A Game Theoretic Model for Privacy-Preserving Genomic Data Sharing

Zhiyu Wan,[1,2,*] Yevgeniy Vorobeychik,[1,3] Weiyi Xia,[1,2] Ellen Wright Clayton,[4,5,6] Murat Kantarcioglu,[7] and Bradley Malin[1,2,8]
[1]Department of Electrical Engineering and Computer Science, Vanderbilt University, [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, [3]Department of Computer Science and Engineering, Washington University in St. Louis, [4]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, [5]Law School, Vanderbilt University, [6]Department of Pediatrics, Vanderbilt University Medical Center, [7]Department of Computer Science, University of Texas at Dallas, [8]Department of Biostatistics, Vanderbilt University Medical Center
*Correspondence: zhiyu.wan@vanderbilt.edu (Z.W.)

ABSTRACT

Emerging scientific endeavors are creating big data repositories from millions of individuals. Sharing data in a privacy-respecting manner could lead to important discoveries, but high-profile demonstrations show that links between de-identified genomic data and named persons can sometimes be reestablished. Such re-identification attacks have focused on worst-case scenarios and spurred the adoption of data sharing practices that unnecessarily impede research. To mitigate concerns, organizations have traditionally relied upon legal deterrents, like data use agreements, and are considering suppressing or adding noise to genomic variants.

In this report, we use a game theoretic lens to develop more effective, quantifiable protections for genomic data sharing. This is a fundamentally different approach because it accounts for adversarial behavior and capabilities and tailors protections to anticipated recipients with reasonable resources.

We demonstrate this approach with a public resource with genomic summary data from over 8000 individuals and show risks can be balanced against utility more effectively than traditional approaches. We further show the generalizability of this framework by applying it to other genomic data collection and sharing endeavors. Recognizing that such models are dependent on a variety of parameters, we perform extensive sensitivity analyses to show that our findings are robust to their fluctuations.

BACKGROUND

Sharing genomic data is for the social good

- Funded investigators by NIH is expected to send de-identified genomic data to NIH Database of Genotype and Phenotype (dbGaP)
- Sharing of genomic data accelerates the discovery of genotype-phenotype associations, especially for rare disease
- Tests based on genomic data assists diagnosis of diseases - that are clinically actionable, and establishment of more effective drug regimens

Big genomic data era

- International HapMap Project (269 individuals) – Started from 2002
- 1000 Genomes Project (2,504 individuals) – Started from 2008
- NIH All of Us Research Program (316,000 individuals enrolled; aims at 1,000,000 subjects) – Started from 2018

Privacy risk of sharing genomic summary statistics

- Sharing allele frequencies (variant of genomic region) about a pool of genomes is useful for research (e.g., GWAS studies)
- Homer's attack demonstrates the risk of sharing allele frequencies in 2008
- NIH stops sharing summary statistics from dbGaP due to the Homer's attack
- More powerful attacks emerge afterwards (Wang's attack, Sankararaman's attack[1])
- Technical countermeasures include SNP suppression, noise addition, etc.
- Legal deterrence includes data use agreement (DUA) and penalty

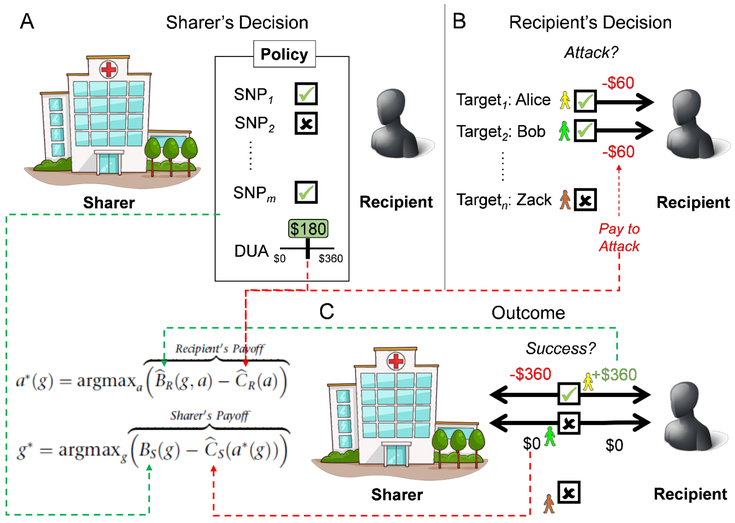
[1] Sankararaman S, et al. *Nature Genetics* 41, 965-967 (2009).

OBJECTIVES

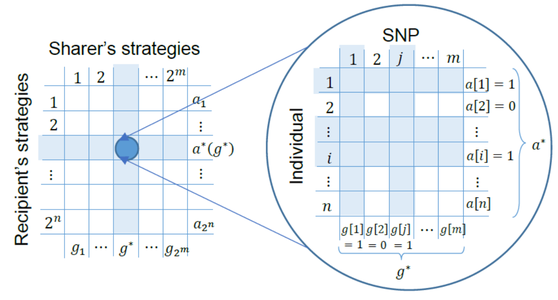
- Risk assessment based on a reasonable adversarial model
- Risk mitigation combining both technical and legal deterrences
- Find the best strategy for the data sharer with a perfect trade-off between sharing utility and privacy

METHODS

- Model the genomic data sharing process as a one-shot Stackelberg (leader-follower) game between the data sharer and the data recipient
- The genomic data sharing process and the game model:



- An illustration of the strategy profile:



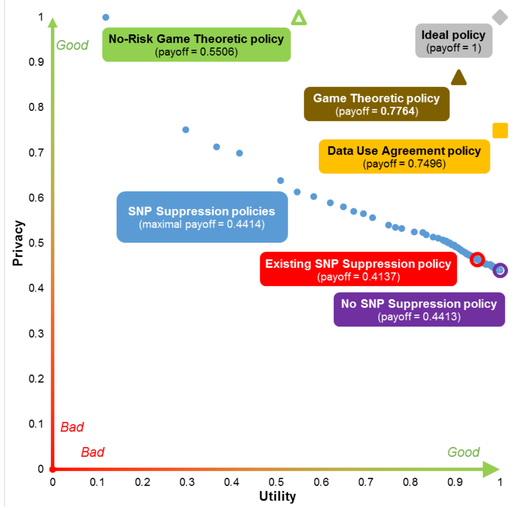
- Search the data sharer's strategy space using genetic algorithm

EXPERIMENTS

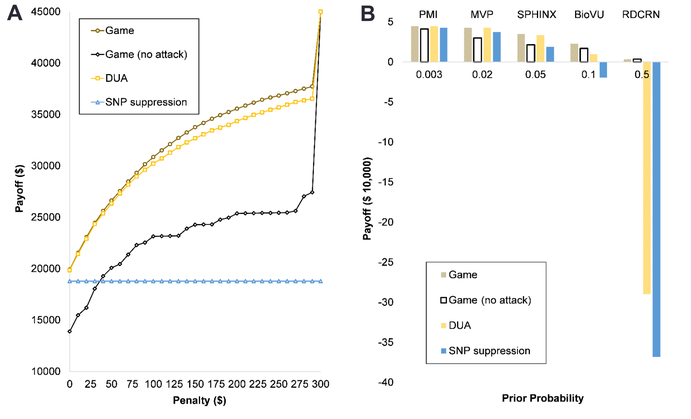
- Dataset
  - 8,194 individuals in SPHINX (https://www.emergesphinx.org/)
  - 2,500 statistically independent SNPs to publish (total of 51,826 SNPs)
  - 2,504 individuals in 1000 Genomes Project
- Valuation settings:
  - $45,000 for grant dollars (or the maximal benefit to the sharer)
  - $360 for the benefit to the attacker for each successfully detected individual
  - $180 for the expected penalty to the attacker per record
  - $60 for the attacker's accessing cost per record

RESULTS

- SPHINX policy Analysis:



- Sensitivity analyses on (A) penalty and (B) prior probability of the detection:



CONCLUSIONS

- The game-theoretical solution achieves the highest payoff for the data sharer.
- The no-attack variation of the game can achieve a payoff higher than the state-of-the-art SNP-suppression strategy while eliminating privacy risk.
- The game-theoretical solution is not sensitive to the changes of parameters such as the penalty and the prior probability.
- Future Directions:
  - Valuation
  - Multiple adversaries
  - Irrational adversaries