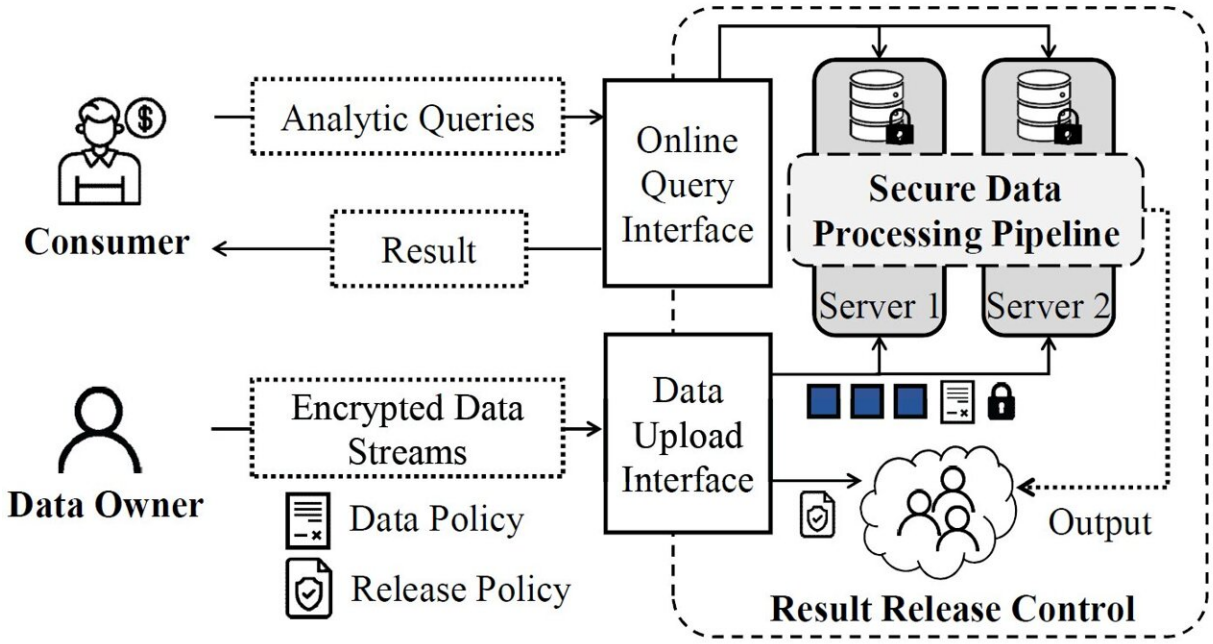
Novel system prevents personal metadata leakage from online behavior for privacy protection

February 6 2023



There are four types of logical parties in Vizard: data owner, data consumer, a secure data processing pipeline, and a result release control committee (RRC). Credit: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022). DOI: 10.1145/3548606.3559349

Privacy preservation is the most challenging issue for data collection. Even if the data is encrypted, metadata, such as users' online behavior,

may lead to identity exposure. A research team from City University of Hong Kong (CityU) recently developed a metadata-hiding analytic system, called Vizard, which enables data owners to securely define their data authorization and control who can use their data, providing potential applications in various sectors, such as precision medical research.

"Imagine if you send a letter to a friend called Alice, and the envelope is sealed so that no one can read the content. But anyone can see that 'you sent Alice a letter' since her address is on the envelope. This is what we called 'side-information', also known as 'metadata' in the virtual world," said Professor Wang Cong, Professor in the Department of Computer Science at CityU.

Examples of very basic metadata for document files are author, date created, date modified and file size. But a wide range of other information, from the frequency of visits to an e-commerce site to a record of participating in a cancer study, is also regarded as metadata.

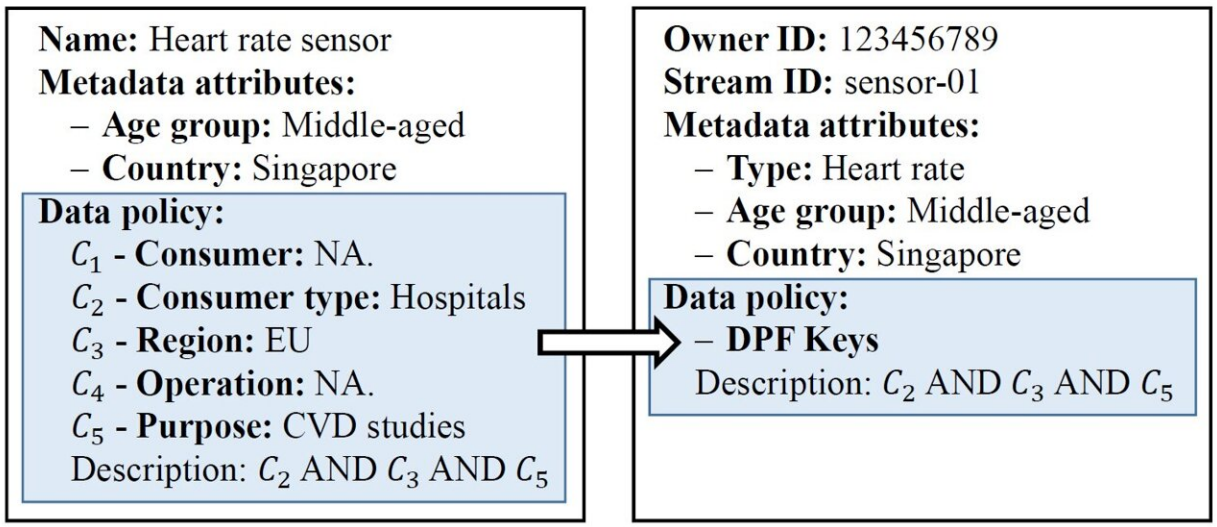
"Third parties may not have access to information on the purchased products or studies, but they will have sufficient metadata almost equivalent to the underlying content," Professor Wang explained. "For example, the person who participated in the cancer study may often visit a health product platform, which may imply that the data owner has cancer or another illness."

Sufficient metadata is almost equivalent to personal tracking

Nowadays, privacy protection relies basically on the practices of data collection platforms, so data owners have no choice but to trust the policies despite the potential risk of data leakage. The research team took on the challenge to remove this "blind" trust with data collection

platforms and big tech companies and developed a novel system, called Vizard, to address metadata leakage concerns.

To design Vizard as a full-fledged metadata-protected data collection and analytical platform, Professor Wang's team utilized a cryptographic tool, called "distributed point function" (DPF). DPF is a generic building block that facilitates secure/encrypted computations, which can be used to anonymously retrieve data during the computation process. Based on DPF, Professor Wang's team developed the Vizard system with stream-specific pre-processing, encryption and throughput enhancement techniques.



Example of data stream descriptions for a heartrate sensor (left) and the secure transformation of its data policies (data access requirements, right). Vizard preserves public metadata attributes (e.g. age group and country in this example) to facilitate grouping and filtering of different data streams. Credit: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022). DOI: 10.1145/3548606.3559349

Second, Vizard is based on an owner-centric control model. Each owner can generate tailored requirements by simply inserting operating keys like "AND", "OR" and "NOT" to control the use of their personal data. For example, owners might want to authorize their data for use only by hospitals in Hong Kong, so the operating keys would be 1) type= hospitals AND 2) region= HK.

The research team demonstrated the efficiency of the novel system. Assuming that Vizard has stored 10,000 owner data ciphertexts and that each owner has specified a data policy that controls which consumers can use their data, it takes only 4.6 seconds for Vizard to handle a data-access query.

Metadata-hiding system enhances data-driven research

This breakthrough design builds on the team's previous work on a practical data analytic system. The system can process encrypted data without decrypting, which is different from existing data processing pipelines and prevents hackers from mining data.

To further protect personal data, a "Result Release Control Committee" (RCC) can be formed by a set of stakeholders, such as data owners and government agencies. Data owners can now jointly set rules related to how the results should be protected before release. For example, they can request correctness verification, privacy protection and reward payments before releasing the results to the data consumers (inquirers). The result release rules will be enforced by the RCC with decentralized trust.

"Our proposed metadata-hiding encrypted data-sharing system can be used in various sectors, like healthcare, business and government, where

big data support is needed for more accurate decision making. For example, hospitals in different regions can securely share their patients' data for more accurate disease diagnosis and precision medicine research," Professor Wang added.

Their findings were presented at the ACM flagship security conference, ACM Conference on Computer and Communications Security (CCS) 2022, under the title "Vizard: A Metadata-hiding Data Analytic System with End-to-End Policy Controls."

More information: Chengjun Cai et al, Vizard, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022). DOI: 10.1145/3548606.3559349

Conference: www.sigsac.org/ccs/CCS2022/

Provided by City University of Hong Kong