Location Privacy and Inference in Online Social Networks

Mirko Zichichi[0000−0002−4159−4269]

[1] Law, Science and Technology Joint Doctorate - Rights of Internet of Everything
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[3] Department of Computer Science and Engineering, University of Bologna, Italy
[4] Department of Law, University of Turin, Italy
mirko.zichichi@upm.es

Abstract. Data protection is about protecting information about persons, which is currently flowing without much control –individuals cannot easily exercise the rights granted by the EU General Data Protection Regulation (GDPR). Individuals benefit from "free" services offered by companies in exchange of their data, but these companies keep their users' data in "silos" that impede transparency on their use and possibilities of easy interactions. The introduction of the GDPR warrants control rights to individuals and the free portability of personal data from one entity to another. However it is still beyond the individual's capability to perceive whether their data is managed in compliance with GDPR. To this regard, in this work the proposed approach consists in using decentralized mechanisms to provide transparency through distributed ledgers, data flow governance by using smart contracts and interoperability relying on semantic web technologies.

Keywords: personal data· distributed ledger technologies· smart contracts· semantic web· linked data

1 Introduction

Ubiquitous connectivity with mobile phones and information posted by users provides Online Social Networks (OSNs) and Internet Service Providers (ISPs) with information about the users' location and activities. This information relates to the personal sphere of the individual and composes a part of the dataset called personal data. Personal data is defined as any piece of information that can identify or be identifiable to a natural person. Digital personal data, in particular, is generated by the interaction of a user with a software or a hardware in form of numbers, characters, symbols, images, sounds, electromagnetic waves, bits, etc. (Kitchin 2014). Collecting this data is surely one of the most important reason for what concerns the improvement of safety and security in citizens surveillance. Nevertheless, the main driving force for the collection and use of individuals' personal data is given by the growth of a "not so new" data-driven economy. A huge business, indeed, lies behind the trade of personal data and several companies make consistent profits operating in this sector. The public awareness of this economy is increasing due to recent scandals on the abuse of personal information, such as the 2018 Cambridge Analytica revelations. But still, further work is needed to let ordinary individuals develop the necessarily practical and interpretive skills (Pangrazio and Selwyn 2019). It is still not possible to feasibly ensure to individuals the sovereignty of their personal data, nor the possibility of an appropriate data interoperability for data consumers.

The current practice of data providers, i.e. entities that collect and manage individuals' personal data, is to centralize users resources in "silos". These providers are usually companies that offer services users pay for with their data, mainly OSNs and ISPs. Hence, they have no incentive to freely share data among each other and to other entities, neither to provide individuals transparency of their data usage. In order to keep the privacy of the European Union citizens, personal data is protected by the EU's General Data Protection Regulation (GDPR) [5] . Even if the GDRP has empowered citizens by radically changing operations carried out by data providers, more efforts are needed to reach both transparency and a balance between privacy and data sharing. Another key point of the GDPR is the requirement of data providers to release to their users the complete dataset they collected on them, when requested. Currently, there are no standards for this kind of requests and there is the tendency to hinder the progress of these, causing the entire process to become almost useless (e.g. requiring to submit the request in physical copy through postal courier).

With particular focus on location information, it is possible to understand how managing this information may be extremely sensitive since it enables the possibility of geotracking and a continuous and dynamic knowledge acquisition. OSNs usually handle the user's location data as a private information that is only available to a certain group of people set with regards to the user's privacy settings (e.g. user's friends). However, accessing this data it is possible to perform spatiotemporal analysis using the disclosed Point Of Interests and reveal sensitive information that the user may not want to expose (Hasan, Ukkusuri, and Zhan 2016). Data privacy is a right that must be granted, but even some standard obfuscation methods may not be enough. Knowing (or reconstructing) the friendship graph it is possible to infer users' fine-grained location, even when they keep their location data private but their friends not (Sadilek, Kautz, and Bigham 2012).

Current problems related to the individuals' personal data management can be summarized in three aspects:

[5] Council of European Union, "Regulation (eu) 2016/679 - directive 95/46"

- Personal data is sometimes concentrated in few points and transacted in opaque transfers without the individual's control or even knowledge.
- Data providers store and maintain data differently through several data silos, and to them it is convenient hampering data exchange and its economical exploitation, when they do not benefit from this.
- Individuals are not capable of determining the fate of their personal data, whereas they may be good willing to offer it for the social good (e.g. better policy making, research) or they want to make direct profit from it.

2 State of the art

One of the most remarkable novelties in GDPR is the concept of data portability. It provides the right to have data directly transferred from one data provider to another making a step towards user-centric platforms of interrelated services (De Hert et al. 2018). This relates to the concept of data interoperability that embodies the complex network of users interaction based on personal data flow. To this scope, it is fundamental the use Semantic Web (Berners-Lee, Hendler, Lassila, et al. 2001), that brings structure to the meaningful contents of the Web by promoting common data formats and exchange protocols. The form of its most successful incarnation is Linked Data: data published in a structured manner, in such a way that information can be found, gathered, classified, and enriched using annotation and query languages.

One of the most recent approach that involves the use of distributed technologies and Semantic Web integration in OSNs is the Solid project (Sambra et al. 2016). Led by the creator of the Web Tim Berners-Lee, the project was born with the purpose of giving users their data sovereignty, letting them choose where their data resides and who is allowed to access and reuse it. Solid provides us a strong reference for this work because it uses Semantic Web technologies to decouple user data from the applications that use this data. Data is, indeed, stored in an online storage space called Pod, a Web-accessible storage service, which can either be deployed on personal servers or on public servers.

A great variety of solutions, instead, involve the use of Distributed Ledger Technologies (DLTs) for the management of general-purpose and personal data and few of them are in compliance with GDPR. A DLT is a software infrastructure maintained by a peer-to-peer network, where the network participants must reach a consensus on the states of transactions submitted to the distributed ledger, to make the transactions valid. The role of DLTs is to provide a trusted and decentralized ledger of data preserving immutability, traceability, transparency and pseudo-anonymity. The concept of DLT is the natural extension of the "blockchain" concept, because it includes those technological solutions that do not organize the data ledger as a linked list of blocks. The buzzword *blockchain* has been presented together with the technology that has changed radically the vision that we have of the Internet, finance technologies, trust in communication and even digital democracy. It was made famous by Bitcoin (Nakamoto 2009), but the decentralized computation enabled by the Ethereum blockchain (Buterin et al. 2013) enhanced the technology allowing the creation of a powerful tool: smart contracts. These contracts are self-managed structures that do not rely on a central control, thus eliminating the presence of single point of failures. The use of smart contracts grants to build Decentralized Applications (dApps) and Decentralized Autonomous Organizations (DAOs) that can realize novel important applications for social good (Zichichi et al. 2019). As an example, these technologies can serve as basis for a Smart Transportation System, with particular focus on users' personal data. Users within the Transportation Network produce various kind of data coming from their vehicle or smartphone, with the same features exploited by OSNs, e.g. user's location and activities (Zichichi, Ferretti, and D'Angelo 2020). Related works for the management of Personal Data using DLTs include various proposals and many of them are also focused on GDPR (Faber et al. 2019; Chen et al. 2019; Farshid, Reitz, and Roßbach 2019). However, these studies do not address DLTs and smart contract challenges, presenting only a conceptual approach. A more technical approach can be found in (Truong et al. 2019). Meanwhile, limiting the view on data privacy, novel works propose the protection of data (or part of it) in DLTs maintaining the ability to verify transactions (Al-Bassam et al. 2017). Several methods of accessing data through Smart contracts have been thoroughly studied in literature (Siris et al. 2019; Maesa, Mori, and Ricci 2017) and still many scenarios are conceivable, including the use of human-readable smart contracts [6].

3 Solution proposed

The solution I propose supports the right of individuals to the protection of their personal data, data interoperability, economic exploitation and social good. In order to avoid the concentration of personal information and its opaque transfers, the system stores and transact personal data in a controlled, transparent and non-centralized manner. Personal data interoperability must be favoured, hence the system uses a set of common languages and protocol, to facilitate cross-domain application and services. In the system an user can define both high-level goals and fined-grained preferences for what regards the access to his data; smart contracts can be used, since these allows to represent and reason with policies. My solution involves the use of Decentralized Technologies together with Semantic Web technologies to satisfy the principles posed above.

3.1 Data that users give to systems

This type of data is given in input by an individual to the system he is using, including self-tracking information (location), OSNs data (including videos, pictures, texts and tweets), emails and videos. It is important to notice the case of OSNs because these have increased the potential to collect personal data that individuals consciously give to systems. In my previous work I proposed novel

[6] https://zenroom.org/

system architecture that allows to create, store and share data generated by users in a Smart Transportation System (Zichichi, Ferretti, and D'Angelo 2020). This use of distributed ledgers and related technologies can serve as the basis to build novel smart services and to promote social good for what concerns individuals' personal data. Part of data within a Transportation Network actually concerns the same features exploited by OSNs, e.g. user's location and activities. The approach I intend to take in this paper stems from the structure shown in the previous one. The shift from current centralized technologies consists in the fact that the user completely controls access on data he generated. This infrastructure gives the opportunity to build a Data Marketplace based on the personal data individuals decide to sell or to set access rules in order to provide data for the social good.

3.2 Data that systems extract and process from users

This type of data is extracted by a system from its user and is collected on behalf of other entities (that may be the same OSNs of the previous case). The reasons could be different, such as surveillance, harvesting people's activities or structurally required (e.g. a Mobile Service Provider must know its users' location in order to redirect telephone calls). It is necessary to stress the fact that the entity that controls the system that generated this data can claim its control (Hearn 2010). The more the data is centralized in "silos" (not communicating between each other), the more individuals lose control over their personal data information. Conversely, in this work my focus in on the use of a unique digital space for each data subject, where data flow is ruled and data providers and consumers can meet to transact. A possible solution to this can be achieved through decentralization and shared standards.

3.3 Distributed Ledger Technologies

In particular, the use of DLTs to represent and transact with personal data would grant data validation and access control, as well as no central point of failure, immutability and most importantly traceability. Moreover, it is possible to use decentralized file systems, that allow continuous data availability. These properties are necessary in order to associate each individual to the digital space that will contain his personal data and that will be used to attend the requests of data providers and data consumers. Crucial is the use of smart contracts, since they provide a new paradigm where unmodifiable instructions are executed in an unambiguous manner during a transaction between two parts. Without the presence of a third party, then, a user may completely control the access to his personal data, being sure that his decisions on how and when to access his data are always observed. Every process is completely traced and permanently stored in the blockchain. For what concerns the expression of legal requirements and privacy preferences, and the compliance with GDPR, smart contracts unintelligibleness (i.e. how their instructions, expressed in a programming language, become a contract) still needs deep investigation. Nevertheless, smart contracts algorithms should be described to satisfy at least the following reasoning tasks: (i) determine if a policy satisfies the legal requirements; (ii) determine if a data request can be satisfied according to the individual's preferences and the legal constraints. On the other hand interoperability can be best achieved if a network of ontologies is used to model the personal data life-cycle and their actors.

3.4 Semantic web based policies

The smart contracts must be thus represented in a language that favours reasoning and a language that eases interoperability. Fortunately, the W3C has published over the last twenty years a set of specifications to describe resources which simultaneously addresses these two design goals: those of the semantic web. In the most spread paradigm, information is represented using RDF (Resource Description Framework). In this framework, resources are identified with URIs and described with collections of triples. The precise meaning of each resource can be formally established with OWL ontologies. Through the use of these ontologies it is possible to convey the meaning of data, hence to facilitate cross-domain applications and services.

The two advantages of 'interoperability' and 'reasoning' can be now well illustrated: first because the aforementioned ontologies are recommended by the W3C and thus universally understood. Second, reasoning with the information represented using these data models is easy because they are mapped in a formal language. An individual may want to say: whenever I am in the province of Lombardy, I want my data not to be transacted. If properly connected to other datasets, the system knowing that the individual is in Milano will infer that the individual is also in Lombardy and should not tranfer the data.

3.5 Zero Knowledge Proof

The use of “suitable” cryptographic techniques, such as Zero Knowledge Proof, may allow to prove that an individual possesses a certain property without revealing his data. For instance, using Zero Knowledge Proof of Location (Wolberger and Fedyukovych 2018) is possible to prove that an individual finds himself in a certain zone without revealing his exact location.

3.6 Vision

After having explained how a unique digital space can be built, it is fundamental to explain its use. The main idea is that this infrastructure can lead personal data flow towards a ”safe” place where the individual can enforce his rights. There are different actors behind the successful implementation of this vision. First of all, the individual is obviously favoured because he assumes the full control over such digital structure. Then, all the actors behind the decentralized structure are incentivized by the use of the technology specification itself, e.g. monetary retribution. Finally, the main actors who use the space both to provide and gather data, i.e. data providers and consumers, are the one to which focus on. In particular, GDPR requires data providers to release personal data to data subjects, but this does not implies the use of the digital space. The use of common standards provided by Semantic web is a necessary incentive, but not sufficient. Hence both providers and consumers must be incentivized by the data market that generates behind the digital space. This is matter of investigation, in particular regarding the complex structure that the system may assume.

4 Objectives

The general objective of this thesis is to design methods and systems to support the right of individuals to the protection of their personal data, at the same favoring its portability and economic exploitation and fostering the social good. By the description given so far, it is clear how there is a tendency to turn an individual to a simple source of data, concentrating the entire decision-making and operational power to the entity that profit by that. My main scope is to invert this trend. In order to attain the general objective, the following sub-objectives must be pursued:

- **O1** - To design methods and systems that store and transfer personal data in a controlled, transparent and non-centralized manner, avoiding thus the concentration of personal information and its opaque transfers.
- **O2** - To identify modeling and evaluation methodologies for the analysis of decentralized and complex systems, such as those considered in this domain, e.g. to understand possible actors and manners to infer data.
- **O3** - To specify languages and protocols that favour personal data interoperability.
- **O4** - To specify the languages and algorithms necessary to represent and reason with policies in smart contracts to govern the access to personal data, enabling the definition of both high-level goals and fined-grained preferences.

5 Hypotheses

The general hypothesis is that a solution to this can be achieved through decentralization and shared standards. Particularly:

- **H1** - The use of DLTs to represent and transact with personal data would grant data validation and access control, as well as no central point of failure, immutability, and most importantly traceability.
- **H2** - It is possible to use decentralized file systems for storage in order to allow continuous data availability.
- **H3** - Location privacy can be guaranteed through “suitable” cryptographic techniques, e.g. Zero Knowledge Proof
- **H4** - Interoperability can be best achieved if data models adapt the W3C specifications for the semantic web.

- **H5** - By means of defeasible deontic logic in smart contracts individuals are able to state how their personal data is managed.
- **H6** - Operating with these technologies is fast enough to ensure the "correct" execution of processes that require individuals' personal data.

6 Research Questions

My main research question is: Are decentralized technologies and semantic web standards able to optimally support individuals' personal data protection and interoperability? A series of sub-questions may derive by this:

- **RQ1** - Is it possible, using these technologies, to handle large quantity of data maintaining privacy and efficiency in indexing and accessibility? And how can it be evaluated?
- **RQ2** - Is the current specification of smart contracts able to assure the correct execution of individuals intentions?
- **RQ3** - Which challenges to the use and diffusion of semantic web technologies do entities, that extract and/or process data from individuals, present?

7 Methodology

Methodology is designed using the "divide and conquer" strategy, i.e. the general objective will be solved decomposing it into different sub-objectives. Then, to solve each sub-objective different strategies and alternatives are provided:

- **O1 M** - An infrastructure based on DLTs and decentralized file system will be specified, where each individual will be associated to a digital space that will contain personal data. This space will be used to attend the requests of data providers and data consumers. The whole architecture and protocols will be designed. The methodology applied in this sub-objective can be considered requirement-driven and empirically validated.
- **O2 M** - The methodology regarding O1 is empirical, in the sense that the approach will be evaluated with a number of test subjects to demonstrate its feasibility and analyze its performance and security. But the evaluation itself must be evaluated because in such case we are dealing with a complex system that does not present a regular structure. Hence, standard system evaluation methods may not be sufficient and the study of compliant methods (e.g. complex networks analysis) is an objective itself.
- **O3 M** - A network of ontologies will be developed to model the personal data life-cycle and their actors. Data will be modeled using W3C specifications such as Resource Description Framework (RDF).
- **O4 M** - The language elements to be reasoned with will be supported by smart contracts to be designed. The design of algorithms will be focused towards the GDPR compliance and the following reasoning tasks: (i) determine if a policy satisfies the legal requirements; (ii) determine if a data request can be satisfied according to the individual privacy preferences and the legal constraints.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN EJD "Law, Science and Technology Rights of Internet of Everything" grant agreement No 814177.

References

Al-Bassam, Mustafa, Alberto Sonnino, Shehar Bano, Dave Hrycyszyn, and George Danezis. 2017. "Chainspace: A sharded smart contracts platform." *arXiv preprint arXiv:1708.03778.*

Berners-Lee, Tim, James Hendler, Ora Lassila, et al. 2001. "The semantic web." *Scientific american* 284 (5): 28–37.

Buterin, Vitalik, et al. 2013. *Ethereum white paper.* `https://github.com/ethereum/wiki/wiki/White-Paper`.

Chen, Yun, Hui Xie, Kun Lv, Shengjun Wei, and Changzhen Hu. 2019. "DEPLEST: A blockchain-based privacy-preserving distributed database toward user behaviors in social networks." *Information Sciences* 501:100–117. ISSN: 0020-0255. doi:`https://doi.org/10.1016/j.ins.2019.05.092`.

De Hert, Paul, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. 2018. "The right to data portability in the GDPR: Towards user-centric interoperability of digital services." *Computer Law & Security Review* 34 (2): 193–203.

Faber, Benedict, Georg Michelet, Niklas Weidmann, Raghava Rao Mukkamala, and Ravi Vatrapu. 2019. "BPDIMS: A Blockchain-based Personal Data and Identity Management System" [in English]. In *Proceedings of the 52nd Hawaii International Conference on System Sciences,* 6855–6864. United States: Hawaii International Conference on System Sciences (HICSS). ISBN: 9780998133126. doi:`10125/60121`.

Farshid, Simon, Andreas Reitz, and Peter Roßbach. 2019. "Design of a forgetting blockchain: A possible way to accomplish GDPR compatibility." In *Proceedings of the 52nd Hawaii International Conference on System Sciences.*

Hasan, Samiul, Satish V Ukkusuri, and Xianyuan Zhan. 2016. "Understanding social influence in activity location choice and lifestyle patterns using geolocation data from social media." *Frontiers in ICT* 3:10.

Hearn, Alison. 2010. "Structuring feeling: Web 2.0, online ranking and rating, and the digital'reputation'economy." *ephemera: theory & politics in organization* 10.

Kitchin, Rob. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage.

Maesa, Damiano Di Francesco, Paolo Mori, and Laura Ricci. 2017. “Blockchain based access control.” In *IFIP international conference on distributed applications and interoperable systems,* 206–220. Springer.

Nakamoto, Satoshi. 2009. *Bitcoin: A peer-to-peer electronic cash system.* `http://www.bitcoin.org/bitcoin.pdf`.

Pangrazio, Luci, and Neil Selwyn. 2019. “‘Personal data literacies’: A critical literacies approach to enhancing understandings of personal digital data.” *New Media & Society* 21 (2): 419–437.

Sadilek, Adam, Henry Kautz, and Jeffrey P. Bigham. 2012. “Finding Your Friends and Following Them to Where You Are.” In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining,* 723–732. WSDM ’12. Seattle, Washington, USA: Association for Computing Machinery. ISBN: 9781450307475. doi:`10.1145/2124295.2124380`. `https://doi.org/10.1145/2124295.2124380`.

Sambra, Andrei Vlad, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. 2016. “Solid : A Platform for Decentralized Social Applications Based on Linked Data.”

Siris, Vasilios A, Dimitrios Dimopoulos, Nikos Fotiou, Spyros Voulgaris, and George C Polyzos. 2019. “Interledger Smart Contracts for Decentralized Authorization to Constrained Things.” *arXiv preprint arXiv:1905.01671.*

Truong, Nguyen Binh, Kai Sun, Gyu Myoung Lee, and Yike Guo. 2019. “GDPR-Compliant Personal Data Management: A Blockchain-based Solution.” *arXiv preprint arXiv:1904.03038.*

Wolberger, L, and V Fedyukovych. 2018. *Zero Knowledge Proof of Location.* `https://platin.io/yellowpaper`.

Zichichi, Mirko, Michele Contu, Stefano Ferretti, and Gabriele D’Angelo. 2019. “LikeStarter: a Smart-contract based Social DAO for Crowdfunding.” In *Proc. of the 2st Workshop on Cryptocurrencies and Blockchains for Distributed Systems.*

Zichichi, Mirko, Stefano Ferretti, and Gabriele D’Angelo. 2020. “A Distributed Ledger Based Infrastructure for Smart Transportation System and Social Good.” In *IEEE Consumer Communications and Networking Conference (CCNC).* Las Vegas, USA.