The Measurement of Data Locations in the Cloud

Bernd Jaeger*, Reiner Kraft‡, Sebastian Luhn†, Annika Selzer‡, and Ulrich Waldmann‡
*COLT Technology Services, Duesseldorf, Germany
Email: bernd.jaeger@colt.net
†Westfaelische Wilhelms-Universitaet Muenster, Muenster Germany
Email: sebastian.luhn@uni-muenster.de
‡Fraunhofer Institute for Secure Information Technology, Darmstadt, Germany
Email: reiner.kraft | annika.selzer | ulrich.waldmann@sit.fraunhofer.de

***Abstract*—If a company uses cloud computing services to process their employees' or their customers' personal data, they need to ensure that the cloud provider complies with the relevant privacy statues. One of the things that need to be ensured is that all personal data are processed only in lawful locations. Data sources that can be used to automatically determine the current location of data processing could help cloud users to fulfill their duty and to strengthen the confidence in a privacy friendly processing of their personal data. For that, data location metrics need to be defined, appropriate data sources need to be determined and the measured data need to be combined reasonable. This paper describes the procedure and system architecture of such data location metrics.**

***Keywords*-Automated Privacy Control, Positioning of Data Processing, Privacy, Metrics.**

I. Challenge and Objective

For companies that want their customers' or employees' personal data to be processed in the cloud, it is essential that these data are adequately protected.[1] Since most cloud computing services are considered to be contract data processing, the cloud user stays responsible for his customers' and employees' personal data even if they are processed by the cloud provider according to European privacy law.[2] Consistently the cloud user needs to verify that the cloud provider takes reasonable technical and organizational measures to protect the relevant personal data. This verification needs to take place before any personal data are being processed by the cloud provider. Additionally, the cloud user needs to ensure on a regular basis that the technical and organizational measures of the cloud provider are sufficient.[3] However, this is problematic because of distributed computing in cloud environments and the geographical distances in between the cloud user and the data centers of the cloud provider. [1]

One approach is the use of automatically verifiable privacy metrics for cloud environments, i.e., metrics for assessing relevant privacy properties based on trustworthy data, with which the degree of implementation of technical and organizational measures of the cloud provider can be controlled continuously.

From a privacy perspective the control of the processing location is particularly significant because the European privacy law regulates strict rules in this field: A processing of personal data by a contract data processor within the European Economic Area [4] is legally privileged and usually permitted when based on a written contract and regular privacy checks. The lawfulness of the processing of personal data outside the European Economic Area is dependent on both a statutory permissive rule or the consent of all relevant data subjects and an adequate level of privacy through international privacy agreement or contracts. [5] This distinction is based on the fact that the privacy level within the European Economic Area is subject to a high and largely uniform protection, that results from the implementation of the European Privacy Directive in the individual states. In contrast, the privacy level of states outside the European Economic Area varies dramatically.

Due to the continuous virtualization of IT systems in globally distributed data centers the clear identification of the processing location is a major challenge for the development of automated privacy checks. Therefore, this paper presents a possible solution for this problem based on automatically verifiable privacy metrics.

II. Theoretical Basis

A framework specifying terms and techniques is used for the development of privacy metrics. Its theoretical basis is rooted in two scientific disciplines. Measurement theory comprises the measurability of aspects of reality and is thus the foundation of all measurements. In social sciences, however, phenomena shall be quantified that are not measurable directly. This also applies to privacy

[1] This paper has been written within the scope of the project VeriMetrix which is funded by the BMBF, registration number: 16KIS0053K.

[2] In general, whoever processes personal data is responsible for this specific data processing. To stay responsible for the processing of personal data of a contract data processor is a special characteristic of the so-called contract data processing. Compare Sec. 11 of the German Federal Data Protection Act as one example for the European point of view.

[3] Depending on the relevant categories of personal data this means, that the cloud user needs to control the cloud provider within time frames of 6 months to 3 years.

[4] The European Economic Area is a free trade area between the European Union, Iceland, Liechtenstein and Norway.

[5] E.g. a data processing in the U.S. is permitted if the data subject gave his consent to process his data in the U.S. and his personal data are protected by international privacy agreements such as Safe Harbor.

metrics, as they measure the adherence of a system to external requirements. E.g., they have to determine whether a virtual machine is located at an admissible location (cf. section I). As this cannot be measured directly, indirect hints about the virtual machine's location have to be used. From social sciences, the concepts of constructs and indicators, detailed in section II-B, are used to combine these indirect hints and approximate the real situation.

A. Measurement Theory

The foundation of all measurements is measurement theory [2]. It comprises the question whether aspects or phenomena of reality can be measured at all and thus is a main part of scientific theory. While statistical methods are used to infer information from a (usually large) data basis, measurement theory is concerned with modeling reality and thus the way data can be gathered in the first place. Established by Stevens in [3], it was significantly augmented by Suppes [2].

There are two different schools of thought in measurement theory, called representative and operational measurement theory, respectively. The former one is also called classical measurement theory [2]. Measuring according to this theory means approximating the *structure of reality* by numbers, i.e., creating a numerical representation of the (empirical) reality. In particular, modeling the structure of reality means that relations of real-world objects are projected onto their numerical representations. Empirical object and measurement method are independent of each other. An important task of representative measurement theory is to show that a chosen method suits the object to be measured. The *operational* measurement theory, opposed to the representative one, states that an object is defined solely by the measurement method and thus is one with it. A relation to reality is optional, as it is not part of the theory [4].

B. Constructs and Indicators

A general problem of representative measurement theory is to determine which mathematical transformations of real-world objects should be allowed without changing the relations of different objects. This is especially the case when dealing with complex phenomena where it may not be clear what is measured exactly. Additionally, there may be phenomena where it is known that no direct measurements are possible. That is especially the case for privacy metrics, as the limitations of direct measurements are given by technology and thus are mostly known. Still, statements on these phenomena may be desirable. Thus, a method is needed to make statements on complex, not directly measurable phenomena on measurement-theoretic foundations.

Social sciences developed the concepts of *constructs* and *indicators* for this. A construct is the phenomenon on which a statement is to be made, but which is not measurable directly. For this, indicators are identified that are semantically related to the construct and are measurable directly. Several indicators are combined with statistical methods and thus allow for an approximation of the construct.

This concept of indirect measurement means that only the operational measurement method can be used, not the representative one [5]. This is due to the fact that the construct is not directly measurable by definition and thus, does not allow for any measurability considerations. Furthermore, the choice, loading, and combination method of the indicators—i.e., the operationalization of the indicators—follows semantic considerations only. Operationalizations can be validated by statistical methods as well. For this, the influence of each indicator on the result is examined via test data. For this test data, the result is known. For example, the processing location of personal data for which measurements are taken could be known. Using this test data, the influence of each indicator on the result can be extrapolated.

C. Related Work

The technology program Trusted Cloud that is funded by the Federal Ministry for Economic Affairs and Energy has published a paper [1] that proposes a legal framework for cloud computing certifications based on a standardized catalogue of legal requirements. The catalogue should be uniformed within the European Economic Area. The certificate should be produced by an independent third party certifying that the inspection has been carried out as required by law. The aptitude of the certifying body should be documented by accreditation. The certifying body should be liable for erroneous certificates. [1] In [6] an automated, continuous privacy certification has been proposed to expand the legal framework suggested by the technology program Trusted Cloud. The automated, continuous certification could be based on secure log data and enables the cloud user to react promptly to potential privacy incidents.

The development of metrics for continuous and automated verification of privacy requirements is a new challenge both theoretically and practically. There are occasional attempts to test privacy metrics for organizational matters, such as the number and processing time of privacy requests and infringements but mostly there are proposals for information security metrics rather than privacy metrics, e.g. [7], [8], [9], [10], [11]. Sowa [12] gives an extensive overview over the definition, development and use of information security metrics. These definitions are mainly influenced by the ones of the National Institute for Standards and Technology (NIST) [10]

To ensure that all personal data are processed only in lawful locations, the events and operations of the relevant virtual machines (VM) can be analyzed. The location of virtual machines can be determined by identifying the data center. [13] This however requires the participation of the cloud provider. By using alternative methods the location of virtual machines can be determined without the participation of the cloud provider: For example, users of VMs can determine the virtual coordinates of surrounding network nodes by measuring the round-trip-time to a

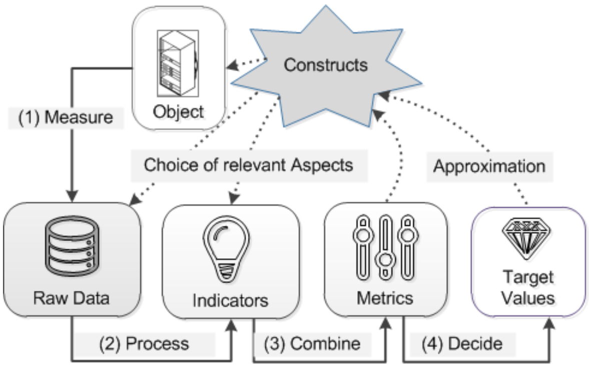

Figure 1. Framework for Privacy Metrics

network node and back ("Round Trip Time", RTT).[14]

This technique is similar to that in the distance-bounding protocols which are introduced by Brands and Chaum [15]. Furthermore, various measurement data can be combined to so-called fingerprints, in order to analyze them as a whole and determine the probable location of the investigated object. For this purpose, statistical tools and machine learning methods [16] are used, in order to correlate currently measured fingerprints to reference data of known locations, without need to semantically understand the fingerprint information. Such heuristically detected fingerprints have already been used to identify neighbored VM locations. [17] [18]

III. A Framework for Privacy Metrics

The goal of the framework presented here is to meet the specific requirements of privacy metrics for cloud applications. The method of using metrics within IT companies has been described in the last section. It can be seen that metrics in this environment are directly measurable most of the time, even if the desired information is not. There is no explicit concept of constructs (cf. section II-B). That metrics can be used to infer the desired information is either implicitly stated or is done verbally. An operationalization (cf. section II-A), i.e., a choice and loading of indicators, is usually not done.

On contrast, the framework presented in this chapter defines metrics such that they approximate constructs, i.e., the operationalization of indicators is part of a metric itself. In the following, the framework is presented. For this, terms and relations of these terms are mentioned first. Afterwards, an example of application is given with the determination of processing locations. Details on the statistical methods used, i.e., the analytical model, follow. Lastly, an exemplary calculation method is described.

A. Terms and Relations

The main goal of the framework is to combine the theoretical foundation and concepts with the domain-specific requirements of privacy metrics for cloud applications. The first step is to define all terms used to construct metrics as well as the relations of these terms. An overview is given in figure 1.

All terms used in this framework are taken from the literature on IT metrics and social sciences, respectively. The foundation of each measurement is the *object of investigation*. This could be a virtual machine on a server in the cloud. The object of investigation has attributes that are measured. *Quality criteria* are used to determine whether a measurement is useful. These are criteria such as reliability and validity. The result of a measurement is *raw data*, which are stored for further processing by *instruments*. Through this processing, they become context-dependent indicators for a construct that is to be approximated.

This construct, already mentioned in section II-B is the main part of the framework and describes the actual desired information. Especially when dealing with privacy metrics, this information is not directly measurable as it mostly based on data protection or regulatory laws and requirements, see section I. An example for this, determining the processing locations of data, is described in the next section. Due to the fact that direct measurements are impossible, the construct has to be approximated by measurable quantities. Metrics derived from these constructs usually report the fulfillment level of the construct's requirements as well as a confidence level stating how reliable the statement on the fulfillment level is. The *goal* of a metric is to fulfil the requirement as good as possible. *Target values* determine whether that goal is reached.

Metrics are constructed by combining several *indicators*. The combination method is determined by the *analytical model*, detailed in section III-C. Part of an analytical model are one or more *decision criteria*, helping to determine the fulfillment level of the metric.

B. Example of Application: Determination of Processing Locations

In this framework metrics are designed by combining top-down and bottom-up approaches. The top-down approach can be exemplified by starting with the metric construct of processing location of customer's data. This construct results from the data protection requirement which stipulates that data must be stored and processed solely in selected countries with an assured high data protection level. This directly results from the legal regulations, but does not include the means by which this question might be answered.

On basis of this construct a metric can be drafted that quantifies the fulfillment level of the requirement that all data are processed solely and throughout in permitted locations. Subsequently, indicators have to be found that have a semantic relationship to this metric. In contrast to the top-down approach the definition of indicators implies that suitable raw information about the object of investigation, i.e. a specific virtual machine, can be collected and processed to meaningful location indicators.

- Measure the minimal latency time to nearby servers, in order to determine the physical location.
- Record the communications paths / hops used while communicating with reference servers.

- Generate virtual machine environment (VME) fingerprints (virtual hardware information, driver software versions, virtual network configuration, local network environment like gateways, ...).
- Carry out DNS requests that result in different responses depending on respective locations.

The possible data sources need to be checked according to their information content, complexity and cloud providers acceptance. For example, the acceptance of cloud providers for scanning the machine environment is probably low because these methods may be considered intrusive and risky to data and infrastructure.

The aim of generating meaningful indicators, however, is accompanied by the aim of establishing a broad reference database. Generally, a variety of indicators is useful to verify that a given location is compliant to the privacy requirements of a cloud user. The more indicators are available, the more reliable is the result of this calculation. Meaningful indicators should be considered even if stronger indicators were available in order to allow cross-validation.

C. Analytical Model

The essential part of a metric is the analytical model which describes the operationalization of indicators, i.e. their combination in the calculation of metric values. The relevance of each indicator and measurement method for the approximation of a given construct is expressed as well as the validity of the resulting metric. The appropriateness of an analytical model depends both on semantical considerations and on its validation with test data.

As constructs in the context of cloud privacy metrics are resulting of data privacy laws and regulatory requirements, the analytical model has to classify the given indicator values according to their degree of requirement fulfilment. Therefore, static classification methods like decision trees are used for this purpose. The semantic of these instruments is simple: for each classification decision the causal indicators are obvious. Furthermore, decision trees are constructed by supervised learning [22] so that the operationalization of indicators can be checked statistically as described in chapter II-C.

Generally, indicators can be grouped into such that give direct information on location (group I) and such that give only fingerprint information on a location (group II). The greater the consistency of both groups of indicators, the clearer is also the determination of location. Optimally, if all indicators of group I and group II lead to identical results, it has a high probability that the location is correctly determined, and also the assessment of its compliance to the privacy requirements of a cloud user. More difficult is the determination whenever the individual indicators give different results. In this case, for example, if the fingerprint comparison (group II) is negative with simultaneous positive result from group I, the confidence of the classification method used for the fingerprint comparison can support a decision.

Ideally, the calculation method for the determination of the data processing location is able to immediately adapt changes in the information which is used for the location fingerprinting. Furthermore, it should be implemented flexible to include new measurement and calculation methods e.g. new algorithms for the comparison of fingerprints.

IV. System Architecture

The metric system is used by cloud users that want to check continuously whether their privacy requirements are currently met. The architecture for measuring the processing location should take this into account. Furthermore it should take into account that the cloud provider and the cloud user may have different interests when it comes to the measurement of the processing location. The provider is usually not interested in transferring raw in-house data about the cloud infrastructure to external evaluating components and may fear the uncontrolled spread of information. He also could be interested to filter out or manipulate critical data. In contrast, the cloud user wants to receive original unaltered information, but does not want to share user-specific data with other cloud users of the same cloud provider.

Figure 2 shows the main components of the system architecture.

The figure shows on the left the cloud provider's environment that hosts a virtual machine (VM) with customer's data and a reference VM of the metric system. These VMs include cloud-provider-independent VeriMetrix agents that measure the environmental parameters which are described in section III and sends them to the central component ("VeriMetrix Collector"). Generally, reference VMs can be deployed at known locations to create comparative values. These are stored in the reference database (shown on the bottom right). The metrics module of the collector processes the measuring data, in order to calculate indicators and the final metric results. It uses both for classifying new measuring data: reference data that are measured in parallel (if available) as well as previously verified data from the reference database, in order to determine the probable VM processing location.

One aim of this metric system is the reliable and sufficient proof of privacy infringements such as unauthorized modifications of the processing location. The ability to configure the appropriate warnings and to get proofs by viewing the underlying database are therefore important features of the user interface. The customer may directly retrieve metric via a web service or may feed the metric results to existing applications. An auditor may be commissioned to evaluate measurement data, to add some data from non-automatic checks of privacy measures and finally issue evaluation reports to the customer. An administration interface on the part of the auditor configures and controls the measuring modules, for example the extend and frequency of measurements.

The question of whether and how the measuring data should be protected, not only affects privacy but also general IT security. This results from the fact that known

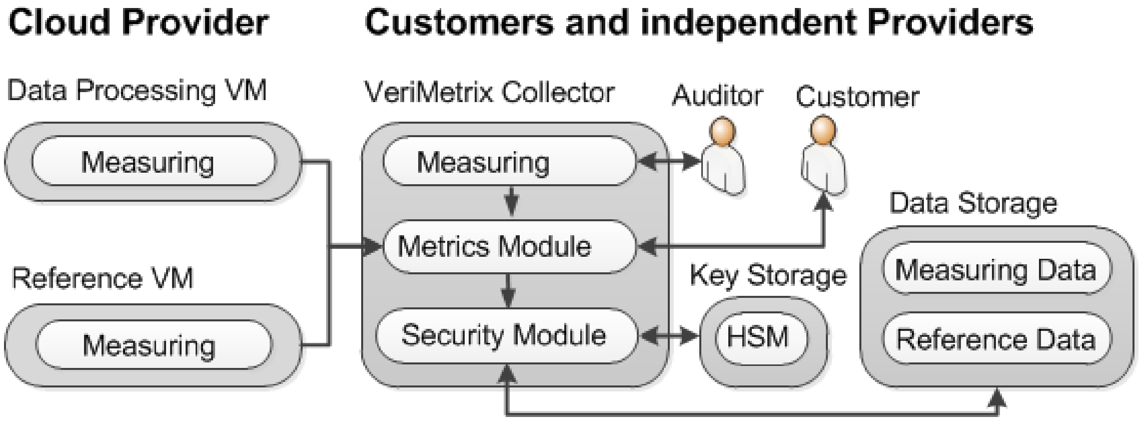

Figure 2. Architecture of the VeriMetrix System

locations of VMs can be the base for an attack.[6] For example, if environmental information of specific VMs become known to hackers, attack VMs may be started in the same environment and therefore with higher chance of success . [17] Therefore, the components use Transport Layer Security (TLS) for data exchange. The collector component provides a security module (see on the bottom center of figure 2) that transfers the measuring and reference data in a secured format for later use.

The secured format is described in [6] and provides for authenticity, integrity and confidentiality of the data according to a method defined by Schneier and Kelsey. [19] This process ensures forward integrity and secure data storage even in untrusted environments. Therefore, each new data entry is concatenated to the previous entry on encryption, hashes and checksums and is then written into a current measurement file. The security module and the VeriMetrix keystore mutually agree on basic keys each time a new data file has to be secured. From these keys and coded access rights the security module derives one-time keys to encrypt the current data entry. All keys are deleted after use. Only the keystore keeps the basic keys and is able to recalculate on demand the specific encryption keys for authorized external entities (e.g. to an auditor).

The collector provider may outsource both, the keystore and the database to a cloud service, even to the tested cloud provider himself. The keystore may work in form of an HSM service, which is configurable by the auditor rather than the cloud provider.[7]

V. Response to unlawful locations

If the result of the metrics reveals abnormalities in regard to the locations, the cloud user needs to legally interpret the result and respond to the infringement. For this, the cloud user should first assess the severity of the privacy infringement based on the individual case. Criteria are e.g. the amount of the relevant personal data, the imminent harm for the data subjects and the circumstances of the infringement.

Depending on the result of the severity rating, different reactions to a detected privacy infringement can be considered. Considering all criteria of the individual case the first reaction may be the demand to immediately eliminate the infringement. The elimination of the infringement needs to be verified within an appropriate time frame. If there are substantial privacy violations or if the cloud provider does not eliminate the infringement immediately, the cloud user should terminate the contract and contractual bind the cloud provider to delete his personal data. [20]

If the cloud user uses a cloud storage service he might have further options to react to the privacy infringement: Since downloading and deleting previously stored data is part of the typical functions of this type of cloud services, the cloud user can actively delete or migrate his data to either his own data center or the data center of another cloud provider.[8] [20]

VI. Outlook

As mentioned above, the cloud user is obliged to control the cloud service on a regular basis to ensure that the personal data are processed in a lawful manner. The system of metrics that was introduced in this paper can help the cloud user to verify the technical and organizational measures taken by the cloud provider. By verifying these measures automatically, it can be ensured that they are implemented by the cloud provider lawfully both in the (near) past and in the present. Like this the cloud user is enabled to respond to privacy-related incidents promptly.

Furthermore, such a system of metrics creates the base for user-friendly marketplaces which enables potential cloud users to find and compare privacy friendly cloud services. It is conceivable that these marketplaces also store

- the privacy requirements of individual sectors in which special laws require more stringent privacy requirements for the processing of personal data,
- the privacy properties of cloud services.

[6] Compare BSI IT-Grundschutz Catalogues, Gefährdung G 4.90.

[7] Such functionality are for example provided by the AWS CloudHSM service.

[8] In all three cases it is recommended to contractual bind the cloud provider to delete backups.

Like this, potential cloud users can easily find out whether or not a cloud service guarantees to fulfill their sector-specific requirements. Therefore, such marketplaces can help building trust in cloud services and thereby contribute to an increased usage of this technology.

References

[1] G. Borges *et al.*, "Solutions in the field of data protection law." Cloud Computing Legal Framework Working Group, Oct 2012. [Online]. Available: http://www.trusted-cloud.de

[2] P. Suppes, *Foundations of Measurement*. Elsevier, 2014, vol. 2.

[3] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, 1946.

[4] D. J. Hand, "Statistics and the theory of measurement," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 445–492, 1996.

[5] G. W. Bohrnstedt, "An overview of measurement in the social sciences," in *Presentation at the National Academy of Sciences Workshop on Advancing Social Science Theory: The Importance of Common Metrics, February*, 2010, pp. 24–25.

[6] T. Kunz, A. Selzer, and U. Waldmann, "Automatic data protection certificates for cloud-services based on secure logging," *Springer, Trusted Cloud Computing*, pp. 59–75, 2014.

[7] ISO, *ISO 27004:2009, Information Security Management — Measurement*. International Organization of Standardization (Hrsg.), 2009.

[8] CIS, *The CIS Security Metrics*. The Center for Internet Security (Hrsg.), 2010.

[9] A. Jaquith, "Security metrics: Replacing fear, uncertainty, and doubt," 2007.

[10] NIST, *NIST 800-55: Performance Measurement Guide for Information Security*. National Institute of Standards and Technology (Hrsg.), 2008.

[11] D. A. Chapin and S. Akridge, "How can security be measured," *Information Systems Control Journal*, vol. 2, pp. 43–47, 2005.

[12] A. Sowa, "Metriken, der schlüssel zum erfolgreichen security und compliance monitoring," 2011.

[13] P. Massonet, S. Naqvi, C. Ponsard, J. Latanicki, B. Rochwerger, and M. Villari, "A monitoring and audit logging architecture for data location compliance in federated cloud infrastructures," in *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*, may 2011, pp. 1510–1517.

[14] T. Ries *et al.*, "Verification of data location in cloud networking," *In: Fourth IEEE International Conference on Utility and Cloud Computing*, pp. 439–444, 2011.

[15] S. Brands and D. Chaum, "Distance-bounding protocols," in *EUROCRYPT'93, Lecture Notes in Computer Science 765*. Springer-Verlag, 1993, pp. 344–359.

[16] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

[17] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM conference on Computer and communications security*, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 199–212.

[18] B. Jäger, A. Selzer, and U. Waldmann, "Die automatisierte messung von cloud-verarbeitungsstandorten," *Datenschutz und Datensicherheit – DuD*, vol. 01, pp. 26–30, 2015.

[19] B. Schneier and J. Kelsey, "Secure audit logs to support computer forensics," *ACM Transactions on Information and System Security (TISSEC)*, vol. 2, no. 2, pp. 159–176, 1999.

[20] T. Kunz, A. Selzer, and S. Steiner, "Konsequenzen festgestellter verstöße bei der auftragsdatenverarbeitungskontrolle," *Datenschutz und Datensicherheit – DuD*, vol. 01, pp. 21–25, 2015.