D. Nuñez, C. Fernandez-Gago, and J. Luna, "Eliciting Metrics for Accountability of Cloud Systems", *Computers & Security*, vol. 62, pp. 149-164, 2016.
http://doi.org/10.1016/j.cose.2016.07.003
NICS Lab. Publications: https://www.nics.uma.es/publications

Eliciting Metrics for Accountability of Cloud Systems

David Nuñez[a], Carmen Fernández-Gago[a], Jesús Luna[b]

[a] *Network, Information and Computer Security Laboratory (NICS Lab)*
Universidad de Málaga, Spain
Email: {dnunez, mcgago}@lcc.uma.es
[b] *Cloud Security Alliance, Scotland, U.K.*
Email: jluna@cloudsecurityalliance.org

Abstract

Cloud computing provides enormous business opportunities, but at the same time is a complex and challenging paradigm. The major concerns for users adopting the cloud are the loss of control over their data and the lack of transparency. Providing accountability to cloud systems could foster trust in the cloud and contribute towards its adoption. Assessing how accountable a cloud provider is becomes then a key issue, not only for demonstrating accountability, but to build it. To this end, we need techniques to measure the factors that influence on accountability. In this paper, we provide a methodology to elicit metrics for accountability in the cloud, which consists of three different stages. Since the nature of accountability attributes is very abstract and complex, in a first stage we perform a conceptual analysis of the accountability attributes in order to decompose them into concrete practices and mechanisms. Then, we analyze relevant control frameworks designed to guide the implementation of security and privacy mechanisms, and use them to identify measurable factors, related to the practices and mechanisms defined earlier. Lastly, specific metrics for these factors are derived. We also provide some strategies that we consider relevant for the empirical validation of the elicited accountability metrics.

1. Introduction

The cloud computing paradigm is complex but at the same time, it entails enormous business opportunities. However, this complexity raises concerns, refraining organizations and users to adopt cloud services. According to a recent study from Forrester [1], the lack of transparency and compliance information is among the major concerns from business cloud consumers. Providing *accountability* across cloud ecosystems can enable trust and break barriers to cloud adoption.

Accountability is a complex concept, whose definition varies depending on the discipline where it is applied. We will use the following definition derived from the A4Cloud project [2]: '*Accountability consists of defining governance to comply in a responsible manner with internal and external criteria, ensuring implementation of appropriate actions, explaining and justifying those actions and remedying any failure to act properly*'. According to this definition, one of the important issues for an organization (e.g., a cloud provider) is to show compliance with rules and obligations in a transparent manner, by providing information about internal procedures and policies. Thus, it is of paramount importance for an organization to have mechanisms in place that allow the assessment of accountability, either by themselves (in the cases of self-assessment) or by external authorities, as this supports the demonstration to interested parties (e.g., data subjects, regulators, auditors, etc.) that accountability practices are available. It is then when metrics for accountability can play an important role. Metrics can serve as a tool to measure that the type of activities that the cloud providers perform are appropriate and effective for a specified context.

Conceptually, the notion of accountability can be decomposed into several properties. Such properties, referred as *attributes of accountability* [2], include transparency, verifiability, observability, liability, responsibility, remediability and attributability. Thus, it would be logical to think that if we are interested in assessing how accountable an organization is we should be able to provide techniques to measure the attributes that influence on accountability. How much or to what extent they should be measured is a key issue.

This is not specific to the concept of accountability. Metrics have a central role in cloud computing, as reflected by the NIST definition of the cloud [3], in which five essential characteristics are identified: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. This last characteristic, embodied by the use of metrics, is of key importance in cloud environments for several reasons. Metrics can be used by cloud consumers to make informed decisions about cloud providers, by helping them to select the appropriate providers depending on their results. Cloud consumers can also use metrics in order to monitor the quality of the services that the providers deliver and check whether the terms agreed on the SLA are met. Metrics can be useful as well to assess cloud governance as they can give some indications to external stakeholders on the suitability and effectiveness of implemented practices. At the same time, the perception of transparency of cloud providers increases as they offer means to measure internal processes.

Hence, metrics for accountability can be considered as a means to show that proper mechanisms for privacy, security and information governance are in place and indeed support accountability. To the best of our knowledge no metrics have been defined for the purpose of accountability in cloud computing systems, let alone, a methodology to elicit them. This paper presents such a methodology, which consists of three different stages. In the first stage we perform a conceptual analysis of the accountability attributes that allows us to derive decomposition of them. Next, we analyze relevant control frameworks designed to guide the implementation of security and privacy practices and mechanisms, and check for their appropriateness to accountability. From these controls we can identify measurable aspects that lead to the definition of accountability metrics.

Once a collection of metrics is extracted, it is important to find a suitable validation method. To the best of our knowledge, there are no standard methods for such validation. In this paper, we also discuss some validation strategies, select the one that we found better for our case and describe how the validation process of the accountability metrics took place.

The structure of the paper is as follows. Section 2 gives an overview on existing related work. Section 3 describes the role of metrics in relation to the concept of Accountability. Section 4 describes the methodology that we have followed to elicit accountability metrics. Section 5 proposes a method for expressing confidence in the measure results whereas Section 6 discusses strategies that we believe are more appropriate for the validation of the accountability metrics and how we performed it. Finally, Section 7 concludes the paper and outlines the future work.

2. Related Work

In the field of information security, the work related to metrics for security is extensive, such as for example, the CIS security metrics catalogue [4]. In the specific context of the cloud, Luna *et al.* present in [5] an approach for quantitative reasoning about cloud security SLAs. The authors do not describe a methodology to elicit the security metrics, but a proposal for how to aggregate and reason about them. With regard to standards, the ISO/IEC 27004 standard [6], which belongs to the ISO/IEC 27000 family on information security, provides guidance on the development and use of metrics for Information Security Management Systems (ISMS), whereas the NIST SP 800-55 publication [7] gives some recommendations for the design of metrics for ISMS, as well as some examples of security metrics. Both standards are exclusively focused on information security, which is only a facet of our definition of accountability.

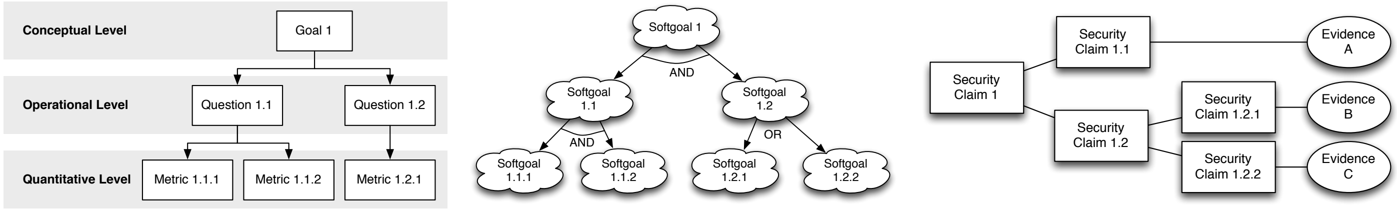
As mentioned in the introduction, the definition of accountability varies depending on the application context. The scope of the notion of accountability we consider in this work is framed within privacy and data governance in cloud services. If we focus exclusively on the privacy part, there is extensive work in the field of metrics for privacy, in particular for anonymity networks, anonymity in databases, and unlinkability for individuals in a communication network. Examples of such metrics are the widely known *k-anonimity* measure for quantifying anonymity of individual records in a database [8], the *size of the anonymity set* for counting the number of set members, which an adversary could be potentially looking for [9], and the *degree of anonymity*, an entropy-based measure that expresses the likelihood that a specific user is the sender of a message in the network [10]. However, security and privacy are only secondary dimensions of our notion of accountability. There are additional core concepts related to accountability, such as responsibility, transparency or remediability, which are seldom tackled. These additional concepts are not only of technical nature, but also operational and organizational, which makes their measurement even more challenging.

Part of the methodology for eliciting accountability metrics that we propose in this paper is in part reminiscent to other similar top-down methodologies for assessing and reasoning about non-functional properties, such as the GQM paradigm, the NFR framework and the security assurance cases.

The Goal Question Metric (GQM) paradigm [12, 13] is a structured approach for the definition and evaluation of the goals of a system, mainly used in the field of software engineering. The GQM paradigm, as depicted in Figure 1a, is based on a top-down decomposition of the needs of the target system, first into goals that suit these needs, next into operational questions associated to these goals, and finally into metrics for answering these questions. Therefore, the GQM can be seen as a top-down approach for the identification of measurements, with three different levels: (i) a conceptual level that corresponds to the identification of goals and that deals with high-level concepts such as business objectives and needs; (ii) an operational level, where each goal is refined into several questions that reflect the operations taken within the system for reaching the goals; and (iii) a quantitative level, where different metrics are assigned to each question. Therefore, GQM proposes a stratified approach, based on the level of abstraction of the treated concepts: high-level concepts (i.e., goals) are refined (i.e., through questions) until reaching quantifiable notions (i.e., metrics).

The Non-Functional Requirements framework (NFR) [14, 15] is a goal-modelling technique that permits the description of softgoals, that is, goals that represent non-functional requirements and that do not have clear-cut satisfaction criteria. According to [15], a softgoal is said to be satisfied when there is sufficient positive evidence and little negative evidence against it. On the contrary, a softgoal is unsatisfiable when there is sufficient negative and little positive evidence. An example of NFR diagram is shown in Figure 1b. This framework enables the recursive decomposition of softgoals in a top-down manner, which enhances the expressiveness and level of refinement of the models. However, it is difficult to say that a softgoal is "satisfied" in a clear-cut sense, depending on the satisfaction of its sub-goals. Furthermore, full satisfaction of softgoals may be impossible because of conflicts and trade-offs between them. The intention behind the NFR framework is to help during the process of finding a set of leaf softgoals that maximizes their positive influence over top softgoals while minimizing their negative influence.

A security assurance case is a method for structuring a set of arguments or claims, supported by a corresponding body of evidence [11, 16]. Security assurance cases (or simply, secu-



(a) GQM Paradigm

(b) NFR Framework

(c) Assurance Cases (adapted from [11])

Figure 1: Top-down methodologies

rity cases) are used to demonstrate, through the provision of evidence, that a system meets certain security properties, represented by security claims. Figure 1c shows a high-level view of a security case that has a top-level claim called Security Claim 1. As stated in [11], the validity of this claim is demonstrated by describing arguments that decompose the top-level claim into subclaims, repeating this process recursively, so that each subclaim is supported by further arguments, until, ultimately, the top-level claim is associated to a body of evidence. How evidence is chosen and organized is key for constructing a security assurance case, as well as defining well-structured arguments that show how this evidence supports the given claims. It is clear that the nature of the claims will constrain the kinds of evidence that will be used, as some pieces of evidence will be sounder than others.

As discussed in Section 4.1, a top-down approach is a natural way for decomposing the problem of metrics elicitation into smaller parts, but used alone does not guarantee the definition of metrics. In this paper, we show how a top-down approach can be supplemented in order to facilitate the extraction of meaningful metrics.

In this work, we also take the confidence in the result of a metric into consideration, and present and approach for reasoning about it. This topic has been already dealt with in the literature in different forms. For example, Ouedraogo *et al.* present in [17, 18] a taxonomy of quality metrics, with the intention of expressing the assurance in the security verification process. In this context, 'quality' is referred to an assessment on the confidence of the constituent aspects of the verification process, namely coverage, rigor, depth and independence. A set of levels for each of these aspects is defined, as well as the criteria of assignment. This work is relevant for us since we follow a similar approach.

There are other relevant frameworks to describe the level of assurance in security evaluation. For instance, the Common Criteria standard (ISO/IEC 15408) [19] defines the notion of *Evaluation Assurance Level* (EAL), an ordinal rating that indicates the thoroughness of the specification, development and evaluation processes of a computer security product; another example is in the family of ETSI standards on trust services for electronic signature infrastructures, which defines different criteria for the evaluation of trust service providers [20]. In this paper, however, the notion of confidence is explicitly focused on the assurance of the metrics evaluation process, rather than on general evaluation procedures.

Finally, we note that in the Metrics Metamodel proposed as previous work [21] (briefly described in Section 4), the quality of the associated evidence is considered as a prospective factor that could influence the metric. Although the concept of confidence is not explicitly mentioned, it is stated that the evidence for metrics may come from sources with different levels of certainty and validity, depending on the method of collection or generation of such evidence. That is, the notion of confidence associated to the source material for applying the metrics (i.e., the evidence) is considered implicitly. However, no further proposal is made to this respect. In this work, we explicitly develop the notion of confidence and describe the criteria for expressing different levels. In addition, the Metrics Metamodel is limited to the conceptual decomposition of the accountability concept, while this work also integrates the analysis of control frameworks into a complete methodology for metrics elicitation.

3. Metrics and Accountability

As mentioned earlier, metrics can serve as a tool for verifying the compliance of high-level requirements, such as security and privacy. Therefore, it is logical to consider the definition of metrics for evaluating how accountable an organization is. Furthermore, the notion of metrics itself is a core element in accountability, as it represents a suitable tool for demonstrating that proper mechanisms are in place, and that, indeed, they support accountability. All these characteristics of metrics can be seen as different facets of their relation to accountability. In this section, we study this relation, first by describing an abstract modelization of the concept of accountability, and next, by explaining how metrics fit in this model.

3.1. A Model of Accountability

Before delving into how metrics are contextualized within the conceptual domain of accountability, it is necessary to consider how the concept of accountability can be modeled. To this end, we consider the three-layer model of accountability from the A4Cloud project [2], which distinguishes between accountability attributes, accountability practices, and accountability mechanisms. In addition, we also consider the notion of accountability evidence, which is of prime importance for properly contextualizing the accountability metrics. These concepts are explained in detail below.

3.1.1. Accountability Attributes

Accountability is conceptually decomposed into attributes, which capture concepts that are strongly related to and support the principle of accountability [2]. The accountability attributes are: observability, verifiability, attributability, transparency, responsibility, liability and remediability.

- *Observability* is a property of an object, process or system which describes how well the internal actions of the system can be described by observing the external outputs of the system.
- *Verifiability* is a property of an object, process or system whose behavior can be verified against a set of requirements.
- *Attributability* is a property of an observation that discloses or can be assigned to actions of a particular actor.
- *Transparency* is a property of a system that it is capable of 'giving account' of, or providing visibility of, how it conforms to its governing rules and commitments.
- *Responsibility* is the state of being assigned to take action to ensure conformity to a particular set of policies.
- *Liability* is the state of being legally obligated or responsible.
- *Remediability* is the state of being able to correct faults or deficiencies in the implementation of a particular set of policies and rules and/or providing a remedy to a party, if any, harmed by the deficiency.

There exist also relationships among attributes (e.g., implication and inclusion), depending on how they can be interpreted (e.g., technical, legal and ethical points of view).

3.1.2. Accountability Practices and Mechanisms

Accountability practices are those behaviors that should be inherent to accountable organizations. In particular, these are: (i) definition of internal rules and policies for complying with pertinent criteria, (ii) implementation of appropriate actions to update governance, (iii) demonstration of compliance with regulations and internal policies, and (iv) remediation and redress in case of any failure.

Accountability mechanisms are those processes and tools that support and implement accountability practices, and that range from risk assessment and auditing to software and hardware systems (e.g., log systems).

Although the model of accountability presented in [2] has a structure of three layers (namely, attributes, practices, and mechanisms), for our purposes we do not need to distinguish between practices and mechanisms, since both are, ultimately, means for implementing accountability, with accountability practices being at an organizational (or behavioral) level, and accountability mechanisms at a technical level (e.g., concrete tools and technologies). We are interested in metrics that assess accountability, both from the organizational and technical perspectives.

3.1.3. Accountability Evidence

In addition to the concepts presented earlier, the notion of *evidence* is necessary for properly contextualizing metrics within the accountability model. The concept of 'Evidence' from the point of view of Accountability is defined by the A4Cloud project [22] in the following way: *'Accountability evidence can be defined as a collection of data, metadata, and routine information and formal operations performed on data and metadata, which provide attributable and verifiable account of the fulfillment of relevant obligations with respect to the service and that can be used to convince a third party of the veracious (or not) functioning of an observable system.'*

This definition is broad enough to permit the consideration of different types of evidence sources, ranging from observations of technical systems (e.g., network logs) to organizational documentation (e.g., internal policies of an organization). This is reflected in the consideration as evidence, not only of data (and metadata), which is usually associated to technical characteristics and observations of a system or process, but also of 'routine information', which comprises information regarding the internal processes of organizations. From now on, and within the context of metrics for accountability, we refer to these elements as 'Evidence'.

As pointed out in [21], the concept of Evidence is central for the process of eliciting metrics. Any assessment or evaluation of a property or attribute can only be made using as input some tangible information. The term 'Evidence' was used in this context to refer to the information used to support the assessment within a metric. Hence, a metric does not directly measure a property of a process, a behavior, or a system, but the evidence associated to them. It can be seen that the notion of evidence in this context is very broad and it is not limited to computerized data (such as a system log), but can be applied to more general information (e.g., the description of a process within an organization, a certification asserted by an external party, etc.), as long as it presents measurable characteristics.

3.2. Contextualizing Metrics within Accountability

Taking into consideration the definition of the accountability attributes, accountability practices and mechanisms, and accountability evidence given earlier, we can informally describe the relationship between these concepts and the notion of metrics, depicted in Figure 2. The concept of evidence conveys all the information supporting the evaluation performed by a metric. As discussed earlier, metrics do not directly measure or evaluate an accountability attribute, but the evidence associated to it, which in turn is the consequence of a practice or mechanism that is present as a support of the accountability attributes.

The process of implementing accountability practices and mechanisms should entail the assessment of the accountability of an organization in a systematic way, acting as a feedback loop. The definition and use of specialized metrics are means for achieving this assessment. The ultimate goal of the use of metrics is to achieve a 'virtuous cycle', where metrics aid to identify deficiencies and inefficiencies in the application of the accountability practices and mechanisms, which in turn, should

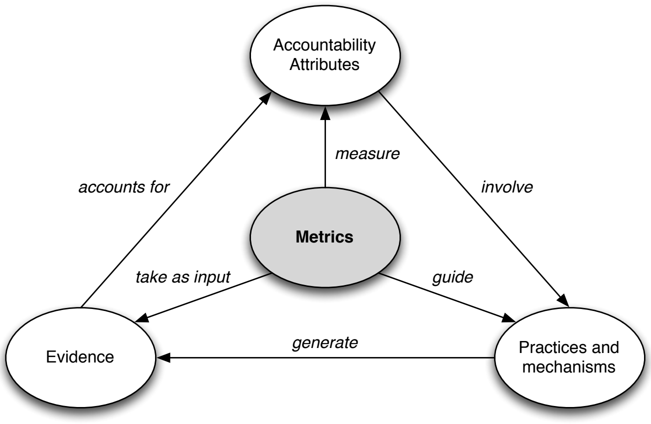


Figure 2: The role of metrics in accountability

provide more and better information, in form of accountability evidence, that can be used during the measurement process. Therefore, metrics can be used by cloud providers as a means to bootstrap accountability, by requiring to implement and improve the processes and behaviors that characterize an accountable organization, described in Section 3.1.2 (i.e., definition of governance, ensuring proper implementation, demonstration of compliance, and remediation and redress). Therefore, the more mature an organization is from an accountability perspective, the greater role is held by accountability metrics.

In addition, accountability metrics can aid to demonstrate whether appropriate practices are in place, fostering this way trust in the cloud ecosystem. This can be done by providing the metrics results as an attestation, either qualitative or quantitative, of the application of these practices by cloud providers. This way, progress in the implementation of accountability mechanisms and practices can be justified continuously and consistently. Therefore, although policy constitutes the actual basis for compliance, metrics are a means for demonstrating its implementation.

4. A Methodology for Eliciting Accountability Metrics

In this section we describe the methodology that we have followed for eliciting metrics for accountability. In order to measure the accountability attributes, we need to have a clear target of the aspects of the attributes that are to be measured. The definitions of the attributes are in some cases vague, subjective or ambiguous, which makes difficult to measure specific aspects. We need a suitable model that allows us to identify measurable factors from the definitions of the attributes. Once these specific factors are identified, we derive metrics for them, relying in addition on the analysis of existing control frameworks. We propose a methodology for eliciting accountability metrics that consists of three main stages:

1. Conceptual analysis. The initial stage of the methodology is devoted to the modelization and decomposition of complex properties, such as the attributes of accountability. For this stage, we use the Metamodel for Accountability Metrics proposed in [21], which enables a top-down and recursive decomposition of these attributes, and the identification of practices and mechanisms that support accountability.
2. Analysis of control frameworks: The previous stage is complemented by an analysis of relevant control frameworks, which are structured collections of controls specifically designed for guiding and assessing the implementation of practices and mechanisms that support security, privacy and information governance. The aim of this stage is facilitating the systematic identification of assessable factors.
3. Definition of metrics. Finally, quantifiable elements are identified from these controls. Thus, metrics can be defined based on them.

Below, we describe the stages of the methodology in more detail. As an illustration of the followed approach, we also provide an example of application of each stage of the methodology, with the objective of explaining the process to obtain a metric, whose result is shown in Table 3. In addition, Figure 3 depicts the different stages of the methodology, along with the provided example.

4.1. Stage 1: Conceptual analysis

The accountability attributes, described in the previous section, belong to the family of *non-functional properties*, which include those properties that are not directly related to functionality, but to a quality or behavioral attribute of a system [23]. Evaluating this kind of properties is widely regarded as a complicated problem because of their nature. Non-functional properties tend to be defined in subjective and ambiguous terms, and to present multi-dimensional aspects. As a consequence, it is often very difficult to assess if this kind of properties have been met, since there is no clear-cut criteria for deciding it. A similar problem occurs with non-functional requirements in the area of requirements engineering [15].

It is clear then that the non-functional nature of accountability attributes is an important hindrance for defining meaningful metrics. As stated earlier, most of the problems we face are related to the level of abstraction of the attributes of accountability. Some of them are defined in a very high-level of abstraction, which is prone to vagueness and ambiguity, and are then not useful from a metrics perspective. Furthermore, there is a disparity in the level of abstraction between different attributes. Thus, a tentative solution is to consider a stratified view of the attributes, where high-level attributes represent more vague and wide concepts and low-level attributes represent more tangible and empirical notions. This would also allow a fine-grained decomposition of attributes, if needed.

To this end, we proposed in [21] a metamodel to describe accountability attributes. The goal of this metamodel is to break down accountability into simpler and lower level concepts, constructing a tree-like model until reaching more concrete elements, such as specific accountability practices and mechanisms. One of the main contributions of the metamodel is that it allows also to identify the evidence that is associated to these

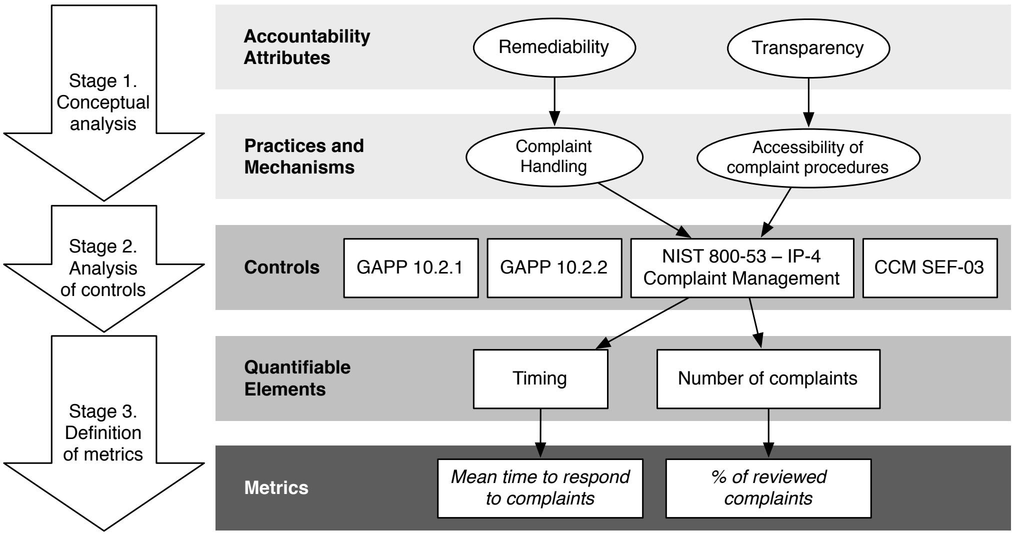


Figure 3: Methodology for Eliciting Accountability Metrics

practices and mechanisms. As explained in Section 3, evidence is a core concept within the contextual model of metrics for accountability, since metrics use evidence to measure accountability. That is, a metric does not directly measure the attributes of accountability but uses the evidence produced by the practices and mechanisms in place, in order to derive a meaningful measure for them.

Although this top-down approach may seem like a natural strategy for reasoning about high-level concepts, such as accountability, it does not guarantee to reach measurable concepts. In fact, when facing the elicitation of metrics using exclusively the metrics metamodel, we found difficult to derive metrics starting from the model of the accountability attributes. Actually, the value of the proposed metrics metamodel lies principally in aiding to correctly identify and specify the supporting practices and mechanisms that are relevant or influence the accountability attributes, rather than being a method for extracting relevant metrics. Alternatively, other modeling methodologies could be applied in this stage, such as the NFR framework, but will face the same difficulties eventually. The next stage can be considered precisely as a complementary strategy to this one.

Example. Let us consider the case of Remediability and Transparency attributes. Remediability is devoted to the establishment of policies and procedures for providing remedy to a party after a failure, and is supported by three main practices, namely, notification, reparation and redress. The redress practice can be further decomposed into more concrete mechanisms, such as procedures for complaint handling, as shown in Figure 3. We are also interested for this example in the accessibility to complaint procedures, which undoubtedly support the notion of Transparency. It is clear that other practices and mechanisms are related to Transparency and Remediability, but we will not consider them in this example, for the sake of illustration.

4.2. Stage 2: Analysis of Control Frameworks

The goal of this phase is to supplement the conceptual analysis produced before, resorting to information located closer to the sources of evidence, in order to facilitate the elicitation of accountability metrics in a systematic way. Taking into consideration that the main objective of the metrics for accountability is to demonstrate that proper mechanisms for privacy, security and information governance are in place, it is necessary to identify quantifiable aspects of the evidence that are directly associated to the accountability practices and mechanisms.

Control frameworks constitute a good source for these aspects. In our context, a control framework is a structured collection of guides and rules specifically designed to assist in and assess the implementation of practices and mechanisms that support security, privacy and information governance. These not only include practices and mechanisms of technical nature, but also organizational. Control frameworks are widely used in organizations during audits and certifications. Evidence of their application can be reasonably extracted from audit records or similar data. Therefore, it is fair to assume that they can be used as sources of assessable factors from where metrics can be derived. This approach could also be extended to consider other sources apart from control frameworks, such as regulations. However, the structured nature of control frameworks facilitates this process.

In our context, we are interested in control frameworks that are relevant for accountability of cloud services, such as the Cloud Control Matrix [24], the Generally Accepted Privacy Principles [25], and NIST SP 500-83 [26]. These control frameworks are specifically focused on the categories of mechanisms that support security, privacy and information governance.

The Cloud Controls Matrix (CCM) by Cloud Security Alliance [24] is a control framework specifically designed for the purpose of cloud security. Its goal is to serve as a guide of security principles for cloud vendors, as well as to aid cloud customers to evaluate the security practices that are implemented

Table 1: Coverage of Control Frameworks to Accountability Attributes

Control Framework	Number of Relevant Controls/Total Number of Controls	Coverage of Controls to Accountability Attributes						
		Observability	Verifiability	Attributability	Transparency	Responsibility	Liability	Remediability
Cloud Controls Matrix (CCM) v3.0.1	61 / 133	6	24	24	46	44	16	36
Generally Accepted Privacy Principles (GAPP)	42 / 73	2	5	3	31	3	0	7
NIST 800-53 Rev 4 - Privacy Control Catalog	21 / 26	1	7	0	11	6	1	3

by cloud providers. The CCM is a fundamental constituent of the Open Certification Framework of CSA [27], whose ultimate goal is to harmonize security expectations from customers, the measures implemented by providers and the evaluation of such measures. In its latest version, CCM v3.0.1 provides a control framework comprised of 133 controls, divided into 16 domains. These controls are in turn related to other widely-accepted security standards and regulations, as well as to other relevant controls frameworks.

The AICPA/CICA's Generally Accepted Privacy Principles (GAPP) [25] is a set of privacy-related principles for guiding the definition and management of privacy programs. Each of these principles has associated a set of criteria of accomplishment, summing up to a total of 73 GAPP criteria. The GAPP control framework is categorized in 10 thematic areas or 'principles', each of them grouping a number of criteria, which we will consider as controls. It is worth noticing that from a very broad perspective, we can find some similarities among the GAPP principles and the A4Cloud accountability attributes.

The NIST Special Publication 800-53 Revision 4 [26] consists of a wide catalogue of security and privacy controls intended for US federal information systems and organizations. The last revision of this publication included a Privacy Control Catalog (under Appendix J), a comprehensive subset of 26 controls specialized in privacy and data protection.

The rationale for the selection of these particular frameworks is that CCM has a focus on cloud security, while the GAPP and NIST frameworks are specialized in privacy. Other control frameworks, such as ISO/IEC 27001:2013 [28] and COBIT 5 [29], could also be considered, although they are less relevant for the scope of this work, in comparison to the selected frameworks.

Before continuing with the metrics elicitation process, we study the coverage of the selected control frameworks with respect to the notion of accountability. For each control from these frameworks, we analyzed its applicability towards the satisfaction of the accountability attributes. More specifically, for each control we evaluated whether it is relevant for accountability or not, and in case it was relevant, we marked the related accountability attributes. Table 1 shows the result of this suitability analysis. Although none of the control frameworks is entirely relevant from the accountability point of view, all of them cover a broad portion of accountability attributes. In particular, the GAPP controls and the privacy controls from NIST SP 800-53 are relatively more in line with the accountability attributes, which is a consequence of their privacy-oriented nature. CCM presents lower coverage, but it is compensated due to the fact that it specifically targeted cloud providers and its greater number of controls.

Next, the results of this initial suitability analysis was extended to produce a mapping between controls and the accountability practices and mechanisms identified during the first stage. The objective of this mapping is that, when metrics are derived in the last stage of the process, they are automatically aligned with the accountability attributes. Thus, any quantitative improvement in the measured results will have beneficial effects on the fulfillment of the controls, which in turn will imply a positive impact on the associated accountability attributes.

Example. Let us consider the control IP-4 from NIST 800-53 Rev. 4 [26], shown in Table 2. This control is titled '*Complaint Management*' and dictates that '*the organization implements a process for receiving and responding to complaints, concerns, or questions from individuals about the organizational privacy practices*'. It is relevant to accountability since it is directly supporting Remediability through its complaint handling practice, as mentioned in the example of the previous stage and depicted in Figure 3. Other accountability attributes, such as Transparency, are also influenced, although to a lesser extent. Once we have identified this control as relevant, in the next stage we study the description of the control in order to identify quantifiable elements. We note also that there may exist other controls relevant to the accountability practices and mechanisms under consideration. In this example, controls GAPP 10.2.1 and 10.2.2 are also relevant to complaint handling.

4.3. Stage 3: Definition of metrics

The last stage of the metrics elicitation process involves the actual definition of the accountability metrics. The previous stages aided to relate the accountability attributes to concrete accountability-supporting practices and mechanisms, which in turn are mapped to controls from relevant control frameworks.

Table 2: Example of Control relevant to Accountability (extracted from NIST 800-53 Rev. 4 [26])

Control ID	IP-4: Complaint Management
Control Description	The organization implements a process for receiving and responding to complaints, concerns, or questions from individuals about the organizational privacy practices.
Supplemental Guidance	Complaints, concerns, and questions from individuals can serve as a valuable source of external input that ultimately improves operational models, uses of technology, data collection practices, and privacy and security safeguards. Organizations provide complaint mechanisms that are readily accessible by the public, include all information necessary for successfully filing complaints (including contact information for the Senior Agency Official for Privacy (SAOP)/Chief Privacy Officer (CPO) or other official designated to receive complaints), and are easy to use. Organizational complaint management processes include tracking mechanisms to ensure that all complaints received are reviewed and appropriately addressed in a timely manner.

Then, the goal of this stage is to evaluate the existence and quality of the practices and mechanisms that are object of the controls, and that are implemented to ensure accountability.

In order to facilitate this process, we propose to incorporate an intermediate step, in which we inspect the nature of the controls that are relevant, and identify whether there are quantifiable element in the description of the control that are susceptible to be measured. Qualitative elements may be identified too, if they have at least an ordinal nature.

For these elements it is necessary also to identify the sources of evidence that will be used for their quantification. Ultimately, this collection of evidence will be the one associated to each derived metric. As described in Section 3.1.3, examples of evidence are system logs, external observations of a particular system or component, certifications asserted by a trusted party, textual descriptions of internal procedures, intermediate reports from audits, etc.

At this point, the quantitative elements identified can be used as parameters in the definition of metrics. This step is mainly driven by the nature of the identified factors, the evidence associated to them, and the experience of the person defining the metrics. However, a rule of thumb is that it is difficult to justify the definition of very complex metrics that have an intricate formulation. An important goal of a metric is that it should be easily computed and comprehended. For this reason, most metrics will be based on the number of occurrences of some events, the computation of rates or percentages, and the definition of several ordinal levels, in the case of qualitative metrics.

An important issue to take into consideration when defining a metric is the notion of scales of measurement. In the classical theory of measurement [30], the scales of measurement are a classification of measurement methods, according to its mathematical characteristics. The correct identification of the scale of measurement is essential for interpreting and analyzing the results of a metric. The scales of measurement can be classified as *nominal* (when there is no relation between values of the scale), *ordinal* (where there is only an order relation between values), *interval* (when, in addition to an order relation, it is possible to compute meaningful differences between values), and *ratio* (when, in addition to the computation of differences, the zero value is meaningful and not arbitrary). For example, a metric that describes the location of a cloud provider (e.g., by country code) is nominal; a metric for risk assessment defined in terms of levels (such as 'Low/Medium/High') is ordinal, whereas a metric for measuring the mean time for notifying data subjects about an incident is ratio. Ordinal and ratio are the more common types of scales used in metrics, since defining levels and counting occurrences of events, respectively, can be described easily. Nominal scales are not very useful since they only provide a set of unrelated categories, and interval scales are seldom used. Nominal and ordinal metrics are often grouped as qualitative metrics, whereas interval and ratio metrics are considered quantitative.

It is important that elements identified in the previous step are correctly classified as quantitative or qualitative. It is not correct, from the point of view of the theory of measurement, to simply assign a number to an inherently qualitative property and perform operations that are not permitted. For example, a typical mistake is to treat ordinal data as ratio values in the [0, 1] interval and to perform computations with them (e.g., multiplication). Instead, it is preferable to analyze the intended formula and to produce a metric with an ordered scale. See [21] for more details regarding the role of scales of measurement in the definition of metrics.

Once the metric is defined, it is necessary to describe it properly following a template or model that captures the main features of the metric, enabling consistency and repeatability of the results. To this end, there are some recent proposals specific to metrics for cloud services that could be applied such as NIST Cloud Computing Service Metrics Description [31] and ISO/IEC 19086-2 [32]. Given that none of these standards are final at this moment, we describe for the moment a generic template based on ISO/IEC 27004 [6], although it can be adapted in the future to more specific templates. The template is as follows:

Table 3: Example of Metric: Mean time to respond to complaints

Metric ID	A4Cloud Metric 27
Name	Mean time to respond to complaints
Description	This metric indicates the average time that the organization takes to respond to complaints from stakeholders
Accountability Attributes	Remediability, Transparency
Associated Evidence	Records of complaints and resolutions
Input	This metric is computed using the following parameters: • T_i : Response time for the i-th complaint • N: Total number of complaints
Formulation and output	Output = $\frac{1}{N}\sum_{i=1}^{N} T_i$
Associated Controls	NIST SP 800-53 (IP-4), CCM (SEF-03), GAPP (10.2.1, 10.2.2)

- Metric ID. A numeric identifier of the metric.
- Name of the Metric. A distinctive name that summarizes the goal of the metric.
- Description. A brief explanation of the purpose of the metric.
- Accountability Attributes. An enumeration of the Accountability Attributes that are influenced by the results of the metric.
- Associated Evidence. A description of which evidence sources could be used for extracting the information necessary to compute the metric.
- Input. The specification of the input parameters that are used for computing the metric.
- Formulation and output. A description of the method used for computing the metric, as well as the identification of what is the output. In most cases, the formulation would be an arithmetic formula (in the case of quantitative metrics) or a description of levels (in the case of qualitative metrics).
- Associated controls. Identification of relevant references, in particular, to those controls that are associated to the metric. In this case, the reference to the control includes the name of the control framework and the identifier of the control.

Example. In the extended description of the NIST control IP-4 it is stated that '*the organization responds to complaints, concerns, or questions from individuals within an organization-defined time period*'. From this description we conclude that timely response of complaints is important to support Remediability. To this end, measuring the actual time of complaint responses could provide a meaningful and quantitative measure of this sub-aspect of Remediability. Next, we define the metric '*Mean time to respond to complaints*' as the average time that it takes for the organization to respond to complaints from affected stakeholders. Table 3 shows the final metric. As shown in Figure 3, other metrics can also be extracted from the analysis of the IP-4 control, for example, for measuring the percentage of complaints that are actually reviewed. This process is repeated for other related controls, such as the CCM SEF-03, and GAPP 10.2.1 and 10.2.2. Our postulate is that an organization that truthfully strives to minimize the result of these metrics would indeed contribute towards enhancing its Remediability state (and Transparency, to a lesser extent), and therefore, its overall support for accountability.

The application of the methodology to all the attributes of accountability and taking the three identified control frameworks into consideration, produced a catalog of 39 metrics, summarized in Table 4. For space reasons we left out the complete catalog, although the interested reader may refer to [22] for a detailed description.

5. Expressing Confidence in Metrics

The definition of the metrics can be extended to convey not only the assessment done by the metric itself, but also a measure of the *confidence* on this assessment. In this context, the term

Table 4: Accountability Metrics Catalog, showing its relation to the accountability attributes (T – Transparency, V – Verifiability, A – Attributability, O – Observability, RM – Remediability, RS – Responsibility, L – Liability) and relevant controls. For the sake of clarity, the metrics have been organized in thematic categories.

ID	Metric	T	V	A	O	RM	RS	L	Associated Controls
Verifiability and Compliance									
1	Authorized collection of PII		×						NIST (AP-1, IP-1)
2	Privacy Program Budget		×						NIST (AR-1)
3	Privacy Program Updates		×				×		NIST (AR-1), GAPP (1.1.2, 1.2.1)
4	Periodicity of PIAs for Information Systems		×						NIST (AR-2)
5	Number of privacy audits received	×	×						GAPP (8.2.7), CCM (AAC-02)
6	Successful Audits received	×	×		×				GAPP (8.2.7), CCM (AAC-02)
7	Record of Data Collection, Creation, and Update		×						NIST (AP-1, IP-1, DM-2), GAPP (5.2.2)
8	Data classification		×						NIST (SE-1), GAPP (1.2.3), CCM (DSI-01)
9	Coverage of Privacy and Security Training		×						NIST (AR-5), GAPP (1.2.7, 1.2.9, 1.2.10), CCM (BCR-11, CCC-02, HRS-10)
10	Account of Privacy and Security Training		×						NIST (AR-5), GAPP (1.2.7, 1.2.9, 1.2.10), CCM (BCR-11, CCC-02, HRS-10)
11	Level of confidentiality		×						CCM (EKM-01, EKM-04)
12	Key Exposure Level		×						CCM (EKM-01, EKM-04)
13	Data Isolation Testing Level		×						CCM (IVS-09)
Transparency, Responsibility and Attributability									
14	Type of Consent	×							NIST (IP-1), GAPP (3.2.1)
15	Type of notice	×							NIST (TR-1)
16	Procedures for Data Subject Access Requests	×							NIST (IP-2), GAPP (6.2.1, 6.2.4)
17	Number of Data Subject Access Requests	×							NIST (IP-2), GAPP (6.2.1, 6.2.4)
18	Responded data subject access requests	×			×				NIST (IP-2), GAPP (6.2.1, 6.2.4)
19	Mean time for responding Data Subject Access Requests	×							NIST (IP-2), GAPP (6.2.3)
20	Readibility (Flesch Reading Ease Test)	×							NIST (IP-1), GAPP (2.2.3, 3.1.1)
21	Rank of Responsibility for Privacy						×	×	NIST (AR-1), GAPP (1.1.2)
22	Certification of acceptance of responsibility		×				×	×	NIST (AR-5), GAPP (1.1.1), CCM (BCR-11, HRS-11, SEF-03)
23	Frequency of certifications		×				×	×	NIST (AR-5), GAPP (1.1.1), CCM (BCR-11, HRS-11, SEF-03)
24	Log Unalterability		×	×					CCM (IAM-01)
25	Identity Assurance		×	×					GAPP (6.2.2, 8.2.2), CCM (IAM-01, IAM-02, IAM-12)
26	Mean time to revoke users			×					CCM (IAM-11, IAM-02)
Remediability and Incident Response									
27	Mean time to respond to complaints	×				×			NIST (IP-4), GAPP (10.2.1, 10.2.2)
28	Number of complaints	×				×			NIST (IP-4), GAPP (10.2.1, 10.2.2)
29	Reviewed complaints	×				×			NIST (IP-4), GAPP (10.2.1, 10.2.2)
30	Number of privacy incidents	×			×	×			GAPP (1.2.7), CCM (SEF-04, STA-05)
31	Coverage of incident notifications	×			×	×			GAPP (1.2.7), CCM (SEF-04, STA-05)
32	Type of incident notification	×				×	×		GAPP (1.2.7), CCM (SEF-04, STA-05)
33	Privacy incidents caused by third parties	×			×	×			GAPP (7.2.4), CCM (SEF-04, STA-05)
34	Number of BCR plans tested		×			×			CCM (BCR-02)
35	Maximum tolerable period for disruption (MTPD)					×			CCM (BCR-09)
36	Sanctions	×				×		×	CCM (STA-02)
37	Incidents with damages	×				×		×	CCM (STA-02)
38	Total expenses due to compensatory damages	×				×		×	CCM (STA-02)
39	Average expenses due to compensatory damages	×				×		×	CCM (STA-02)

confidence refers to a measure of the assurance of the reliability of the metrics results. In other words, a measure of how reliable the result of a metric is. In this section we describe an approach that can be used to extend the proposed metrics (and other metrics as well) in order to express a measure of the confidence on the assessment done by the metrics.

5.1. A Formal Approach

In this section, we consider how to formally define the idea behind the informal description of *confidence* presented earlier. We formalize the idea of confidence in the metrics by describing it as an orthogonal dimension to the 'raw' measurement result, i.e., the measure itself. Figure 4a illustrates this concept as a bidimensional space, where 'Raw Measure' and 'Confidence' are orthogonal dimensions. In this space, two identical raw measures (denoted by a common value r) could have associated different levels of confidence (c and c', respectively), as shown in the figure. Therefore, the tuples (r, c) and (r, c') represent two different values in the bidimensional space. This kind of tuples, which are the conjunction of a raw measure and a confidence level, will be referred to as *measures* in this new context; that is, we are augmenting the notion of measure. Intuitively, it can be seen that the measure $M' = (r, c')$, which has a higher level of confidence, is preferable to measure $M = (r, c)$.

We could formalize this intuitive preference by inducing a partial order over this bidimensional space. An example of partial order that captures this idea is the 'product order'. This type of partial order defines that, given two pairs (r, c) and (r', c') in the bidimensional space $\mathcal{R} \times C$, then $(r, c) \leq (r', c')$ if and only if $r \leq r'$ and $c \leq c'$ (assuming that r and r' are at least ordinal measures). That is, consider two different measures, M and M', resulting from different applications of a particular metric, then for the result M' to be greater than or equal to result M, it has to have a greater or equal raw measure result and a greater or equal confidence level. Once a partial order has been defined, certain pairs of measures can be compared. For example, let M, M', and M'' be different measures in the bidimensional space, as shown in Figure 4b. According to the product order defined earlier, measure M can only be compared to elements in the gray area, such as element M'. Other elements, such as M'' cannot be compared to M if we establish the product order.

The product order is a conservative but safe choice, since that way we ensure that in order to being capable of asserting that a measure is greater than another, both their raw measure and confidence level should be greater or equal than the other's. In other words, if its raw measure is greater but the confidence level is lower, then one cannot assert anything about the order relationship between the two measures. It is possible, however, to define other kinds of partial orders, different to the product order (and even, if necessary, total orders).

5.2. Factors of Confidence

Once we have formalized how the confidence dimension fits with respect to the measures done by the metric, we have to define how the confidence level is devised. In order to facilitate the process of measuring the confidence in the metrics, a suitable approach is decomposing this concept into relevant factors. Recall that the notion of confidence we are referring to should indicate the quality of the evaluation processes associated to the metric. We follow a similar approach to Ouedraogo *et al.* [17, 18] when distinguishing what are the factors that influence the confidence in the metrics results. We identify two factors of confidence that we describe next.

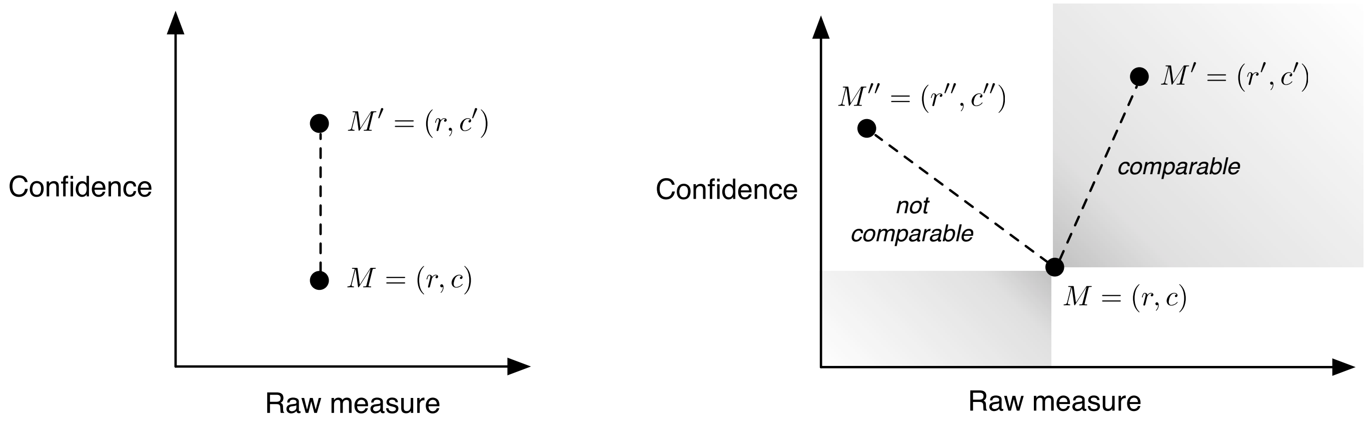
5.2.1. Source of Assessment

This factor is devoted to the identification of the source of the assessment, that is, the actor that performs the evaluation of the metric. Depending on the independence of this actor with respect to the object of evaluation, we can identify several levels of confidence:

- Level 1 (Self-assessment). In this case, the evaluation is performed by the same individuals or organizations that manage the object of evaluation. It is clear that, although one may trust the validity of the assessment and trustfulness of the actors that perform the evaluation, independence is not formally fulfilled. Setting aside the quality of the assessment, the source of assessment in this case implies a lower level of confidence. An example of this kind of source of assessment is CSA Open Certification Framework (OCF) Level 1 [27], which corresponds to self-assessment questionnaires from actual cloud service providers [33].
- Level 2 (Third-party assessment). This level corresponds to an evaluation that is performed by a specialized and trusted third party, such as an auditor or certification body. In this case, the assessment is partially or fully performed by an independent entity. An example of this kind of source of assessment is CSA OCF Level 2, which corresponds to certification or attestation by authorized auditors.
- Level 3 (Consumer/Publicly Verifiable). This level refers to evaluations that can be directly performed by the interested stakeholders. Although in level 2, full independence is achieved from the point of view of the verification process, it is clear that freeing from the need of an intermediary entity is preferable and should be supported when possible, by providing technical and organizational means for the interested stakeholders to perform the evaluations by themselves.

Note that we are only defining an ordinal scale, so trying to reason about the 'distance' between level 2 and level 3 makes no sense. That is, level 2 represents already a high level of independence. The important fact here is that we consider level 3 to be greater than level 2, given the aforementioned reasons. Level 3 is more aligned with the concept of Accountability since it facilitates the demonstrational facet of Accountability, by directly supporting some of its core attributes such as Observability, Verifiability and Transparency.

The first ideas towards this factor are mentioned in [21], where it is identified as an aspect that influence the definition of accountability metrics. Independently to this work, a similar



(a) Raw Measures and Confidence in a bidimensional space

(b) Comparable elements in the Measure-Confidence space

Figure 4: Measure and Confidence Space

factor called 'Independence' is proposed in [17, 18], although our definition differs in that it does not take into consideration partial independence, and that it distinguishes between evaluations performed by a third party and those that are verifiable publicly or by the consumers.

5.2.2. Consistency

This factor describes the level of regularity of the evaluation. For an evaluation to be consistent, a systematic and structured procedure must be followed. However, this is not always true in reality, and certain degrees of relaxation exist during assessments. This factor is identified in [17, 18] as 'Rigour'. It is also reminiscent to the concept of Evaluation Assurance Level (EAL) in the Common Criteria framework [19]. In Common Criteria, these levels reflect the assurance requirements that must be fulfilled in order to achieve Common Criteria certification; a higher level implies better confidence in the security test procedures (among other aspects, such as design and development). The Consistency factor is similar to that concept, although focused exclusively on the evaluation aspects. The following levels of Consistency are proposed:

- Level 1 (Informal procedure). In this level, there is no formal procedure specified for performing the evaluation, or if it exists, no proofs of adherence to the procedures are provided.

- Level 2 (Structured procedure). In this level, a formal procedure is defined, but the means for performing the evaluation are manual and possibly open to interpretation and subjectivity. Proofs of adherence to the procedures are available.

- Level 3 (Automated procedure). The highest level of rigor corresponds to the case when a formal evaluation procedure exists and it is performed in a standardized fashion by means of an automatic mechanism. Proofs of adherence to the evaluation procedures are also available. An example of this kind of rigor level is CSA OCF Level 3, which corresponds to certification based on continuous monitoring.

Note that an evaluation with level 2 of consistency could be as exact as an automatic assessment. However, as noted in [17, 18], the latency of such evaluation would be much lower than an automatic one. Another important difference is the repeatability of the results, since an automatic method is presumably more accurate than a manual procedure. For this reason, the differentiation of level 2 and level 3 is made, since an automated procedure for evaluation would be preferable in a cloud-based setting like ours. As in the rest of cases, it is meaningless trying to figure out the magnitude of the difference between levels, as this is only an ordinal measure.

Although the two factors are theoretically independent, there may exist certain correlation between them. For example, a third-party assessment (Level 2 of Source of Assessment) would probably have a consistency level of 2 or 3, since in most cases the entity that performs the evaluation is a certified and professional organization, which presumably will follow high-quality procedures for the evaluation. Another example is a publicly verifiable assessment (Level 3 of Source of Assessment) performed by an interested stakeholder that, however, does not count with the resources to perform a strict and rigorous evaluation, thus achieving an informal level of consistency (Level 1 of Consistency). These were mere examples, and all the permutations between these two factors are possible. However, these examples reflect different possibilities that affect the global confidence on the metrics results, and justify the identification of the factors of confidence that were presented in this section.

5.3. Establishing the Level of Confidence

Once the factors of Confidence are defined, we can aggregate both factors into a single measure of Confidence. Given that there are two independent factors, we can set up a 'Confidence matrix', very similar in structure to the 'Risk Matrix' typically used in the field of Risk Assessment. The Confidence Matrix is defined in Table 5.

Table 5: Metric Confidence Matrix

		Consistency		
		Informal (Level 1)	Structured (Level 2)	Automated (Level 3)
Source of Assessment	Self-assessment (Level 1)	0	1	1
	Third party assessment (Level 2)	1	2	2
	Consumer/Publicly Verifiable (Level 3)	1	2	3

In this matrix, each combination of confidence factors produces a single Confidence Level that ranges from 0 to 3. These levels are defined as follows:

- Level 0 (Unreliable). There is almost no confidence in the metrics results, since both the independence and the consistency of the assessment are very low.
- Level 1 (Insufficient). In this case, one of the two factors only achieves the lowest level, so the global confidence value will be considered as insufficient. It is clear that confidence in metrics is insufficient when the assessment is self-made or the process is informal.
- Level 2 (Essential). This level is the minimum desired level of confidence. The assessment guarantees an acceptable level of independence and consistency.
- Level 3 (Maximum). This is the preferable level of confidence. However, achieving this level is presumably a costly procedure, since it implies automating the evaluation and making it publicly verifiable.

It is clear that both, self-assessed and informally performed evaluations, are not sufficient for providing a reliable metric, thus, the maximum attainable level of confidence for these two levels is 1. In particular, when the evaluation is both informal and self-assessed, the confidence is considered non-existent (level 0). Once both factors reach a level of 2, then an acceptable level of confidence is achieved (level 2). For the particular case when the evaluation is both publicly verifiable and automated, a maximum level of confidence is reached (level 3). Note that the Confidence level defined above is only a coarse-grained indicator of the aggregation of the two factors of confidence. A finer grained indicator could be possible, but it would have more levels, which complicates its interpretation. Thus, the selection of this scale was done for the sake of simplicity and clarity.

6. Validation

Once the methodology for eliciting metrics for accountability has been defined, it is desirable to test how useful they are and how they can be applied. According to the PMBOK Guide [34], '*Validation is the assurance that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers*'. To this end, there are several approaches that could be followed in order to analyze the suitability of the accountability metrics. In this section we discuss some of these approaches and describe the one we followed in our particular case.

6.1. Validation Strategies

In general, validating metrics is a complex process as it usually deals with abstract concepts. Besides that, there is not a standard methodology designed for validating metrics, but different ways of doing it that are more appropriate than others, depending on the context or user of the metric. There are some validation strategies that are applicable to accountability metrics, based on either simulations or opinions.

6.1.1. Simulation-based Validation

This kind of validation strategy is based on the definition of controlled experiments where simulated data is used for feeding the proposed metrics. The results are then analyzed in order to identify inconsistencies and points for improvement. The disadvantages of this kind of validation strategy is that the simulation of the metrics itself is not an easy task, and in addition, the subsequent analysis does not guarantee to induce a judgement on the validity of the studied object (in this case the metrics). However, it could give a valuable insight on the operation and feasibility of the proposed metrics.

For the case of the metrics derived from the methodology presented in Section 4, we discarded this approach. There are some metrics that, because of their nature, cannot be validated through the use of public or simulated data. For example, most of the metrics that involve purely subjective measures, such as users' satisfaction or users' perception, should be validated with information provided by the users themselves.

Nonetheless, real data that could be used as evidence for some of the elicited metrics is already available in some cases. In particular, the CSA STAR registry [35], which corresponds to CSA OCF level 1 [27], is a public repository where assessment reports of cloud providers are gathered. However, from the point of view of the confidence on the metrics results, these assessments are not ideal, as the information comes directly from the cloud provider, in the form of self-assessed questionnaires. Additionally, only part of the metrics could be simulated this way, since not all the metrics are associated to controls from the Cloud Control Matrix, in particular, to its previous version (v.1.1). Therefore, a simulation with this data would only obtain partial results.

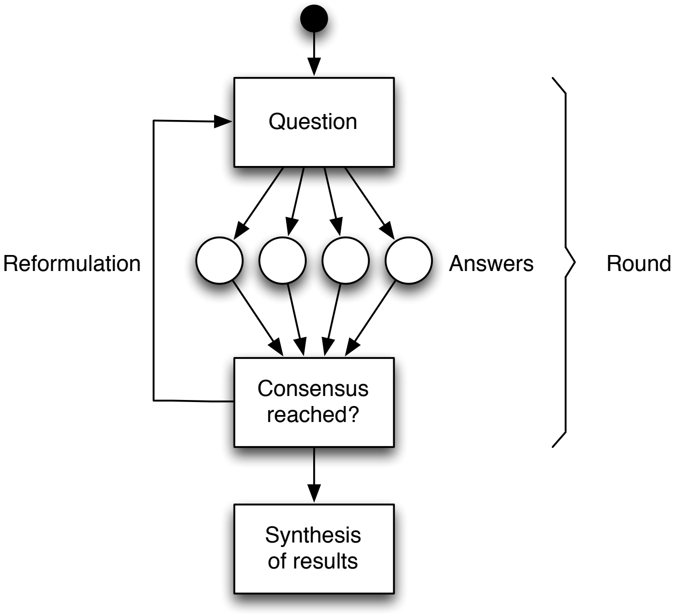


Figure 5: Delphi Methodology

6.1.2. Opinion-based Validation

In the cases where simulations are not the best option, interview-based methods for gathering users' opinions and experience could be very useful for validating metrics. Such methods include expert interviews, group discussions, etc. The goal of these methods is to gather direct feedback and suggestions for improvement from relevant stakeholders. These approaches have the added value of the possibility of counting with a sample of the intended stakeholders of the metrics, such as IT professionals or end-users.

One of the most prominent examples of this kind of methods is the Delphi methodology [36], a structured procedure based on surveys of expert opinions that is usually used in forecasting and decision-making processes. The Delphi methodology requires the participation of a moderator (or a group of moderators), who prepares questionnaires and reviews the responses, and a group of experts, which responds anonymously to the questionnaires. The procedure in the Delphi methodology is iterative. In each round, experts' opinions about a certain subject are surveyed by means of questionnaires. At the end of the round, the moderator reviews the responses, and refines the questions based on the identified consensus and disagreement. The process is repeated several times, until a reasonable consensus is reached or the moderator believes it is enough. Figure 5 shows the general process of the Delphi methodology.

The Delphi methodology supports the refinement of the surveyed questions, but at the same time, it requires that experts participate in several rounds, which can be difficult and time consuming. Ideally, this process should be done in person and in one session, but the methodology is flexible, so it could be performed on-line and in different time periods. The only objective is to iteratively refine the research questions based on the opinions of the experts. In order to elicit the answers to the research question, the use of questionnaires, rather than direct interviews, is preferable in order to avoid the 'interviewer effect', which is any impact (positive or negative) that the interviewer characteristics and behavior induce on the responses of the interviewees.

A secondary means to validating the catalogue of metrics is through expert feedback. This method basically consists of presenting the object of validation to a selection of experts, in this case, belonging to the community of security and privacy metrics, who give feedback based on their expertise. The difference with the previous method is that there is no defined structure at all in the process for asking for feedback.

6.2. Experience from the Validation of Accountability Metrics

In this section, we justify which validation strategy is used for the accountability metrics. The first aspect that we must take into consideration is that the proposed metrics are of diverse nature, ranging from technical to organizational aspects. Hence, it would be difficult to find a 'perfect' approach for validating all the metrics. Ideally, one could choose one strategy or another depending on the nature of the analyzed metric, the associated evidence that is needed for realizing it, and the feasibility of the validation itself. This would imply using a combination of methods, depending on these factors, so the validation would be supposedly easier or more adequate for certain metrics than for others. However, it is impractical to come across a validation strategy for each proposed metric, and therefore a trade-off solution that covers the most of them should be found.

Simulation-based validation is in principle an appealing approach, since it would add an experimental spirit to the validation phase. However, the analysis that has to be performed on the results of the simulation is far from trivial and does not guarantee that the metrics are correctly validated, rather than merely executed. Given these reasons, we conclude that simulation-based validation is not suitable for the validation phase, although it may be a very interesting way to supplement the validation once a suitable simulation environment exists, so resulting data from simulations is adequate, from both the viewpoints of volume and relevance.

On the contrary, the great advantage of opinion-based validation is that, once metrics are defined, it is possible to gather opinions from different stakeholders regarding a wide variety of metrics, regardless its nature and technical complexity, since the opinion will be based on the description of the metric, rather than on an actual implementation. This implies that the a priori feasibility of the implementation of the metric is an aspect to be elicited, together with the rest of validation questions, in order to compensate for the lack of actual implementation. With regard to the risks of opinion-based validation, one of the main issues is the possible lack of variety of the feedback due to the appearance of the interviewer effect, which can produce a bias on the result. To this end, it is preferable to resort to methods that offer a structured and anonymous approach, such as the Delphi methodology. Therefore, our preferred option for validation is based on the Delphi methodology. This decision is based on the following rationale:

- Flexibility: this methodology can be adapted freely to the characteristics of the research question and the characteristics of the participant groups.

- Feasibility: we can ask for opinions for metrics whose simulation would be complicated or impractical.

Additionally, unstructured feedback from experts is also considered, as an auxiliary approach of eliciting feedback. Comments from some experts consulted were examined

With regards to the content of the validation sessions, we prepared a set of questions regarding to the accountability metrics catalogue. Given the size of the catalogue (39 metrics), we strived to keep the questions short. In our approach, the experts evaluate the metrics catalogue through some general questions, but at the same time they were given the liberty of asking or discussing about any particular metric. This way, the size of the questionnaire is kept short, but there is room for discussing specific aspects if needed. The questionnaire contained three questions in the form of statements about the respondents' opinions with a five-point scale: strongly disagree (1), disagree (2), neither agree nor disagree (3), agree (4), strongly agree (5). The questions were the following:

- Q1: '*This set of metrics contains meaningful and relevant measures for Accountability in the Cloud*'. With this question, we wanted to analyze the level of appropriateness of the catalogue for measuring the concept of Accountability in the Cloud.

- Q2: '*The use and application of this set of metrics would be easy, in general*'. The goal of this question is to assess the perceived degree of feasibility of the metrics proposed.

- Q3: '*This set of metrics can be easily understood by a professional audience*'. The goal of this question is to evaluate the degree of usability of the catalogue with respect to the facility of being understood by professionals.

The questionnaire was distributed during August and September 2014 to 18 IT security professionals. We focus on professionals as this type of stakeholders is the one that most likely will apply and benefit from the metrics for accountability, due to the specialization of some of these metrics. Additionally, the questionnaire was distributed through the regular publicity channels of the Cloud Security Alliance, and communicated to security experts in the way of an online survey.

The motivation behind the election of these questions was twofold. Firstly, past experience has shown that it is difficult to gather responses to surveys if there are too many questions. Thus, questions should be concise and kept to the minimum. Secondly, we wanted to evaluate the metrics with respect to the most relevant quality criteria for validation. It is clear that there are several aspects that could be assessed for facing the validation. In [37], Savola identifies three core quality criteria for security metrics, namely correctness, measurability, and meaningfulness; a fourth criterion, usability, is also found to be very relevant. In relation to the quality criteria proposed by this work, we find some parallelism with our questions. Our first question is oriented towards eliciting the opinion regarding the correctness and meaningfulness of the metrics. The second question tries to evaluate the metrics with respect to the measurability dimension, and the third question is focused on the usability aspects of the metrics catalogue.

In the original Delphi methodology, the participants are involved through several rounds. However, given the difficulty of engaging a moderately big group of participants during the whole process, we adapted the methodology for only two rounds. For the first round, we organized a workshop where we could interact with the experts. The results of this round were analyzed and some changes on the catalogue of metrics were made in order to refine the input for the next round. For example, for some qualitative metrics we included more levels that were suggested by the participants as they believe thus the metrics would be more meaningful. This was the case for instance of a metric (14. Type of notice), where new intermediate levels were introduced other than consent is given or not. Thus, one of the intermediate levels introduced considered the behavior of the data subject for granting consent.

The most common concern form the expert group was regarding the acceptance of the proposed metrics by cloud actors, and in particular cloud providers, since these metrics may expose weaknesses of their internal processes and policies. This is true to a certain extent, although, at the same time, this would imply supporting the Observability and Transparency attributes, and hence, the goal of Accountability. As discussed in Section 3.1.3, an objective of the accountability metrics is to demonstrate whether appropriate policies and practices are in place, fostering this way trust in the cloud ecosystem.

The second round of validation was performed individually, in an ad hoc manner. A refined version of the catalogue, together with better explanation of its objectives and motivation, was distributed individually, and responses were gathered one at a time. In parallel to this process, an online survey was available from the beginning of the validation. Thus, its results could not influence the questions of the second validation session since we did not close the survey until the end of the validation period. Hence, from the perspective of the validation, we consider the online survey as part of the first round.

The final results after these rounds show that the three questions were in average answered in ratings considered as 'Agree' or 'Strongly agree'. In particular, question Q1 was rated higher (4.07) than the 'Agree' level, which corresponds to a rating of 4. Question Q2 also increased, although it did not surpass the agree level. Rating of question Q3 was 3.71, almost an 'Agree' level.

7. Conclusions

In this paper we have presented a methodology to elicit metrics for accountability in the cloud. Accountability is a broad concept that needs special attention, given its importance to governance, compliance and trust, and metrics are a key mechanism, not only for demonstrating accountability, but to build accountability. Our proposed methodology consists of three stages. The nature of the accountability attributes that we have to measure is very abstract and it is difficult to identify specific concepts to be measured. Thus, the first stage consists of a conceptual analysis that helps us identifying concrete aspects of the accountability attributes that can be assessed, as well as practices and mechanisms associated to the attributes. The second stage of the proposed methodology consists of an analysis of the existing control frameworks that are related to the concepts identified in the earlier phase. In our case we have used as reference for this analysis the Cloud Controls Matrix (CCM) v3.0.1, the Generally Accepted Privacy Principles (GAPP), and the privacy controls from NIST SP 800-53. From the results of this analysis, measurable aspects (either quantitative or qualitative) are identified in the last stage of our methodology. Metrics derived through this process are automatically aligned with the principles of accountability, since a quantitative improvement in the measured results has a beneficial effect on the fulfillment of the controls, which in turn implies a better implementation of accountability-supporting practices and mechanisms. We have also established a way to express confidence, not only on the metrics itself, but also on the assessment process. The empirical validation that we carried out for the elicited metrics is discussed as well in this paper.

We believe that even though this methodology is initially proposed for eliciting metrics for accountability, it can be used for the elicitation of metrics in other fields. Exploring the ambits where the approach can be applied, or furthermore, whether it can be considered as a general approach for eliciting metrics, remains as future work. For example, ETSI standards on trust services for electronic signature infrastructures [20] describe mechanisms that can be used by certification providers to express their internal policies and processes, in order to build trust in the certification infrastructure, as well as the requirements for assessing them. For these evaluations, methodologies for metrics elicitation like ours can be very useful to derive meaningful metrics that aid during the assessment process.

Finally, we note that metrics derived by this methodology are being considered for standardization by ISO/IEC 19086-2 [32]. For more information about the progress of this work we refer the reader to CSA Working Group on CloudTrust [38]. These metrics are going to be included as well in the definition of an Accountability Maturity Model (AMM) that it is a work in progress.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Program (FP7/2007-2013) for the project A4Cloud under grant agreement n° 317550 and the European Training Network NeCS (H2020-MSCA-ITN-2015- 675320). The second author has been supported by the Spanish Ministry of Economy co-funded by FEDER through the project PRECISE (TIN2014-54427-JIN).

References

[1] Is your cloud provider keeping secrets? Demand data transparency, compliance expertise, and human support from your global cloud providers. Technical report, Forrester Research, 2015.

[2] Daniele Catteddu, Massimo Felici, Giles Hogben, Amy Holcroft, Eleni Kosta, Ronald Leenes, Christopher Millard, Maartje Niezen, David Nuñez, Nick Papanikolaou, et al. Towards a model of accountability for cloud computing services. In *Pre-Proceedings of International Workshop on Trustworthiness, Accountability and Forensics in the Cloud (TAFC)*, 2013.

[3] Peter Mell and Timothy Grance. The NIST definition of cloud computing. Technical Report SP 800-145, 2011.

[4] The Center for Internet Security. The CIS Security Metrics (v1.1.0). `https://benchmarks.cisecurity.org/downloads/metrics/`, 2010.

[5] Jesus Luna Garcia, Robert Langenberg, and Neeraj Suri. Benchmarking cloud security level agreements using quantitative policy trees. In *Proceedings of the 2012 ACM Workshop on Cloud computing security workshop*, pages 103–112. ACM, 2012.

[6] ISO/IEC 27004:2009 – Information Technology – Security techniques – Information Security Management – Measurement, 2009.

[7] Elizabeth Chew, Marianne Swanson, Kevin Stine, Nadya Bartol, Anthony Brown, and Will Robinson. NIST SP 800-55 – Performance measurement guide for information security. Technical report, National Institute of Standards and Technology, 2008.

[8] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[9] Andreas Pfitzmann and Marit Köhntopp. Anonymity, unobservability, and pseudonymity—a proposal for terminology. In *Designing privacy enhancing technologies*, pages 1–9. Springer, 2001.

[10] Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In *Privacy Enhancing Technologies*, pages 54–68. Springer, 2003.

[11] John Goodenough, Howard Lipson, and Chuck Weinstock. Arguing security – creating security assurance cases, 2007.

[12] Victor R Basili. Software development: A paradigm for the future. In *Computer Software and Applications Conference, 1989. COMPSAC 89., Proceedings of the 13th Annual International*, pages 471–485. IEEE, 1989.

[13] Rini Van Solingen, Vic Basili, Gianluigi Caldiera, and H Dieter Rombach. Goal question metric (gqm) approach. *Encyclopedia of Software Engineering*, 2002.

[14] John Mylopoulos, Lawrence Chung, and Eric Yu. From object-oriented to goal-oriented requirements analysis. *Communications of the ACM*, 42(1):31–37, 1999.

[15] John Mylopoulos, Lawrence Chung, Stephen Liao, Huaiqing Wang, and Eric Yu. Exploring alternatives during requirements analysis. *Software, IEEE*, 18(1):92–96, 2001.

[16] Howard Lipson and Charles Weinstock. Evidence of assurance: Laying the foundation for a credible security case, 2007.

[17] Moussa Ouedraogo, Djamel Khadraoui, Haralambos Mouratidis, and Eric Dubois. Appraisal and reporting of security assurance at operational systems level. *Journal of Systems and Software*, 85(1):193–208, 2012.

[18] Moussa Ouedraogo, Reijo M Savola, Haralambos Mouratidis, David Preston, Djamel Khadraoui, and Eric Dubois. Taxonomy of quality metrics for assessing assurance of security correctness. *Software Quality Journal*, 21(1):67–97, 2013.

[19] ISO/IEC 15408:2009 – Information technology — Security techniques — Evaluation criteria for IT security, 2009.

[20] Electronic Signatures and Infrastructures (ESI); The framework for standardization of signatures: overview. Technical Report TR 119, ETSI, 2015.

[21] David Nuñez, Carmen Fernandez-Gago, Siani Pearson, and Massimo Felici. A metamodel for measuring accountability attributes in the cloud. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, volume 1, pages 355–362. IEEE, 2013.

[22] The Cloud Accountability Project (A4Cloud). `http://www.a4cloud.eu/`.

[23] Norbert Siegmund. *Measuring and Predicting Non-Functional Properties*

of Customizable Programs. PhD thesis, Otto-von-Guericke-Universitat, Magdeburg, Germany, 2012.

[24] Cloud Security Alliance. Cloud controls matrix (v3.0.1). `https://cloudsecurityalliance.org/research/ccm/`, 2014.

[25] AICPA-CICA. Generally Accepted Privacy Principles.

[26] G. Locke. NIST SP 800-53 – Recommended Security Controls for Federal Information Systems. Technical Report NIST 800-53v4, National Institute of Standards and Technology, 2013.

[27] Cloud Security Alliance. Open Certification Framework (OCF)). `https://cloudsecurityalliance.org/research/ccm/`, 2014.

[28] ISO/IEC 27001:2013 – Information Technology – Security techniques – Information Security Management systems – Requirements, 2013.

[29] Information Systems Audit and Control Association. *COBIT 5: A Business Framework for the Governance and Management of Enterprise IT*. ISACA, 2012.

[30] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[31] Elizabeth Chew, Marianne Swanson, Kevin Stine, Nadya Bartol, Anthony Brown, and Will Robinson. NIST SP 500-37 – Cloud Computing Service Metrics Description. Technical report, National Institute of Standards and Technology, 2011.

[32] ISO/IEC NP 19086-2 – Information Technology – Cloud computing – Service level agreement (SLA) framework and technology – Part 2: Metrics, Under development.

[33] Cloud Security Alliance. Consensus Assessments Initiative Questionnaire. `https://cloudsecurityalliance.org/research/cai/`.

[34] A guide to the project management body of knowledge (pmbok guide) , 4th edition, 2009.

[35] Cloud Security Alliance. Security, Trust & Assurance Registry (STAR). `https://cloudsecurityalliance.org/star/`.

[36] Norman Dalkey and Olaf Helmer. An experimental application of the delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

[37] R. Savola. Quality of security metrics and measurements. *Computers & Security*, 2013.

[38] CSA CloudTrust Working Group. `https://cloudsecurityalliance.org/group/cloudtrust-protocol/`.