Allowing Privacy-Preserving Smart City Open Data Linkage

Francesco Buccafurri
DIIES Dept.
University "Mediterranea" of Reggio Calabria
Reggio Calabria, Italy
Email: bucca@unirc.it

Celeste Romolo
DIIES Dept.
University "Mediterranea" of Reggio Calabria
Reggio Calabria, Italy
Email: celeste.romolo@gmailcom

***Abstract*—Open data are a crucial component of the smart-city ecosystem. In this scenario, many subsystems exist, like transport, shopping, cinema, theatre, utilities consumption, etc. Data coming from the interaction of citizens with these subsystems can be anonymized and published as open data, according to normative requirements and best practices. Therefore, any third-party (even government) entity can perform isolated data analytics, but it is not able to relate open data referring to the same citizen thus missing a lot of potential powerful information. In this position paper, we present a cryptographic approach aimed at allowing cross-correlation of smart-city open data only to authorized parties yet preserving citizens' privacy. The solution leverages the public digital identity system compliant with eIDAS (the European framework) by giving to the Identity Providers the role of Trusted Third Party.**

Keywords–Open Data; Privacy; Linked Data.

I. INTRODUCTION

The concept of *Smart City* is wide and involves in a truly integrated fashion all the components of a community like transport, shopping, cinemas, theaters, utilities consumption, etc. Among the other aspects, the capability of managing and exploiting the information flow underlying the working of the various components of a city is fundamental to make the community really *smart*. According to this paradigm, a very important task is to publish in an interoperable form data coming from the interaction between citizens and the various components of the city. Indeed, any third party can develop applications and perform powerful analysis by exploiting these data. This is basically the principle underlying the concept of *open data*, which both normative enforcement [1] and common best practices require to adopt in a smart community. Evidently, for privacy reasons, data can be published in an anonymous form, hopefully also by satisfying robust privacy protections, like *k-anonymity* [2] or *l-diversity* [3]. However, this as a negative side effect. Indeed, no correlation between data belonging to different subsystems can be done, thus missing a lot of potential powerful knowledge [3].

In this paper, we propose a new strategy, based on a multi-party cryptographic protocol, which allows us to keep anonymity of citizens, but enables data linkage for authorized parties, to recover the knowledge gap and increase the benefit of open data. The solution is practical, because it identifies how to map into real-life entities the different roles of the model, also by considering a public digital identity system compliant with eIdas (the European framework) [4] and by giving to the Identity Providers the role of Trusted Third Party.

The structure of the paper is the following. The next section contextualizes the proposal in the related literature. In Section III, background notions are provided. In Section IV, the problem is formulated and the proposed model is described. The detailed description of the solution is given in Section V. Finally, we draw our conclusions in Section VI.

II. RELATED WORK

A wide scientific literature exists highlighting the importance of open data in the context of Smart Cities. In [5] the role of big data and open data in smart cities is well explained and analyzed. In [6], the correlation between big data, smart cities and city planning is studied. The work [7], discusses how Mobile Application Clusters can be developed through competitions for innovative applications. The Smart City services that are developed in competitions benefit both the Mobile Application Cluster and the citizens. The function of the competition mechanism to encourage the development of new mobile applications utilizing Open Data is described with examples from the Helsinki Region. The authors of [8] sketch the rudiments of what constitutes a smart city, which we define as a city in which ICT is merged with traditional infrastructures, coordinated and integrated using new digital technologies. They highlight how to build models and methods for using urban data across spatial and temporal scales, and to apply them to many subsystems like transport and energy.

A considerable attention has been devoted to the problem of privacy in the context of Smart Cities mainly regarding the protection of information stored and managed by the City entities. In [9], the authors leverage some concepts of previously defined privacy models and define the concept of citizens privacy as a model with five dimensions: identity privacy, query privacy, location privacy, footprint privacy and owner privacy. By means of several examples of smart city services, we define each privacy dimension and show how existing privacy enhancing technologies could be used to preserve citizens privacy. The work [10] deals with problem of data over-collection. This problem arises from the fact that smartphones apps collect users' data more than its original function while within the permission scope. For the authors, this is rapidly becoming one of the most serious potential security hazards in smart city. In the above paper, the authors study the current state of data over-collection and study some most frequent data over-collected cases. The problem of security and privacy is deeply investigated in [11]. One of the main points of this paper is the observation that privacy can be achieved (i) by imposing high security requirements onto the used technology to avoid third party abuses; and (ii) by decoupling technical smart city data streams from the personal one to avoid abuse of data by insiders.

 ISBN: 978-1-61208-707-8

The problem of open data linkage has been considered in a number of papers in the past, especially in the field of health. The paper [12] is an evolution of the W3C SWEO community project, with the purpose of linking Open Data coming from various open datasets available on the Web as RDF, and to develop automated mechanisms to interlink them with RDF statements. In [13], the authors argue that Linked Data technology, created for Web scale information integration, can accommodate XBRL data and make it easier to combine it with open datasets. This can provide the foundations for a global data ecosystem of interlinked and interoperable financial and business information with the potential to leverage XBRL beyond its current regulatory and disclosure role.

Although privacy and linkability have been recognized, as shown above, as important problems in the field of Smart Cities, to the best of our knowledge, there is no paper trying to reach a compromise between the two features, which is, instead the contribution of this paper.

III. Background

Our solution leverages any public digital identity system compliant with the European framework eIdas [4]. Among these, we choose the Italian system SPID [14] to describe a concrete implementation of the general framework. SPID is based on the language Security Assertion Markup Language (SAML) [15], which is an XML-based, open-standard data format designed to exchange authentication and authorization messages between identity and service providers. It uses assertions (signed XML messages) to transfer information in such a way that federated authentication and authorization systems can be implemented. SAML messages included into the HTTP GET request, while longer messages exploits the mechanism of HTTP POST Binding. For security reasons, in SPID, HTTP must be used only on combination with TLS.

The SPID framework includes the following components:

1) **Users**. They are people using the system to authenticate for a service delivered by a Service Provider (see below). Besides an ID and all personal identifying information (such as social security number, name, surname, place of birth, date of birth and gender), other attributes can be associated with the users (like for example a professional status).
2) **Identity Providers**. They identify people in the registration phase (either frontally or remotely), create and manage IDs, and grant the assertion to the Service Providers to authenticate the users at the required level of assurance. The strength of the authentication of the user at the Identity Provider depends on the requested level of assurance. Identity Providers are private or public subjects certified by a Trusted Third Party.
3) **Service Providers**. They are public or private organizations adhering to the SPID system providing a service to authorized users and requiring a given level of assurance.
4) a **Trusted Third Party (TTP)**. It is a government entity (Agency for Digital Italy – AGID), which guarantees the standard levels of security required by the regulation and certifies the involved entities.
5) **Attribute Providers**. They are optional entities whose role is to certify attributes, such as possession of a degree, membership of a professional body, etc.

IV. The Smart-City Open Data Model and Problem Formulation

In this section, we introduce the Smart-City Open-data model which the solution proposed in this paper is applying to. The Smart City is composed of a number of subsystems, denoted as $\{S_1, S_2, \ldots S_n\}$. They are for example transport system, health facilities, schools, universities, shops, cinemas, theaters, utilities consumption, etc. Assume that each subsystem x has a pair $\langle I_x^i, D_x^i \rangle$ where I_x^i is the real identity of an individual i as known to the subsystem x and D_x^i is the set of data that every day (week, month, etc.) the subsystem x collects about that individual. Assume that each subsystem publishes as *open data* the pair $\langle P_x^i, \bar{D}_x^i \rangle$, where P_x^i, is a *pseudonym* of the real identity (as known to x) and $\bar{D}_x^i$ is a suitable transformation of the original data. Let us assume that:

1) $P_x^i = \alpha(I_x^i)$ where α is an anonymization function with the purpose of disguising the actual identity.
2) $\bar{D}_x^i = \delta(D_x^i)$ a transformation of the original data with the double purpose: (i) to hide useless details, and (ii) to make it difficult data de-anonymization (therefore, the function δ takes into account all threats contrasted by state-of-the-art approaches like *k-anonymity* [2] or *l-diversity* [3].

The current situation in real-life systems, and, to the best of our knowledge, in the scientific literature, is the following. The subsystems are independent each other and they use different functions α and δ. Therefore, there is no way to understand that the various data refer to the same individual, so one data are unlinkable. Observe that this, according to this model, this is an expected feature, because it is fundamental to really protect citizens' privacy. Indeed, each subsystem knows the real identity of the individuals, so linking its data with that of other subsystems may result in a potentially very dangerous information leakage. However, the side effect is that it is impossible, for any party, to reconstruct, even in anonymous form, the behavior of a single individual.

The aim of this paper, thus the problem faced by this work, is to recover the above gap of knowledge, without compromising citizens' privacy. The proposed solution is presented in the next section.

V. The Proposed Solution

In this section, we present a possible solution of the problem formulated in the previous section. The solution is both theoretical and practical, because is aware about how to map the entities of the model to concrete parties already playing a role in digital communities. The solution is tailored to a Country belonging the the European Union. Obviously, a more general case could be considered just by identifying different normative and infrastructural components.

We assume subsystems belong to the same Member State. Suppose this State adopts a Public Digital Identity System compliant with eIDAS regulation [4]. For example, in Italy

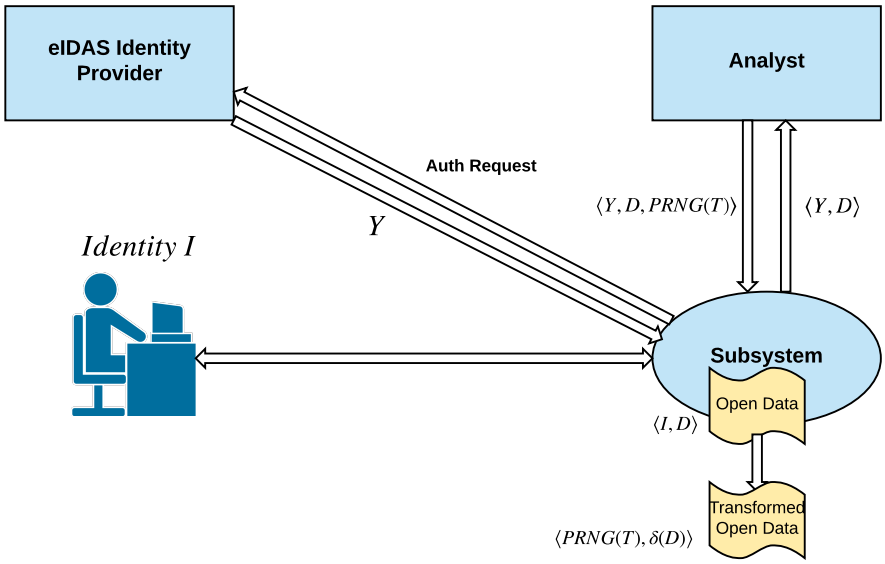

Figure 1. A simplified description of the protocol.

such a system exists and is named SPID (SAML-based authentication system) [15]. So, we assume that the users involved in the system own an eIDAS identity, so each is registered at an Identity Provider (for free). This is a realistic assumption because it is a scenario to which EU aims, according to the eIDAS regulations and others related Acts.

The actors of the systems are the following:

1) Users (i.e., citizens)
2) eIDAS (accredited) identity providers
3) subsystems, playing the role of (special) eIDAS service providers
4) Analysts $\{A_1, \ldots, A_t\}$, a dynamic group of parties empowered to data analysis.

Let U be a user whose identity is managed by the Identity Provider IP.

Let S be the subsystem that U is accessing. This transaction will result in a new (set of) open data. We focus on a single open data $\langle I, D\rangle$. Our proposal aims to modify the function α (see the previous section) in such a way that only for an Analysts Party it is possible to link open data published by different subsystems and referring to the same user. No change is required for the function δ.

The function is modified according to the following mechanism. According to the eIDAS system, the authentication request to a service provider S given by U is forwarded by S to IP. In our case, the request should be modified to enable the open-data mechanism (so, some modifications of the format of SAML messages of the eIDAS system are required). We could also require that the service provider (i.e., the subsystem) that wants to generate such open data must be previously registered and adhere to some common procedural rules.

When IP receives the authentication request, it activates the standard SAML mechanism, but the returned assertion (granting the authentication) will include also (here a modification of the SAML message is required):

1) An order number N, denoting the number of open-data authentications required so far by the user U
2) A value $Y = MAC(IDU, SIP)$, where MAC is a secure message authentication code (like for example HMAC [16]), IDU is the eIDAS identification number (and it can be considered as the identity value i above), and SIP is a secret owned by IP (this is done to avoid that S can invert Y and can find the identification number, which is not public). Moreover, as Y is the output of an hash function, no collision can be found, so Y is uniquely identifying the user.

The subsystem S, once the assertion is received, proceeds as follows:

1) Chooses an Analyst A_x (even at random);
2) Sends the triple $\langle Y, N, Id\rangle$ to A_x, where Id identifies the open data.

At this point, the Analyst A_x, computes:

1) $T = MAC(Y, X)$, where X is a secret shared with all the analysts (this can be obtained by using a dynamic group key agreement protocol – there are a number of efficient extensions of Diffie Hellman to do this [17])
2) T is the seed of a LEcuyer's PRNG [18]. So, A_x, computes $PRNGN(T)$.
3) A_x sends to S the message: $\langle Y, Id, PRNG(T)\rangle$. This, in other words, means that the new function α is defined as $\alpha(IDU) = PRNG(T)$.

At this point, the subsystem S matches the message $\langle Y, Id, PRNGN(T)\rangle$ to the corresponding open-data D and publishes it in the form $\langle PRNG(T), \delta(D)\rangle$.

Observe that when the same user U accesses another subsystem, say G, the protocol will associate the new open data with the pseudonym $PRNG(PRNG(T))$, so that the two open data are both anonymous but linkable. But they are linkable only for those that know the seed T (there is a seed for each user). So, full analytics can be performed only by any Analyst. No other party can do this.

Concerning the adoption of the public digital identity system, as observed earlier, it seems a realistic hypothesis, as the idea underlying this framework in EU member states is to use it as Single-Sign-On system for all the interactions between citizens and the public sector (eventually, with a unique interoperable system over the entire Europe). This is a practical solution because, thanks to the public identity, there is no need of a specific Registration Authority (Identity Providers play this role within their functions).

The proposed solution is summarized in Fig. 1, in which some messages are simplified for the sake of presentation.

VI. Conclusions

Open Data are a fundamental component of the Smart-City ecosystem. They allow transparency, e-participation, but also the development of application able to integrate different subsystems of the community, thus fulfilling the Smart-City paradigm. Typically, for privacy reasons the are anonymized in such a way that they are also unlinkable. However, a lot of potential powerful knowledge may derive from the correlation between data of the same user belonging to different subsystems. In this paper, we proposed a solution based on a multi-party cryptographic protocol also relying on the public digital identity system which appears as good compromise between privacy requirements and information power of data. This is a still work-in-progress paper. Therefore, a number of aspects need to be analyzed in more detail. Among these, the problem of de-anonymization of data, when linkable, assumes a different form than the case of unlinkable data. This is what we plan to do in the near future about this work, together with a *proof-of-concept* implementation of the system.

References

[1] "EU (2003): Directive 2003/4/EC of the European Parlia ment and of the Council of 28 January 2003 on Public Access to Environmental In formation." *In: Official Journal of the European Union, L 41/26,*, 2003.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical report, SRI International, Tech. Rep., 1998.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006, pp. 24–24.

[4] C. Cuijpers and J. Schroers, "eIDAS as guideline for the development of a pan European eID framework in FutureID," *Open Identity Summit 2014*, vol. 237, pp. 23–38, 2014.

[5] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, 2016.

[6] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, 2013.

[7] H. Hielkema and P. Hongisto, "Developing the helsinki smart city: The role of competitions for open data applications," *Journal of the Knowledge Economy*, vol. 4, no. 2, pp. 190–204, 2013.

[8] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, 2012.

[9] A. Martinez-Balleste, P. A. Pérez-Martínez, and A. Solanas, "The pursuit of citizens' privacy: a privacy-aware smart city is possible," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 136–141, 2013.

[10] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.

[11] A. Bartoli, J. Hernández-Serrano, M. Soriano, M. Dohler, A. Kountouris, and D. Barthel, "Security and privacy in your smart city," in *Proceedings of the Barcelona smart cities congress*, vol. 292, 2011.

[12] C. Bizer, T. Heath, D. Ayers, and Y. Raimond, "Interlinking open data on the web," in *Demonstrations track, 4th european semantic web conference, innsbruck, austria*, 2007.

[13] S. O'Riain, E. Curry, and A. Harth, "Xbrl and open data for global financial ecosystems: A linked data approach," *International Journal of Accounting Information Systems*, vol. 13, no. 2, pp. 141–162, 2012.

[14] "SPID-Agenzia per l'Italia Digitale," `http://www.agid.gov.it/sites/default/files/regole_tecniche/spid_regole_tecniche_v0_1.pdf`, 2015.

[15] "Security Assertion Markup Language (SAML)," `http://it.wikipedia.org/wiki/Security_Assertion_Markup_Language`, 2015.

[16] H. Krawczyk, R. Canetti, and M. Bellare, "Hmac: Keyed-hashing for message authentication," 1997.

[17] M. Steiner, G. Tsudik, and M. Waidner, "Key agreement in dynamic peer groups," *IEEE Transactions on Parallel and Distributed Systems*, vol. 11, no. 8, pp. 769–780, 2000.

[18] P. L'Ecuyer, "Uniform random number generation," *Annals of Operations Research*, vol. 53, no. 1, pp. 77–120, 1994.

 ISBN: 978-1-61208-707-8