

[Bedi et al. 2012] Bedi, P., Kaur, H., and Gupta, B. (2012). Trustworthy service provider selection in cloud computing environment. In *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*, pages 714–719. IEEE.

[Bernardi et al. 2011] Bernardi, S., Merseguer, J., and Petriu, D. C. (2011). A dependability profile within marte. *Software & Systems Modeling*, 10(3):313–336.

[Biggs et al. 2016] Biggs, G., Sakamoto, T., and Kotoku, T. (2016). A profile and tool for modelling safety information with design information in sysml. *Software & Systems Modeling*, 15(1):147–178.

[Brickell and Shmatikov 2008] Brickell, J. and Shmatikov, V. (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 70–78. ACM.

[Cho et al. 2015] Cho, J.-H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):28.

[Douglas Miller 2017] Douglas Miller (2017). Think cloud compliance: an introduction to cloud computing for legal and compliance professionals. https://download.microsoft.com/download/0/D/6/0D68AE95-6414-4074-B4B8-34039831E2BF/Introduction-to-Cloud-Computing-for-Legal-and-Compliance-Professionals.pdf.

[Dujmovic and Elnicki 1982] Dujmovic, J. and Elnicki, R. (1982). A dms cost/benefit decision model: mathematical models for data management system evaluation, comparison, and selection. *National Bureau of Standards, Washington DC, No. GCR*, pages 82–374.

[Dwork 2008] Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer.

[ElEmam et al. 2011] ElEmam, K., Paton, D., Dankar, F., and Koru, G. (2011). De-identifying a public use microdata file from the canadian national discharge abstract database. *BMC medical informatics and decision making*, 11(1):53.

[GDPR.ORG 2017] GDPR.ORG (2017). Eu general data protection regulation (gdpr) portal: Site overview. http://www.eugdpr.org/.

[Habib et al. 2011] Habib, S. M., Ries, S., and Muhlhauser, M. (2011). Towards a trust management system for cloud computing. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pages 933–939. IEEE.

[IBM Corporation 2006] IBM Corporation (2006). An architectural blueprint for autonomic computing. *IBM White Paper*, 31:1–6.

[Kim et al. 2017] Kim, S.-H., Ko, I.-Y., and Kim, S.-H. (2017). Quality of private information (qopi) model for effective representation and prediction of privacy controls in mobile computing. *Computers & Security*, 66:1–19.

[Kuehnhausen et al. 2012] Kuehnhausen, M., Frost, V. S., and Minden, G. J. (2012). Framework for assessing the trustworthiness of cloud resources. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference on*, pages 142–145. IEEE.

[Martinez et al. 2014] Martinez, M., De Andres, D., and Ruiz, J.-C. (2014). Gaining confidence on dependability benchmarks' conclusions through"back-to-back"testing (practical experience report). In *Dependable Computing Conference (EDCC), 2014 Tenth European*, pages 130–137. IEEE.

[Office of the Privacy Commissioner of Canada 2018] Office of the Privacy Commissioner of Canada (2018). The personal information protection and electronic documents act. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/.

[Planalto.gov.br 2018] Planalto.gov.br (2018). Lei geral de proteo de dados (lgpd). http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Mpv/mpv869.htm.

[Prasser and Kohlmayer 2015] Prasser, F. and Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The arx data anonymization tool. In *Medical Data Privacy Handbook*, pages 111–148. Springer.

[Saaty 1988] Saaty, T. (1988). What is the analytic hierarchy process? *Mathematical Models for Decision Support*, 48:109–121.

[Sedayao 2012] Sedayao, J. (2012). Enhancing cloud security using data anonymization. *White Paper, Intel Coporation*.

[Silva et al. 2018] Silva, H., Basso, T., Moraes, R., Elia, D., and Fiore, S. (2018). A re-identification risk-based anonymization framework for data analytics platforms. In *2018 14th European Dependable Computing Conference (EDCC)*, pages 101–106. IEEE.

[U.S. Department of Health & Human Services 2017] U.S. Department of Health & Human Services (2017). Health information privacy. https://www.hhs.gov/hipaa/index.html.

[Xiao et al. 2007] Xiao, Z., Meng, X., and Xu, J. (2007). Quality aware privacy protection for location-based services. In *International Conference on Database Systems for Advanced Applications*, pages 434–446. Springer.

[Zarrabi et al. 2012] Zarrabi, F., Pavlidis, M., Mouratidis, H., Islam, S., and Preston, D. (2012). A meta-model for legal compliance and trustworthiness of information systems. In *International Conference on Advanced Information Systems Engineering*, pages 46–60. Springer.