Acknowledging Value of Personal Information: a Privacy Aware Data Market for Health and Social Research

Francesco Bruschi[1], Vincenzo Rana[1], Alessio Pagani[1], and Donatella Sciuto[1]

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, Italy
<francesco.bruschi, vincenzo.rana, alessio.pagani, donatella.sciuto>@polimi.it

Abstract

Gathering information to perform health or social research is a complex endeavour. Users are wary of sharing medical and, more generally, personal data. Furthermore, as they grow more conscious about privacy concerns (which is socially desirable) and of the value of their own sensitive data, obtaining information even for research purposes will become increasingly harder. On the other hand, as automatic data analysis and inference tools and techniques become more and more effective, the potential value of having greater amounts of data available increases. In this paper, we present a scenario that encompasses recent technologies to create a personal data market in which users are spurred to gather and record personal information in a secure way, maintaining ownership through cryptography. The main incentives come from the fact that research actors acknowledge user data value by purchasing it: when a research actor needs users personal information, he makes a bid, to which users respond providing the information required. We explore the possibility of using a set of technologies, such as smart contracts and trusted computing, to guarantee both the information buyer about the data quality and authenticity, and the seller that the contract will be honored, even in the total absence of reciprocal trust (the parties could be unknown to each other, or even completely anonymous).

1 Introduction

In the next future, people will have increasing opportunities of gathering information about themselves, recording everything coming from their wearable sensors, or from sensors in the environment that surrounds them. This is producing a sort of personal omniscient diary. For one extreme, but not so fictional take, check the Black Mirror episode "The entire history of your life", in which a chip interfacing directly with the nervous system is used to record all the visual perception of the person in which it's implanted. One can think of the same pervasive recording, only generalized and extended to accelerometric sensors, heartbeat sensors, etc. Some remarkable examples of wearable sensors can be found in [8], where Huang et al. show how it is possible to record arterial diameter changes by means of a wearable ultrasound sensor.

Currently, samples to be used for correlation purposes are mainly gathered in one of the two following ways:

- 1. People are asked to participate in a research, are observed, and are asked the permission to use the observation gathered.

- 2. Anonymized data are used.

For example, Fitabase offers researchers a platform in which data gathered from patients using fitbit (a popular tracking personal device) is collected and can be analyzed. As of now, 482 researches have been carried out using Fitabase. The service works this way: patients are given a fitbit device (if they don't have one already), and instructions on how to connect it to the platform. They are then asked to provide other information needed for the study (does Fitabase collect those information?). The platform then collects everything, and offers the researchers tools to automatically represent and analyze the data. The more datasets regarding a patient can be accessed, the better it is of course to investigate correlation among factors and variables. Recently, many approaches have been explored to enable selective and secure sharing of medical records using blockchain technology. In short, blockchains [12, 13] are digital data structures that employ distributed consensus protocols to implement immutable append-only logs, with particular resistance to tampering. The most famous blockchain is the one powering the Bitcoin cryptocurrency, which uses a particular mechanism to reach consensus and resist to tampering and attacks such as sybil ones. The extremely powerful abstraction that the blockchain offers, that is that of an append-only, publicly accessible log, is being considered for applications different from recording untamperable money accounts. In particular, its use is being advocated among the others in healthcare data management, protection and sharing. In [9] authors build, upon a blockchain, an immutable log that allows users to access personal medical information, across providers and treatment sites. In [18] authors describe an articulated platform architecture for clinical trial and precision medicine, based on a blockchain. The platform addresses different problems in the implementation of clinical trials: data management and integration, parallel computing, verifiable anonymous identity management and trusted data sharing.

In Estonia [19], Guardtime [https://guardtime.com/], a provider of enterprise blockchain solutions, has collaborated to the creation of a digital health infrastructure, that allows citizens, healthcare stakeholders (providers, insurances, etc.) to digitally retrieve all the medical information about treatments in a controlled manner.

In this paper, as depicted in Figure 1, we propose a perspective scenario in which users have market incentives to gather as much information as possible, keeping ownership (not only legal) of it, and selectively disclosing it to research entities, in change of some benefits (e.g., money). In particular, we give an overview of a possible approach for implementing a personal data marketplace based on the blockchain technology and analyse the current status of the related technologies, while we leave the definition of a complete proposal for future work.

2 Personal data market architecture

The first element considered is data gathering and storage. Currently, data produced by sensors, whether personal, wearable, environmental or whatever, are transmitted by the hardware and stored locally on the device, on personal storage (e.g. pc hard disk), or sent in the cloud (eg: google maps). For instance, Fitbit buffers data locally, and requires syncing with a cloud warehouse when the buffer is full. The syncing exploits user's smartphone or pc as a gateway. Even though the communication with the PC/phone via bluetooth is encrypted, data on the server seems to be not. This means that the user doesn't have exclusive access to the data. In our scenario, all the information from the devices (and applications) is encrypted with a key that is known only by the user. Ideally, the information is encrypted since its gathering, on the device hardware, and then stored, on personal storage devices, or in the cloud. This only

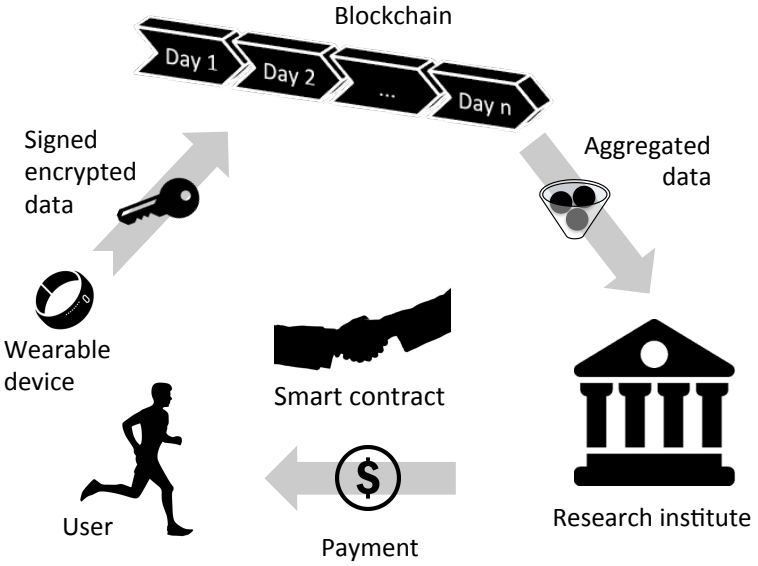

Figure 1: Schema of the proposed approach.

requires standard cryptography. Possibly, the information can be notarized on a blockchain for timestamp and/or subsequent verification of ownership. Every user then securely records all of her life this way. Now let's imagine that a research entity wants to investigate the correlation between "amount" of running and blood pressure trends in users. What they need is to have, for as many users as possible:

- 1. info from a tracker such as fitbit.
- 2. blood pressure measurements, at a sampling rate above a certain threshold (say once a week).

Currently, such a research entity would ask users to enroll in a platform such as Fitabase, and to provide blood pressure information. Users, on the other hand, have no incentive whatsoever in doing so, other than the ethical urge to help in a research. Moreover, they could be reluctant to give away sensitive information such as complete records of their position, or detailed heart rate recordings.

Now an alternative way, core in the perspective we're imagining, would be for the research institution to *buy* the data from users. Moreover, since they probably don't need all the track details, and the users aren't willing to disclose them completely, they could ask for a subset, or a *function* of the complete information (such as, for instance, the average heart rate over periods of 10 minutes, devoid of localization information). The possibility to monetize this information would of course create an incentive for users to record more and more information. This in turn creates a problem: how can the research institute be sure that the user didn't provide fake information (for instance, that he generated fake tracks or blood pressure measurements?). The answer could be a technology that is called *zero knowledge proofs*. Zero knowledge proof is a set of techniques that allow to generate a mathematically convincing proof, with arbitrary confidence, of facts such as: this set of values a1, a2, ... has been obtained by decrypting this

data D, and then averaging windows of 200 values. Crucially, the proof doesn't need to reveal the key! In this way, the researcher could be sure about one feature of the data (namely, that they are the average of a set D, owned by the user).

But how it is possible to trust that data D has been indeed gathered by a Fitbit during user's exercise, and not generated synthetically? This could be guaranteed with the set of techniques of the so called "trusted computing". The tracker could be equipped with a trusted processor, that would in turn guarantee that a certain data set D was generated by a trusted tracker, with a given software, by signing it with a key embedded in its hardware. Combining zero knowledge proof and trusted computing, a user could provide a proof that a set of data is indeed the average of his heart beat, as measured by a trusted device.

Next problem to consider is how the user can guarantee that his blood pressure readings are "real". To this aim, pharmacies could provide a service in which they read pressure and produce a (digitally) signed certificate of the reading. The certificate would contain the reading, the time, and a proof of the user identity. He will then be able to attach it to the tracker data, again in a certifiable manner.

The final issue regards how the user can be sure that he will be paid, once he provides his certified data. This doubt can have varying importance, according to the credibility and reputation of the research institution. One desirable feature (since everyone would gain from that) would be to minimize reputation requirements for the data buyers. This can be achieved by using *smart contracts*: the research entity deploys a smart contract that guarantees, upon execution, that whenever data with some certified feratures (that it was generated by a trusted tracker by averaging values, that the pressure readings are certified and belong to the same user, etc) will be provided, the owner of the data will be paid a certain amount of (digital) money.

This mechanism could create a data market where:

- 1. users will be encouraged, through actual economic incentives, to record everything about their lives, and to retain ownership of the gathered data;
- 2. the amount of potentially valuable data will increase in size and pervasiveness;
- 3. research institutions will be able to obtain required certified data by paying them to users.

3 Analysis and state of the technologies involved

In this section we analyze the state of the mentioned technologies and their availability, both technically (e.g., are they computationally feasible?) and "commercially" (is there an off the shelf implementation that can be used right now?).

3.1 Zero knowledge proofs

Zero knowledge proofs are mathematical/logical tools that allow to provide evidence of the truth of a statement, without giving any other information. To make an example, suppose that Bob knows the solution to a complex sudoku puzzle. Through zero knowledge proofs, it is possible for Bob to convince Alice that he indeed knows a solution, without revealing anything about it. Moreover, Alice won't be able to "highjack" the proof and convince anyone else that *she* knows a solution. Zero knowledge proofs were first introduced, as a theoretical possibility, by Micali et al. [20] in 1985, as a form of *interactive* proof: the verifier of the proof repeatedly

challenges the prover in various ways; each challenge overcome increases the confidence of the verifier in the truth of the assertion, without revealing anything else. That is repeated until the required confidence is reached. Remarkably, the possibility of zero-knowledge proving is very general: every Non-deterministic Polinomial (NP) problem (that is, every decision problem for which a solution can be tested in polynomial time by a deterministic Turing machine or, equivalently, every decision problem that can be solved by nondeterministic Turing machines in polynomial time or) admits zero knowledge proofs. In their original version, as said, the proofs are *interactive*, that is, they require a verifier and a prover to communicate back and forth challenges and responses. In 1988, Micali et al. introduced the notion of Non Interactive Zero Knowledge Proofs (NIZK) that, at least in theory, greatly simplifies the verification process. Early protocols for NIZK, on the other hand, required prohibitive computational power. As the problem begun to show applicative interest, efforts were poured into refining proof protocols, to obtain practical feasibility. In the scenario proposed, NIZK would allow the owner of the data to selectively disclose part of the information about themselves (e.g.: only average heart rate over windows of ten minutes instead of punctual information, annotated with geo tagging). They could in fact provide the selected/computed data, together with a proof that it was obtained from a sequence provably produced by a trusted device. NIZK today are employed in real world application, the most notable probably being z-cash (https://z.cash/). In z-cash, zero knowledge proofs are used to transfer money from one wallet to another, without disclosing any information about the amount and the source and destination of the transfer [5]. Z-cash exploits a software library for the generation and verification of non interactive zero knowledge proofs that is freely available, and can be used to design other applications.

3.2 (Embedded) Trusted computing

The term *trusted computing* includes a set of technologies that aim at making it possible to guarantee that a digital system is behaving in expected ways. The technology is being defined and proposed by the Trusted Computing Group, an association comprising AMD, Hewlett-Packard, IBM, Intel and Microsoft. The technology is articulated into several key concepts: Endorsement key, Secure input and output, Memory curtaining / protected execution, Sealed storage, Remote attestation, Trusted Third Party (TTP) [21]. Almost all of them are relevant to our scenario.

Endorsement Key is the private component of a public/private couple key that is generated at the time of device production, and that is embedded into the hardware (and nowhere else), while the public part is (guess!) made public. In this way, it is possible to produce commands that only the device can decipher and, most relevant for the problem at hand, it is possible for the device to sign data in a way that cannot be faked (i.e., it is not possible to run a program that fakes a running session on a PC, since it could not sign it). If this was embedded in the sensors, they would be able to generate provably original data series.

Memory curtaining refers to protection mechanisms that make parts of the memory inaccessible from the outside, and even to part of the running software. In particular, encryption keys shouldn't be readable even by the operating system. In our scenario, this would mean not being able to extract the private key and then forging fake tracking data.

Remote attestation refers to the fact that the device can produce a certificate of the particular hardware and software it is currently running. In our scenario, this would prevent malicious users to substitute the software/firmware of the device, with the aim for instance of making it generate fake data instead of gathering it from the sensors.

All these technologies are currently implemented and shipped on almost all PC compu-

tational platforms (Intel, AMD), and are also available on some mobile/embedded hardware platform, such as ARM processors [22]. It would then be possible, as of now, to build sources of personal and environmental data protected from faking, and then of greater value in a market scenario.

3.3 Smart contracts

Smart contracts are protocols that aim at automating enforcement of agreements among parties. The most remarkable thing about smart contracts is that they are programs, and that their execution doesn't require any trusted intermediary, but is guaranteed by a decentralized system, such as a blockchain. They were first introduced by Nick Szabo, who defined them '*a computerized transaction protocol that executes the terms of a contract*' [6]. At the moment, the most active community is that around the Ethereum platform [23]. In Ethereum, smart contracts are accounts that hold a balance (the currency in the Ethereum platform is called *ether*) and contain code functions that define how to interact with other contracts. They can take decisions, change state, and send value to other contracts, in response to users invoking their functions. Correct execution of contracts, and consensus on its effects, is provided by the Ethereum network itself. In the proposed scenario, smart contracts would implement the bid that the research actors make to buy information from users. A smart contract could be triggered by a user publishing a data set, encrypted with the public key of the buyer, and of the proofs of authenticity previously discussed. The code of the smart contract would verify the proof, and proceed with the payment of the agreed amount.

4 Conclusions

In this paper we proposed a scenario in which users are encouraged to gather, store, and conserve ownership of personal data, due to incentives coming from a market in which they can valorize them by selling selectively (and securely) disclosed information to actors that can exploit them with research aims. We showed how it could be possible, using technology that is either consolidated or under heavy development, to build reciprocal guarantees between sellers and buyers of information (such as data genuinity and integrity, contract execution) without any a priori trust requirements (in an extreme situation, both users and information buyers could remain anonymous).

The mechanism proposed could potentially have a deep impact on the design of clinical and social research, both quantitatively (much more information could be available) and qualitatively (users would retain full ownership of their data, and could valorize it monetarily. Almost all of the technologies are in a developmental stage that would allow to implement the features considered. The only critical ones are non interactive zero knowledge proofs, since efforts to reduce the computational burden needed to forge proofs are the focus of intense current research. Future planned work concerns the theoretical analysis and simulation of the potential market, and the implementation of a full-stack prototype.

References

[1] M. R. Powell and W. J. To. Redesigning the research design: Accelerating the pace of research through technology innovation. 2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH), Orlando, FL, pp. 1-5.

[2] Xia, Qi and Sifah, Emmanuel and Asamoah, Kwame and Gao, Jianbin and Du, Xiaojiang and Guizani, Mohsen. MeDShare: Trust-less Medical Data Sharing Among Cloud Service Providers Via Blockchain. 2017 IEEE Access. PP. 1-1.

[3] Cyr, Britt, Webb Horn, Daniela Miao and Michael A. Specter. Security Analysis of Wearable Fitness Devices (Fitbit). 2014.

[4] Goldwasser, S.; Micali, S.; Rackoff, C. The knowledge complexity of interactive proof systems. 1989. SIAM Journal on Computing, Philadelphia: Society for Industrial and Applied Mathematics, 18 (1): 186-208.

[5] Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, Madars Virza. Zerocash: Decentralized Anonymous Payments from Bitcoin. 2014. IEEE Symposium on Security and Privacy pp. 459-474

[6] Nick Szabo. Smart Contracts: Building Blocks for Digital Markets. 1996. www.fon.hum.uva.nl. Retrieved 2017-07-29.

[7] The Entire History of Your Life?, http://www.imdb.com/title/tt2089050/x

[8] A. Huang, M. Yoshida, Y. Ono and S. Rajan.Continuous measurement of arterial diameter using wearable and flexible ultrasonic sensor. 2017 IEEE International Ultrasonics Symposium (IUS), Washington, DC, 2017, pp. 1-4.

[9] A. Azaria, A. Ekblaw, T. Vieira and A. Lippman. MedRec: Using Blockchain for Medical Data Access and Permission Management. 2016. 2nd International Conference on Open and Big Data (OBD), Vienna, 2016, pp. 25-30.

[10] Bob Marvin. Blockchain: The Invisible Technology That's Changing the World. 08/2017. PC MAG Australia. ZiffDavis, LLC. Retrieved 25 September 2017.

[11] Popper, Nathan. Ethereum, a Virtual Currency, Enables Transactions That Rival Bitcoin's. 03/2016. New York Times. Retrieved 2017-02-07.

[12] Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System. 2008.

[13] Bonneau J, Miller A, Clark J, Narayanan A, Kroll JA, Felten EW. SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies. . S&P 2015.

[14] Vitalik Buterin. Ethereum: A next-generation smart contract and decentralized application platform. 2014.

[15] Wood G..Ethereum: A secure decentralised generalised transaction ledger. 2014.

[16] Andrychowicz M, Dziembowski S, Malinowski D, Mazurek. Fair Two-Party Computations via Bitcoin Deposits. 2014. FC.

[17] Delmolino K, Arnett M, Kosba A, Miller A.. Step by Step Towards Creating a Safe Smart Contract: Lessons and Insights from a Cryptocurrency Lab. 2016. Shi E. FC.

[18] Z. Shae and J. J. P. Tsai. On the Design of a Blockchain Platform for Clinical Trial and Precision Medicine. 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, pp. 1972-1980.

[19] O. Williams-Grut, Estonia is using the technology behind bitcoin to secure 1 million health records, March 2016, [online] Available: http://www.businessinsider.com/guardtime-estonian-health-records-industrial-blockchain-bitcoin-2016-3?r=UK&IR=T.

[20] Blum, Manuel; Feldman, Paul; Micali, Silvio. Non-Interactive Zero-Knowledge and Its Applications. 1988. Proceedings of the twentieth annual ACM symposium on Theory of computing (STOC 1988): 103-112.

[21] Chris Mitchell. Trusted Computing. 2005. Institution of Electrical Engineers.

[22] B. Ngabonziza, D. Martin, A. Bailey, H. Cho and S. Martin. TrustZone Explained: Architectural Features and Use Cases. 2016. IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA, pp. 445-451.

[23] https://ethereum.org/