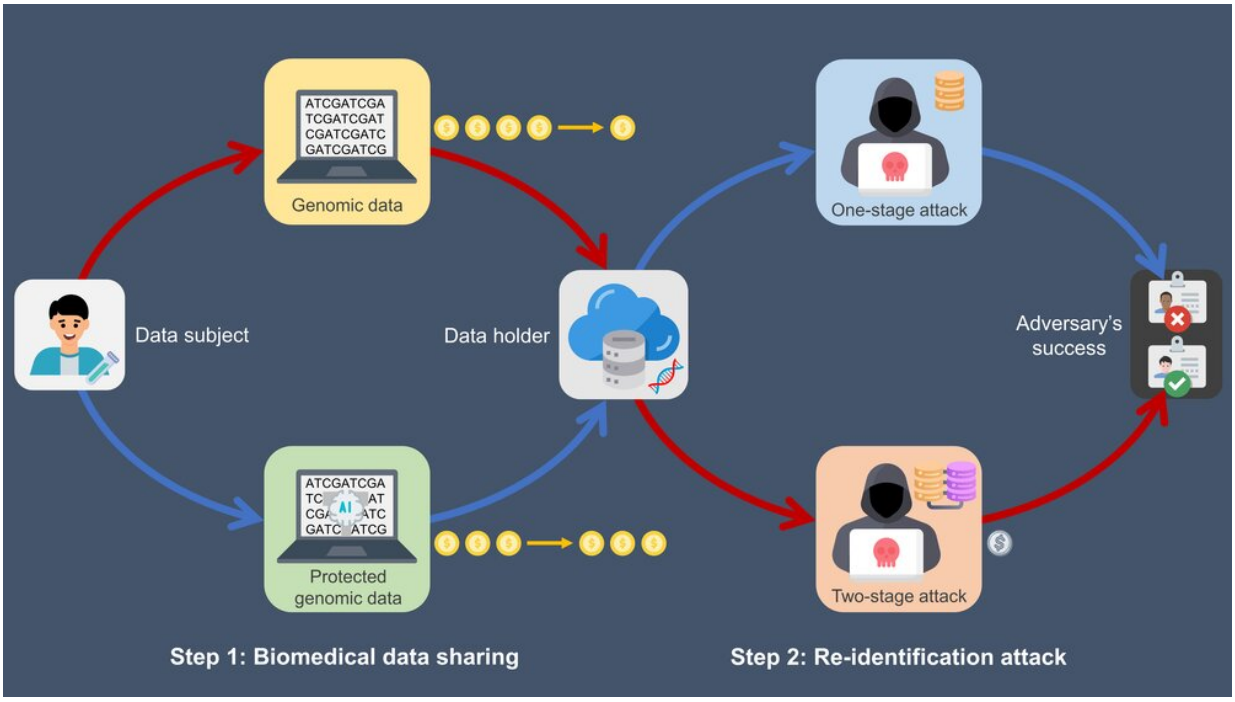
The best way to protect personal biomedical data from hackers could be to treat the problem like a game

December 16 2021, by Zhiyu Wan, Bradley Malin, Yevgeniy Vorobeychik



Poorly protected genomic data attacked by someone with access to multiple data sources (red path) is the most at risk, while better-protected genomic data attacked by someone without access to other sources (blue path) is the least at risk. Credit: Vanderbilt University Medical Center, CC BY-ND 4.0

Game theory, which tries to predict how the behavior of competitors influences the choices the other players make, can help researchers find the best ways to share biomedical data while protecting the anonymity of the people contributing the data from hackers.

Modern biomedical research, such as the National COVID Cohort Collaborative and the Personal Genome Project, requires large amounts of data that are specific to individuals. Making detailed datasets publicly available without violating anyone's privacy is a critical challenge for projects like these.

To do so, many programs that collect and disseminate genomic data obscure personal information in the data that could be exploited to re-identify subjects. Even so, it's possible that residual data could be used to track down personal information from other sources, which could be correlated with the biomedical data to unearth subjects' identities. For example, comparing someone's DNA data with public genealogy databases like Ancestry.com can sometimes yield the person's last name, which can be used along with demographic data to track down the person's identity via online public record search engines like PeopleFinders.

Our research group, the Center for Genetic Privacy and Identity in Community Settings, has developed methods to help assess and mitigate privacy risks in biomedical data sharing. Our methods can be used to protect various types of data, such as personal demographics or genome sequences, from attacks on anonymity.

Our most recent work uses a two-player leader-follower game to model the interactions between a data subject and a potentially malicious data user. In this model, the data subject moves first, deciding what data to share. Then the adversary moves next, deciding whether to attack based on the shared data.

Using game theory to assess approaches for sharing data involves scoring each strategy on both privacy and the value of the shared data. Strategies involve trade-offs between leaving out or obscuring parts of the data to protect identities and keeping the data as useful as possible.

The optimal strategy allows the data subject to share the most data with the least risk. Finding the optimal strategy is challenging, however, because genome sequencing data has many dimensions, which makes it impractical to exhaustively search all possible data sharing strategies.

To overcome this problem, we developed search algorithms that focus attention on a small subset of strategies that are the most likely to contain the optimal strategy. We demonstrated that our method is the most effective considering both the utility of the data to the public and the data subject's privacy.

The worst-case scenario, where an attacker has unlimited capabilities and no aversion to financial losses, is often extremely unlikely. However, data managers sometimes focus on these scenarios, which can lead them to overestimate the risk of re-identification and share substantially less data than they safely could.

The goal of our work is to create a systematic approach to reason about the risks that also accounts for the value of the shared data. Our game-based approach not only provides a more realistic estimate of re-identification risk, but also finds data sharing strategies that can strike the right balance between utility and privacy.

Data managers use cryptographic techniques to protect biomedical data. Other approaches include adding noise to data and hiding partial data.

This work builds on our previous studies, which pioneered using game theory to assess the risk of re-identification within health data and

protect against identity attacks on genomic data. Our current study is the first to consider an attack in which the attacker can access multiple resources and combine them in a stepwise manner.

We are now working to expand our game-based approach to model the uncertainty and rationality of a player. We are also working to account for environments that consist of multiple data providers and multiple types of data recipients.

Provided by The Conversation