MedChain: Accountable and Auditable Data Sharing in Distributed Medical Scenarios

Juan Ramón Troncoso-Pastoriza, Jean Louis Raisaro, Linus Gasser, Bryan Ford, and Jean-Pierre Hubaux

École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Introduction. The current trend towards personalized medicine creates an urgent need to share data among different hospitals and health institutions, which endangers the privacy of the data subjects if not done with the appropriate precautions. Conversely, the frequency of data breaches in the healthcare industry has been rising since 2010 [3, 5], severely holding back health institutions from exposing and sharing their data for the fear of being the next target of cyberattacks. In this landscape, the ability to provide strong auditability, accountability and traceability of the system events plays a role as important as data confidentiality for the purpose of enabling secure and privacy-conscious data sharing, breach detection and fast recovery. National and international regulations (e.g., HIPAA [7] in the United States and the GDPR in Europe [1]) impose strong requirements both in terms of confidentiality, i.e., prevention of undue data leakages and restriction of data access, and accountability, i.e., recording of all data accesses and exchanges carried out by any entity with the purpose of identifying misbehaving individuals. This is especially relevant for medical and genomic data, whose (un)intended leakage can severely harm individuals' privacy and institutions' reputation. Current operational systems for medical data sharing are lacking in terms of privacy protection and/or transparency guarantees that can address these challenges, and they provide a weak federated or centralized model of identity and access control that can endanger the whole network if only one of the sites is breached.

In this talk, we propose *MedChain*, a novel system featuring distributed, flexible and fully decentralized identity management and access control mechanisms based on distributed ledger technologies, that enable (a) full traceability, auditability and accountability of all system events through immutable logs with no single point of failure, particularly dealing with the access to and usage of medical and genomic data, and (b) fine-grained configurable and privacy-conscious access control enforced through smart contracts (protocols to digitally enforce and verify the execution of a set of agreed actions). We exemplify the use of the system through an application to distributed feasibility studies, by integrating it in the currently most widespread cohort explorer tools (i2b2 [4] and SHRINE [8]).

System and attacker model. *MedChain* supports a network of mutually distrustful health institutions comprising universities, research laboratories, and hospitals (data providers) holding patient data that is securely shared with the researchers (queriers) of the network. In this environment, authorized researchers can run queries over the distributed database for the purpose of feasibility studies and data analytics. The same model in [6] is applicable in terms of confidentiality requirements, which can also be governed by *MedChain* policies and addressed through secure computation protocols. Additionally, *MedChain* must cope with internal or external attackers who might try to break into each of the institutions and perform malicious actions in the system, which have to be traceable, auditable, and reversible, even if the attacker gets full control over the local logs of the breached node. Conversely, in order to enforce liability as defined by the applicable regulatory frameworks and internal organizational rules, each institution must remain self-sovereign over their patient's data and the management of their users (affiliated researchers and IT staff, among others).

Proposed solution. *MedChain* employs a private permissioned blockchain to address the aforementioned issues by means of distributing trust among different independent servers (which represent the institutions in the network) and implementing consensus rules and smart contracts to define the identity management, access control, and consent enforcement; hence, any misbehavior can be detected and traced. The underlying blockchain in *MedChain* is based on OmniLedger [2], a scalable and highly efficient distributed ledger that enables atomic transactions and low-latency validation, by means of skipchains (a combination of blokchains and skiplists) and BFT (Byzantine Fault Tolerance) consensus rules.

Current blockchain-based approaches to identity management and access control employ a blockchain solely to enforce transparency of a centralized identity manager, therefore hindering self-sovereignty. Contrarily, identities in *MedChain* are completely distributed and organized in a hierarchical graph of identities that can be collectively stored and verified but independently managed and atomically updated, therefore (a) avoiding trust in a centralized look-up system and (b) overcoming race conditions present in traditional federated identity systems.

Regarding access control, *MedChain* links each data source with a structure that defines the identities and/or groups of identities (as defined in the hierarchical identity graph) authorized to perform determined actions on the data, as access policies expressed in a JSON-based language. This structure also defines the set of identities that are authorized to perform changes or updates in each of the policies; these updates are executed atomically and are recorded in the underlying blockchain. This approach provides the nodes in the network with a flexible access control whose expressiveness goes beyond the rigidity of currently used role-based access control systems.

When an access request by a researcher arrives to *MedChain*, a smart contract is run to check the identity and access rights of the researcher and enforce the applicable policies on the queried data source(s) before granting or denying the requested access to the data in a privacy-conscious way. All the requests and their authorizations are also atomically and immutably stored in the underlying blockchain, providing a secure and tamper-proof evidence for auditability and accountability purposes.

Application in feasibility studies. We implement and showcase *MedChain* in an operational setting within a representative medical and genomic data sharing scenario. For this purpose, we integrate *MedChain* into i2b2 [4] and SHRINE [8], which are the most widespread and established tools for cohort exploration in clinical research; these tools are lacking in terms of data protection guarantees, and they enable only a federated access control with institutional-based access, in which the data sources can only grant, log, and revoke the access to their resources based on the credentials of the requesting institution (instead of the individual researcher); additionally, the logs are kept locally at each institution. The integration of *MedChain* addresses this shortcomings in an efficient way. In order to enable data confidentiality, we also integrate the cryptographic modules of *MedCo* [6], a privacy-preserving layer on top of i2b2. *MedChain* is seamlessly integrated into the i2b2/SHRINE workflow by replacing the access control and identity modules (implemented in the so called Project Management -PM- cell) by those provided by *MedChain*, and distributedly maintained by the set of institutions participating in the network.

In this scenario, authorized researchers affiliated with one of the registered institutions have a unique identity in the system and a private-public key pair (accessible through a smart card); they can use *MedChain*'s modified i2b2 web interface to login to the system and query for patients satisfying a set of inclusion/exclusion criteria at each of the data sources. The actions that can be performed in this scenario range from obtaining differentially-private (obfuscated) patient counts, exact patient counts, to lists of patient pseudoidentifiers, depending on the querier rights. *MedChain* access control structures enable full auditability with no single points of failure, and provide support for the governance of privacy-conscious sharing. This is achieved by storing the information of the access policies and the contracts defining (a) the access rights and (b) the computation of the applicable privacy budgets and privacy measures needed to authorize a privacy-conscious query. In this talk, we will show the concrete majority rules use for practical deployments, the key management mechanisms, and the achievable trade-offs in terms of efficiency, security and privacy; we will also exemplify the different configurations and use cases that *MedChain* enables, and the flexible definition and setup of policies, access rights and privacy guarantees governing the system operation.

Conclusion. *MedChain* addresses the privacy and security challenges in highly distributed medical data sharing networks by avoiding single points of failure by providing the following unparalleled properties: (a) full auditability and accountability of data access and usage, stored in distributed immutable logs that cannot be modified without trace; (b) flexible identity and key management supporting revocation and updates of the individual keys in a consistent transactional way across the whole network; (c) reproducibility of queries and tracking of dataset and system state at a network level; (d) breach tracing and quick breach recovery to a previous accepted system state as recorded in the distributed immutable log; this involves precise and accurate revocation of access instead of full institutional revocation as in current systems; (e) fine-grained distributed access control with increased versatility and flexibility with respect to the traditional role-based access control used in current systems; finally, (f) data confidentiality can be preserved through collective operation keys and distributed protocols (relying on *MedCo*), governed and enforced by the policies defined in *MedChain*.

References

[1] EU Parlament. The EU General Data Protection Regulation (GDPR). `http://www.eugdpr.org/`. Last Accessed: June 26, 2018.

[2] E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, and B. Ford. Omniledger: A secure, scale-out, decentralized ledger via sharding. In *2018 IEEE Symposium on Security and Privacy (SP)*, volume 00, pages 19–34.

[3] G. Martin, P. Martin, C. Hankin, A. Darzi, and J. Kinross. Cybersecurity and healthcare: how safe are we? *BMJ*, 358, 2017.

[4] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010.

[5] L. Ponemon Institute. Sixth annual benchmark study on privacy & security of healthcare data. Technical report, 2016.

[6] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. A. Gomes de Sá E Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. A. Ford, and J.-P. Hubaux. Medco: Enabling privacy-conscious exploration of distributed clinical and genomic data. *Accepted at IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[7] U.S. Department of Health & Human Services. The health insurance portability and accountability act (hipaa). `https://www.hhs.gov/hipaa/index.html`. Last Accessed: June 26, 2018.

[8] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and I. S. Kohane. The shared health research information network (shrine): a prototype federated query tool for clinical data repositories. *Journal of the American Medical Informatics Association*, 16(5):624–630, 2009.