Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220682

An Improved Infrastructure for Privacy-Preserving Analysis of Patient Data

Djura SMITS [a,*], Bart VAN BEUSEKOM [b,*], Frank MARTIN [b], Lourens VEEN [a], Gijs GELEIJNSE [b] and Arturo MONCADA-TORRES [b,1]
[a] *Netherlands eScience Center; Amsterdam, the Netherlands*
[b] *Netherlands Comprehensive Cancer Organization; Eindhoven, the Netherlands*

Abstract. Incorporating healthcare data from different sources is crucial for a better understanding of patient (sub)populations. However, data centralization raises concerns about data privacy and governance. In this work, we present an improved infrastructure that allows privacy-preserving analysis of patient data: `vantage6 v3`. For this new version, we describe its architecture and upgraded functionality, which allows algorithms running at each party to communicate with one another through a virtual private network (while still being isolated from the public internet to reduce the risk of data leakage). This allows the execution of different types of algorithms (e.g., multi-party computation) that were practically infeasible before, as showcased by the included examples. The (continuous) development of this type of infrastructure is fundamental to meet the current and future demands of healthcare research with a strong emphasis on preserving the privacy of sensitive patient data.

Keywords. Federated learning, multi-party computation, vantage6

1. Introduction

Healthcare data are usually scattered over many silos, making bringing them together a real challenge. In recent years, we have seen the rise of privacy-preserving frameworks that tackle this issue [1]. These allow analyzing the data without centralizing them, leaving them at their original source. Researchers can send their queries and receive an aggregate (e.g., model coefficients) in return. These pose a drastically smaller risk of data leakage (since they do not contain any personal information) while still being practically as accurate as their centralized counterparts.

Most of these frameworks lack flexibility and can only be used for specific data partitions. Moreover, they generally require all machines to be directly accessible within the same network. Lastly, many healthcare environments have strict data firewalls which block direct incoming network communication from outside the organization to protect the data, but that can also hamper the execution of certain (non-malicious) algorithms.

In this paper, we introduce the latest version of our framework for privacy-preserving analysis, `vantage6 v3`. In this new version, we added secure direct node-to-node (n2n) communication, which enables the usage of existing libraries for privacy

[*] These authors declare equal contributions to this study and should be considered joint first authors.

[1] Corresponding author: Arturo Moncada-Torres. R&D Department, Netherlands Comprehensive Cancer Organization, Zernikestraat 29, 5612 HZ, Eindhoven, the Netherlands. E-mail: a.moncadatorres@iknl.nl

preserving analysis with the additional security that our framework provides. First, we outline the design of the framework. After that, we describe our implementation of n2n communication and illustrate how this can be used to incorporate other libraries for privacy preserving analysis. We close our paper with some real use cases that have benefited from the new functionality and future outlook.

2. Methods

In `vantage6`, a researcher can send a question to the central server, which provides a communication interface, handles administrative tasks (e.g., user authentication), and communicates with the nodes. The nodes execute the chosen algorithm and report results back to the server, which can then be retrieved by the researcher. In a collaboration, each party hosts a node and has full control over which data may be accessed by which algorithms. A complete description of `vantage6`'s basic architecture can be found in [2].

In the original infrastructure, algorithms were completely isolated to minimize the risk of data breaches. This was achieved by placing the algorithm containers in a (Docker) network that only allows for communication within the network itself. We have upgraded `vantage6`'s infrastructure to version 3.0 (`v3`) to allow algorithms running on different nodes to communicate with one another via a virtual private network (VPN) connection. Importantly, the nodes can *only* communicate via the VPN network, and are otherwise still isolated from the public internet. This n2n communication broadens the range of applications that can be executed using `vantage6` (Sec. 3.1).

2.1. Infrastructure

Figure 1 shows how we extended `vantage6` with n2n communication. When a node is started, a separate VPN client container is also initialized, which handles the VPN traffic of all algorithms that run on that node for a particular collaboration. This way, each algorithm has its own communication channel over the VPN network.

For such a connection, the node requests a VPN configuration file from the `vantage6` server on startup, which is fetched by `vantage6` from the VPN server (which uses OpenVPN) and is sent back to the node. Next, the node starts a VPN client container that establishes a VPN connection. The VPN IP address for this node is sent to the `vantage6` server where it is stored.

The next step is to enable the VPN client containers on different nodes to communicate with one another. First, the VPN client container is configured to drop all internet traffic except for VPN traffic. Then, the container is attached to the isolated Docker network. Finally, the host network on the node machine is configured to direct incoming VPN traffic to the isolated network bridge and *vice versa*.

When an algorithm is initiated, its container is configured to communicate incoming and outgoing traffic via VPN. Note that the actual algorithm is not started until the VPN connection is fully set up, to prevent an algorithm trying to communicate too soon.

For incoming traffic, the algorithm developer may specify on which port(s) the algorithm may be reached by exposing them in the Docker file that is used to build the algorithm image. Note that exposing a port makes it available to the VPN client container, but does *not* expose the port outside of the (safe) isolated Docker network. By default,

one port is exposed on the algorithm container. For each port that is exposed, a free port on the VPN client container is reserved. All incoming traffic on that port is forwarded to the paired port on the algorithm container. The VPN client's port number together with the port label are sent to the `vantage6` server, where they are stored. Algorithm containers can request the addresses of other algorithm containers involved in the same task, which allows them to communicate with other nodes at any point in their execution. Outgoing traffic is routed via the VPN container, which is possible because the container is in the same isolated Docker network.

At this point, all VPN communications are set up. The execution of the algorithm can be started and will be able to communicate with algorithms on other nodes over VPN.

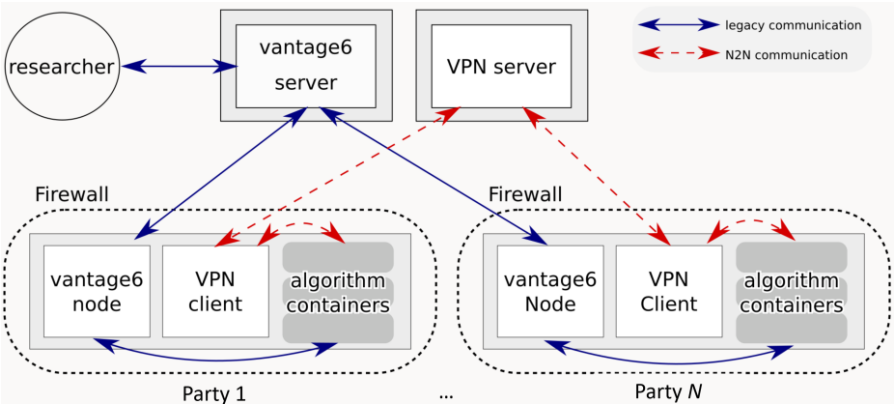

Figure 1. The improved infrastructure of `vantage6 v3`. Notice that nodes can *only* communicate through the VPN network and are still isolated from the (public) internet to minimize the risk of data breaches.

3. Results

The presented (and latest) version of `vantage6` is being actively used by several parties in a variety of projects that leverage its new features.

3.1. Applications

This new version allows the implementation and execution of algorithms based on multi-party computation (MPC). In short, MPC is a group of techniques (based on homomorphic encryption and secret sharing) that perform computations on encrypted data while protecting the privacy of the data at hand. Usually, the overhead of these computations can be quite large (especially on large amounts of data), making the execution of accurate MPC algorithms impractical or even infeasible. `vantage6`'s new ability to allow n2n communication drastically reduces the execution time for MPC-based computations, making them more accessible and usable for real-life scenarios. This has allowed the development of MPC tools for regression models and survival analysis [3], which would not have been feasible with `vantage6`'s previous version.

In another interesting use case, the University of Maastricht is collaborating with the Netherlands eScience Center to find out how socio-economic and medical factors influence the risk of heart disease by using data from different parties. Taking advantage of `vantage6 v3`'s new features, a privacy-preserving *n*-party scalar product as well as other MPC-based methods are currently being developed for secure data analysis [4].

4. Discussion and Conclusions

This new release consolidates `vantage6`'s functional principles of autonomy (allowing each party to be in control of their own data), heterogeneity (permitting differences across parties), and flexibility (allowing analysis of horizontally- or vertically-partitioned data using either federated learning or MPC techniques). Moreover, its new features have a huge potential for the creation and further development of privacy-preserving algorithms. Many external analysis libraries that require n2n communication can now be used in a secure environment, increasing the scenarios in which `vantage6` may be used.

We are continuously improving the `vantage6` infrastructure. Currently, we are working on facilitating the use of common data models and standards such as OMOP-CDM [5] and FHIR [6] to enable research FAIRification. We are also working on horizontal scaling to ensure that the server can handle workloads more efficiently, as well as developing a graphical user interface, which will make it more accessible and easier to use by the (healthcare) scientific community. We are also working on extending the tools for integrating the input of the `vantage6` community. Besides our main website, we launched a Discourse group, where users can find tutorials, showcase their own projects, ask questions, and even connect with other members. We believe that the community's support will make `vantage6` a better platform for everyone.

In this paper, we explained the changes and improvements of `vantage6`'s latest release. We also presented a few examples of projects and initiatives where these new features have allowed for a wide variety of novel privacy-preserving analysis techniques. We believe that the development of this type of infrastructure is crucial to meet the current and future demands of healthcare research with a strong emphasis on preserving the privacy of sensitive patient data.

References

[1] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. J Healthc Inform Res. 2021;5(1):1-19.

[2] Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Geleijnse, G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. In AMIA Annu Symp Proc; 2020 Nov 14-18; Virtual. Bethesda (MD): AMIA; p. 870-877

[3] Worm D, Kamphorst B, Rooijakkers T, Veugen T, Cellamare M, Geleijnse, G, Knoors D, Martin F. CONVINCED – Enabling privacy-preserving survival analyses using Multi-Party Computation. The Hague, the Netherlands: TNO; 2020. 28 p. Report No.: R11342

[4] van Daalen F, Bermejo I, Ippel L, Dekkers A. Privacy preserving n-party scalar product protocol [Internet]. arXiv [Preprint]. 2021 [cited 2022 Mar 11]: 12 p. Available from: https://arxiv.org/abs/2112.09436

[5] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.

[6] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proc IEEE Int Symp Comput Based Med Syst; 2013 June 20-22; Porto, Portugal. Los Alamitos (CA): IEEE Press; p.326-331.