

Obtaining Strong Identifiers Through Attribute Aggregation

Walter Priesnitz Filho, Carlos Ribeiro

Inesc-id, Instituto Superior Técnico, Universidade de Lisboa

Lisboa, Portugal

Email: {walter.filho, carlos.ribeiro}@tecnico.ulisboa.pt

Abstract—The development of services and the demand for resource sharing among users from different organizations with some level of affinity motivate the creation of identity management systems. An identifier can be a single name or a number that uniquely identifies a person, although this is often just a representation of a facet of the person. In a federation, services may require user facets comprised of attributes managed by different identity systems which may then be perceived as two facets of two distinct users and not as belonging to the same user. This problem can be handled by adding a new entity type to the traditional architecture thereby creating links between users from different Identity Providers (IdPs), or by using ontologies in order to establish relations between user attributes from several IdPs. In this paper, we propose a solution consisting of obtaining strong identifiers by combining user attributes within IdPs using direct attribute matching and ontologies. Our application context is the Stork 2.0 Project, an eGovernment Large Scale Project (LSP).

Keywords—Privacy; Identity Management Systems; Attribute Aggregation.

I. INTRODUCTION

The development of services and the demand for resource sharing among users from different organizations with some level of affinity motivate the creation of identity federations. An identity federation features a set of common attributes, information exchange policies and sharing services, allowing for cooperation and transactions between the Federation's members [1].

Although there is no definitive architecture, an identity federation is frequently described as being comprised by: an Identity Provider (IdP), a Relying Party (RP), and a Service Provider (SP) [2]. An IdP is responsible for establishing, maintaining, and securing the digital identity associated with a subject, it may also verify the identity and sign up of that subject. A RP makes transaction decisions based upon receipt, validation, and acceptance of a subject's authenticated credentials and attributes within the Identity System. Relying parties select and trust the identity and attribute providers of their choice, based on risk and functional requirements. Finally, the SP controls the access to the services and resources relying on authorities [3].

An identity is composed by a set of attributes, of which at least one identifies its owner. Although an identifier is often seen as a single name or number that uniquely identifies a person, this is often just a representation of a person's facet,

characterizing the person as authorized to access a service (e.g., employer of, member of). Within a federation, this kind of identities are not relevant for authorization, given that different services require different user facets. Therefore, within a federation each person is characterized by a number of attributes that may be combined to create several facets, which are released whenever necessary to SPs. IdPs manage these attributes, releasing them to SPs according to a security policy, often when required by the authenticated user.

Inside a federation there might be services requiring user facets comprised by attributes managed by different identity systems, which is a problem because often those facets are perceived as belonging to different users rather than the same user. Facets composed by attributes managed by different IdPs may be required for functionality reasons (e.g., checking the curriculum vitae of a person with degrees in several different universities) or it might be required just to increase the strength of the identity.

According to [4], a strong identifier is capable of uniquely identifying a subject in a population by minimizing multiplicity (i.e., the size of the subset of subjects that match that identifier) within a group of subjects, thereby improving the quality of the identification attributes. When considering the overall strength of the identifier, in addition to the multiplicity of the identifier, the Assurance Level of an attribute must also be considered. Assurance Levels (ALs) [5] are the levels of trust associated with a credential and depend of several factors, namely associated technology, processes and policy and practice statements controlling the operational environment.

In some cases, in order to build a strong identifier to satisfy a service's requirement it may be necessary to use a larger set of attributes than the ones present in any IdP. However, incorrect merging of attributes could result in credentials of different persons being attributed to a single user if, for instance, they share the same name and birthday or have other matching attributes.

In this paper, we propose a solution to build strong identifiers by combining the users' attributes within IdPs using direct attribute matching and ontologies in order to find correspondences in users' attributes distributed on IdPs.

This paper is structured as follows: Section II describes some recent proposals on attribute merging. Section III describes open issues on building strong identifiers, while Sec-

tion IV presents particular considerations and possibilities for solving the problem. Finally, Section V considers future research and remaining issues, and Section VI concludes the paper.

II. RELATED WORK

The integration of diverse sources of attributes has been the subject of research by several authors [4], [6]–[9]. Several approaches have been proposed to overcome the challenges discussed above. Some of the proposed solutions include: Aggregation Entities, Aggregation with Focus on Privacy, and Ontologies.

A. Aggregation Entities

Aggregation entities are specifically designed and run to aggregate attributes from several sources.

The Linking Service (LS) is a special kind of aggregation entity proposed in [6] and [7]. The LS acts as an intermediary between the IdP and SP creating links, through user interaction, so that attributes that are present in more than one IdP can be linked and used to identify the user of a particular service. The solution proposed by the authors allows the users to safely establish links between their accounts on several IdPs. The LS connects the different identities and also manages the authentication of different IdPs, so that the user is not required to authenticate separately on each IdP.

The work proposed by [8] enriches the Linking Service concept with some privacy properties identified in Federated Identity Management Systems (FIMS). Firstly, an IdP should not be able to profile the users' actions, therefore, direct links between IdPs and SPs are not allowed and direct interaction between IdPs and SPs is prevented by specific services pseudonyms. Secondly, the disclosure of personal information is controlled by multiple parties, preventing that any single entity from compromising user privacy. SPs cannot obtain the users' personal information from IdPs without prior consent of the users.

B. Aggregation with Focus on Privacy

When working with various sources of integrated data one should take into account the mechanisms in use to which control attributes and data sources should have its access released or denied. This briefly very describes a pertinent issue namely the privacy of the users involved in these processes of data source integration and attribute aggregation.

In [10], authors present lookup tables, dictionaries, and ontologies to map vocabularies and customers. They use aggregated zero knowledge proofs of knowledge (AgZKPK) to allow users to prove ownership of multiple attributes of their identity, without disclosing anything else. The proposal features an authentication service architecture (User-SP-IdP) with Registrars (Rs) entities, which store and manage information regarding reliability/strength of the identifying attributes used in their approach.

Another proposal focused on privacy [11] uses an extension to the Oblivious Commitment Based Envelope (OCBE) protocol. The proposed extension is a version of OCBE protocol for equality predicates (Agg-EQ-OCBE) that analyses multiple functions simultaneously without a significant increase in computational cost. The proposed extension also uses less bandwidth compared to the EQ-OCBE.

C. Ontologies

The use of ontologies allows a higher degree of automation in the process of attribute merging/aggregation. Through its application, it is possible to deal with heterogeneity, which is one of the problems related to aggregating data from different sources.

In [12], authors define four classes of heterogeneity: heterogeneity of the system, that occurs due to technical differences between platforms; syntactic heterogeneity, which is related to representation and formats of data; structural heterogeneity, which results from differences in schemas; and semantic heterogeneity, which refers to differences in meaning generated by different vocabularies and terminologies used.

Ontologies are used in order to share and reuse knowledge [13]. In this context, an ontology is a specification used to create ontological commitments, which are agreements to use a certain vocabulary so that it is consistent with the theory specified in that ontology.

In [14], the authors analysed the requirements of the Pan European e-Services and other features related with integration. The analysis applies basic concepts from a generic model of public service of Governance Enterprise Architecture (GEA) and the Web Service Modelling Ontology (WSMO) to the semantic description of e-Services.

In spite of findings related to defining ways of achieving reliable attribute aggregation processes, to solutions providing privacy, and on ontology mapping, a solution that integrates all these properties has yet to be found.

III. OPEN ISSUES

Approaches used for attribute aggregation are beneficial with regards to obtaining data from various sources, as they are intended to be. However, there are issues that could be improved in these approaches such as the availability of users data aggregators or the use of a single aggregation point for instance.

According to [10] attributes of strong/reliable identification are those capable of uniquely identifying a subject in a population (low multiplicity and high quality), and weak identification attributes are those that may correspond to several subjects (high multiplicity and low quality). Although, two strong identification attributes may separately be able to uniquely identify a subject, their intersection may form a weak identification attribute set, which is not enough to uniquely identify a subject, and therefore is not enough for merging

two identities. Procedures using different IdPs, weak links, etc. could decrease the confidence level of merged identities.

The solutions presented in [10] relate to the treatment of name heterogeneity, mainly with regards to variations in wording, and restrict the language to English. In more heterogeneous environments using the Lookup Tables, as the authors propose, would not be feasible.

The use of ontologies is an interesting resource that we can use to aggregate users' attributes, but when considering the works mentioned above there is one aspect that must be taken into account: in the solutions presented the use of ontologies was only applied to a small number of databases.

The solutions proposed have not yet been tested in a heavy environment. Thus the proposals present no data showing how they perform in a production environment with multiple IdPs and a large number of users.

IV. PROPOSED SOLUTION

A. Application Context

Our application context is the Stork Project, which is one of five eGovernment Large Scale Projects (LSP). The LSPs eCodex, epSOS, PEPPOL and SPOCS carry information regarding justice, health care, procurement and generic business processes, respectively, from one Member State (MS) service to the other. These services communicate with each other through a network of gateways. Stork aims to provide a fundamental building block of any application or service: Authentication. From this perspective, the Stork Project may share four different building blocks with the other LSPs: Authentication, Authorization, Electronic Signatures (long term authentication), and Document Credentials (long term authorization).

B. Solution

As mentioned earlier, previous solutions make no attempt to build strong identifiers by merging identities. Or even try to increase the assurance level of the identification process by joining attributes from several IdPs.

The Stork Project aims to be a basic building block for eGovernment services, providing services such as: Authentication, Authorization, etc. Our proposed mechanism will act in the Citizen "Pan European Proxy Service" (C-PEPS), the Stork gateway. C-PEPS takes on the task verifying citizen credentials and obtaining additional data, e.g., from the represented person and mandates. This role also entails three business processes: Authentication on behalf of, Powers (digital signature), and Business Attributes. Each PEPS includes functionalities specific to its Member State, which are typically the interfaces with the local ID providers, national and business attribute providers.

We propose a mechanism, named User Identification Strengthen (UsIdS), which performs an open search through users IdPs finding correspondences in the users' attributes

(UAs) in order to improve user identification strength. Through an iterative process with the user, he/she will specify which of the IdP(s) that can be used to authenticate him/her in SP does he/she want to use.

A search performed in all pointed IdPs can find matches that are able to certify, with a greater level of assurance, that a user is, in fact, who he/she claims to be. The greater the number of matches found, the greater the strength of the identifier, both because the number of attributes comprising the identifier becomes bigger, but also because some attributes with low assurance levels are repeated by several IdPs. An overview of our mechanism can be observed in Fig. 1.

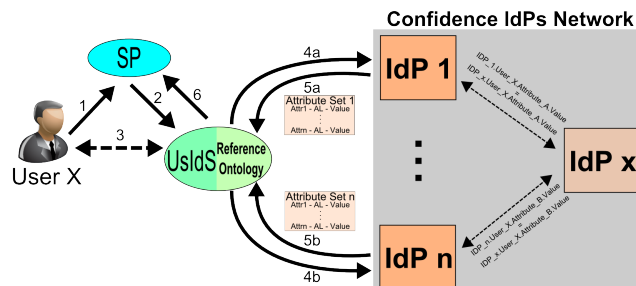


Fig. 1. Proposed search mechanism to find correspondences in user identification attributes

The mechanism assumes, as start point, that user provides a list of IdPs where UAs can be found. As can be observed in Fig. 1, the user sends a service request (Fig. 1 - step 1), indicating the UsIdS as IdP. The SP redirects those instructions to UsIdS (Fig. 1 - step 2). Then, the user sends authentication attributes/authenticates in UsIdS (Fig. 1 - step 3), and attribute set requests are sent to all of the user's IdPs (Fig. 1 - steps 4a, and 4b), the IdPs will then send responses to the UsIdS (Fig. 1 - steps 5a, and 5b) with user attributes sets, each set containing at least the attribute name, Assurance Level (AL), and value. The purpose is to find direct attribute matches, intersections, in attribute sets that can confirm and strengthen the user's identity. The answers received can be handled in two ways depending on the result of the UsIdS analysis of the IdP response.

If an attribute name match is found, the next step is to verify if the attribute values correspond. Otherwise, when attribute names do not match, the next step is to verify on the reference ontology if there is any Ontological Relation (OR) which may established between IdPs involved. If that is the case, attribute values are verified for correspondence. When attribute names do not match, and no ontological relations can be established, UsIdS tries to establish a trusted IdPs network.

To find IdPs, the UsIdS proceeds to search stored ontological relations looking for previously used IdPs. Then a request for user information is sent to that IdPs, and attribute sets are returned in response. UsIdS looks for attribute relations (ARs) between each of the two first IdPs (e.g., IdP₁ and IdP_n) and the new one (e.g., IdP_x). Once an AR is found between i.e., IdP₁ and IdP_x (e.g., IdP₁.Attr_X = IdP_x.Attr_X), the existence

of AR between IdP_n and IdP_x is then is verified. This is repeated until an AR be can found among three, or more, IdPs. When this occurs, the IdP_x attribute set search for presence of any attribute that may be used to improve the strength of the aggregated identity.

When a match is found, the AL of each attribute is verified and the UsIdS sends, as the AL of aggregation, the lowest value within the aggregated value pairs.

In a more schematic way, the process can be seen as follows:

1) Structural Level Verification

- a) With naming conflicts: verifies whether or not similar values, from different user attributes sets, have the same attribute identification.
 - i) Reference ontology-based strategy: ontological relations must be established/verified to solve naming conflicts and help find attribute value correspondences.
- b) Without naming conflicts: when there are correspondences in attributes identification names.

2) Verification Matches

- a) Direct matches: a search is performed in the attribute sets that looks for matches in attribute values. i.e.: $Set_1.Attr_1.Value=Set_2.Attr_1.Value$?
- b) Ontological Relation matches: once a UsIdS finds ontological relations (step 1(a)i) it performs a search through those association sets looking for correspondences in attribute values. i.e.: $Set_1.OR_1.Value=Set_2.OR_2.Value$?
- c) Through trusted IdPs network establishment: it is necessary to obtain user' IdPs in order to create such a network. Then, the process restarts from step 1.

Once UsIdS has performed its searches if correspondences were found an indicator of trust on the User Identity is provided to SP (Fig. 1 - step 6).

As previously described, these matches can be through direct attribute matching or obtained from ontological resources. These ontological resources use an ontology-reference based strategy due to the reduced mapping requirements.

As a start point, the ontologies are used to solve Schema-Level conflicts. According to [14], this kind of conflicts involve differences at the structural level of domain models that need to be mapped. The conflicts can be divided into following categories: naming conflicts, entity identifier conflicts, schema-isomorphism conflicts, generalization conflicts, and aggregation conflicts. We will keep our focus on naming conflicts. This type of conflicts arise when similar concepts are labelled in a different way, or when different concepts are labelled in a similar way.

All established ontological relations are stored in C-PEPS, to improve matching performance in the following searches involving the same users and IdPs, although no private data is kept.

When no matches can be found, the UsIdS tries to establish a trusted IdP network path. The purpose of this network is to find data associations with a third or fourth IdP that can be used to establish a relation among the others IdPs. However, finding the necessary IdPs may be a problem. One possible solution is to search in established ontological relations previously stored. It is also possible to ask the user to indicate where the UsIdS may find more attributes that can lead to matches.

C. Privacy

In order to keep users' privacy, we define a protocol considering the model where partners are "honest but curious", or "semi-honest" [15]. This protocol will be used in communications between IdPs and UsIdS to prevent disclosing of user information in the process of trying to find matches; the entities involved should not gain more information than the one authorized by the user. For instance, if for creating a link between two sets of attributes it is necessary to use another attribute that both IdPs know, this linking attribute should only be revealed, to each attribute source, if the attributes match (i.e., the link is possible), otherwise the attribute source would become aware of user private information for which it was not authorized.

The protocol is defined as follows:

Let p and q be two large prime numbers such that q divides $p-1$, G_q be the unique subgroup of \mathbb{Z}_p^* or order q , and g and h be generators of G_q .

Let x, y , and c be random numbers in \mathbb{Z}_q .

Let id be the identifier/attribute that both IdPs know but don't want to share.

The identifiers id_1 and id_2 are private within the attributes. The protocol must prove that these two identifiers were generated from the same id but it should not be possible to know the exact id value.

Assuming that both the IdP_1 and the IdP_n follow the protocol:

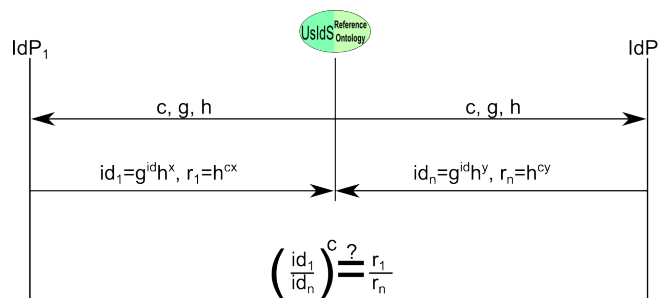


Fig. 2. Proposed privacy preserving protocol

UsIdS sends a challenge c , and generators g and h to both IdP_1 and IdP_n . They reply with a Security Assertion Markup Language (SAML) [16] assertion containing an identifier inside, $id_1 = g^{id_1}h^x$ and $id_n = g^{id_n}h^y$, respectively, but these

identifiers are not equal $id_1 \neq id_n$. They also include in response $r_1 = h^{cx}$ and $r_n = h^{cy}$. Finally, the UsIdS must verify that the following equation holds: $(\frac{id_1}{id_n})^c = \frac{r_1}{r_2}$.

Following this privacy protocol, it is possible to verify if the attribute values correspond without disclosing them.

V. REMAINING PROBLEMS AND FUTURE RESEARCH

There is still, room for further research on how to apply ontologies in UsIdS i.e., an evaluation of how accurate are ontology mappings. A formal definition of collusion resistance must be specified (UsIdS x IdPs, and SPs x IdPs). Some accuracy validations need to be performed on aggregations in order to verify how efficient the UsIdS is. Once there are prototypes it will be possible evaluate and validate the proposed ideas.

VI. CONCLUSIONS

We have proposed a solution to increase the strength of user identifiers by combining facets (i.e., sets of attributes) from several IdPs. The strength of the identifiers results both from an increase in the assurance level of attributes repeated in both sets and an increase of the number of attributes that comprise the combined facet.

Ontologies solve the problem of “Naming Conflicts” that occur when combining sets of attributes. Our chosen Reference Ontology fits our application context (STORK Project), in which there are several languages being used and user data definition on IdPs also has different designations.

Our decision to store ontological relations is due to the fact that the process of establishing these relations could be computationally heavy. So storing the results can improve future searches and can be used to discover IdPs to use in IdPs networks.

The communication process between UsIdS and IdPs uses a privacy protocol in order to assure that user attribute values are not disclosed when IdPs network establishment is being perform. Furthermore, no user attribute values are stored in UsIdS, it just acts as an User Identity Aggregator by relaying IdPs attributes and establishing relations among IdPs and Users.

ACKNOWLEDGMENT

This work was partially supported by CAPES Proc. Num. BEX 9096/13-2 and EU project Stork 2.0 CIP-ICT-PSP-2011-5-297263.

REFERENCES

[1] S. Carmody, M. Erdos, K. Hazelton, W. Hoehn, B. Morgan, T. Scavo, and D. Wasley, “Incommon technical requirements and information,” 2005.

[2] T. W. House. (2009, February) National strategy for trusted identities in cyberspace: Enhancing online choice, efficiency, security, and privacy. URL: http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf [accessed: 2014-09-02].

[3] M. Ates, J. Fayolle, C. Gravier, and J. Lardon, “Complex federation architectures: Stakes, tricks & issues,” in *Proceedings of the 5th International Conference on Soft Computing As Transdisciplinary Science and Technology*, ser. CSTST '08. New York, NY, USA: ACM, 2008, pp. 152–157, ISBN: 978-1-60558-046-3, URL: <http://doi.acm.org/10.1145/1456223.1456258> [accessed: 2014-09-02].

[4] E. Bertino, F. Paci, and N. Shang, “Keynote 2: Digital identity protection - concepts and issues,” in *Availability, Reliability and Security, 2009. ARES '09. International Conference on*, March 2009, pp. lxix–lxxviii, ISBN: 978-1-4244-3572-2, URL: <http://dx.doi.org/10.1109/ARES.2009.176> [accessed: 2014-09-02].

[5] “Identity assurance framework: Assurance levels,” 2010.

[6] D. Chadwick and G. Inman, “Attribute aggregation in federated identity management,” *Computer*, vol. 42, no. 5, pp. 33–40, May 2009, ISSN: 0018-9162, URL: <http://dx.doi.org/10.1109/MC.2009.143> [accessed: 2014-09-02].

[7] D. W. Chadwick, G. Inman, and N. Klingenstein, “A conceptual model for attribute aggregation,” *Future Gener. Comput. Syst.*, vol. 26, no. 7, pp. 1043–1052, Jul. 2010, ISSN: 0167-739X, URL: <http://dx.doi.org/10.1016/j.future.2009.12.004> [accessed: 2014-09-02].

[8] J. Vossaert, J. Lapon, B. Decker, and V. Naessens, “User-centric identity management using trusted modules,” in *Public Key Infrastructures, Services and Applications*, ser. Lecture Notes in Computer Science, J. Camenisch and C. Lambrinouidakis, Eds. Springer Berlin Heidelberg, 2011, vol. 6711, pp. 155–170, ISBN: 978-3-642-22632-8, URL: http://dx.doi.org/10.1007/978-3-642-22633-5_11 [accessed: 2014-09-02].

[9] M. Barisch, E. Garcia, M. Lischka, R. Marques, R. Marx, A. Matos, A. Mendez, and D. Scheuermann, “Security and privacy enablers for future identity management systems,” in *Future Network and Mobile Summit, 2010*, June 2010, pp. 1–10, ISBN: 978-1-905824-18-2.

[10] F. Paci, R. Ferrini, A. Musci, K. Steuer, and E. Bertino, “An interoperable approach to multifactor identity verification,” *Computer*, vol. 42, no. 5, pp. 50–57, May 2009, ISSN: 0018-9162, URL: <http://dx.doi.org/10.1109/MC.2009.142> [accessed: 2014-09-02].

[11] N. Shang, F. Paci, and E. Bertino, “Efficient and privacy-preserving enforcement of attribute-based access control,” in *Proceedings of the 9th Symposium on Identity and Trust on the Internet*, ser. IDTRUST '10. New York, NY, USA: ACM, 2010, pp. 63–68, ISBN: 978-1-60558-895-7, URL: <http://doi.acm.org/10.1145/1750389.1750398> [accessed: 2014-09-02].

[12] V. Kashyap and A. P. Sheth, *Information Brokering Across Heterogeneous Digital Data: A Metadata-based Approach (Advances in Database Systems)*. Springer, 2000, ISBN: 0792378830.

[13] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, Jun. 1993, ISSN: 1042-8143, URL: <http://dx.doi.org/10.1006/knac.1993.1008> [accessed: 2014-09-02].

[14] A. Mocan, F. M. Facca, N. Loutas, V. Peristeras, S. K. Goudos, and K. A. Tarabanis, “Solving semantic interoperability conflicts in cross-border e-government services,” *International Journal on Semantic Web & Information Systems*, vol. 5, no. 1, pp. 1–47, 2009, DOI: 10.4018/jswis.2009010101, URL: <http://dx.doi.org/10.4018/jswis.2009010101> [accessed: 2014-09-02].

[15] Y. Lindell and B. Pinkas, “Secure multiparty computation for privacy-preserving data mining,” *Journal of Privacy and Confidentiality*, vol. 1, no. 1, pp. 59–98, 2009, URL: <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=jpc> [accessed: 2014-09-02].

[16] “Saml specifications,” 2013, URL: <http://saml.xml.org/saml-specifications/> [accessed: 2014-09-08].