COMPUTER SCIENCE

# Using game theory to thwart multistage privacy intrusions when sharing data

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

**Zhiyu Wan[1,2]*, Yevgeniy Vorobeychik[3], Weiyi Xia[2], Yongtai Liu[1], Myrna Wooders[4], Jia Guo[1], Zhijun Yin[1,2], Ellen Wright Clayton[5,6,7], Murat Kantarcioglu[8,9,10], Bradley A. Malin[1,2,11]**

**Person-specific biomedical data are now widely collected, but its sharing raises privacy concerns, specifically about the re-identification of seemingly anonymous records. Formal re-identification risk assessment frameworks can inform decisions about whether and how to share data; current techniques, however, focus on scenarios where the data recipients use only one resource for re-identification purposes. This is a concern because recent attacks show that adversaries can access multiple resources, combining them in a stage-wise manner, to enhance the chance of an attack's success. In this work, we represent a re-identification game using a two-player Stackelberg game of perfect information, which can be applied to assess risk, and suggest an optimal data sharing strategy based on a privacy-utility tradeoff. We report on experiments with large-scale genomic datasets to show that, using game theoretic models accounting for adversarial capabilities to launch multistage attacks, most data can be effectively shared with low re-identification risk.**

## INTRODUCTION

Person-specific biomedical data are now collected on a large scale in a wide range of settings. For instance, in the clinical realm, personal information is routinely stored in electronic health records. The biomedical research domain now supports studies that collect data on a diverse array of participants (*1*). In addition, most recently, the commercial setting has led to a number of ventures where data are collected, such as direct-to-consumer genetic testing companies that collect data from various consumers and build repositories that now cover more than 10% of the U.S. population (*2*). Many believe that sharing these data beyond their initial point of collection is crucial to maximizing the societal value of the data. However, data sharing efforts are often limited by privacy concerns, particularly over the identifiability of data subjects, the individuals to whom the data correspond (*3*).

Genomic data, which are shared in various settings in the United States, provide a clear illustration of the threat of data re-identification and the concern over the possibility. Linking genomic data to explicit identifiers (i.e., re-identification) poses a threat to the anonymity of data subjects. While data managers remove explicit identifiers (e.g., personal names and phone numbers) to adhere to de-identification guidance (*4*–*6*), numerous demonstration attacks have shown that data, and particularly genomic data, can be re-identified through a variety of means (*7*–*10*). Although individuals are incentivized to share their genomic data (*11*–*13*), they usually lack the ability to identify and assess privacy risks properly to make the informed sharing decisions (*12*, *14*).

It is important to recognize that not all re-identification attacks are equally easy to execute and that an oversimplified attack model can lead to an inaccurate measure of risk. Moreover, this inaccuracy is not biased in any particular direction; thus, risk may be underestimated in some cases but overestimated in others. Initially, attacks were based on a single stage (*15*–*17*), where the adversary linked two datasets, one de-identified and one identified, using attributes shared by these datasets (e.g., residual demographics or DNA sequences). More recently, attacks have evolved into multistage forms (*18*–*22*), where each stage reveals another piece of information about a targeted individual.

Here, we introduce the first approach to assessing and strategically mitigating risks by explicitly modeling and quantifying the privacy-utility tradeoff for data subjects in the face of multistage attacks. In doing so, we bridge the gap between more complex models of attack and informed data sharing decisions.

For illustration, we rely on the well-known two-stage attack model of Gymrek *et al.* (*18*), which, to re-identify genomic data, combines surname inference with linkage. Their attack specifically targeted 10 participants in the Center for the Study of Human Polymorphisms (CEPH) family collection, whose genomes were sequenced as part of the 1000 Genomes Project (*23*), by performing surname inference through public genetic genealogy databases made accessible by Ysearch and the Sorenson Molecular Genealogy Foundation (SMGF), now owned by Ancestry.com. In response, in consultation with the local institution managing the CEPH study, the National Institutes of Health moved certain demographics about the participants in the corresponding repository into an access-controlled database (*24*). Although the Ysearch and SMGF databases are no longer accessible to the public (*25*), it is not unreasonable to assume that similar databases may be made publicly accessible in the future.

Various approaches to preventing biomedical data re-identification have been developed from regulatory (*26*–*28*) and technological perspectives (*29*–*36*). However, most of these approaches focus on worst-case scenarios; thus, their impacts on data utility and privacy risk in practice are unclear. For example, the adversary considered

[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37212, USA. [2] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA. [3] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA. [4] Department of Economics, Vanderbilt University, Nashville, TN 37235, USA. [5] Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN 37203, USA. [6] School of Law, Vanderbilt University, Nashville, TN 37203, USA. [7] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN 37232, USA. [8] Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA. [9] Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [10] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA. [11] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.
*Corresponding author. Email: zhiyu.wan@vanderbilt.edu

in these approaches always attacks without taking into account the attack costs (*7*, *8*), which may lead to an overestimation of the privacy risk. In addition, the parameters in technical protection models [e.g., *k*-anonymity (*37*) or differential privacy (*38*)] are usually set without or before measuring their impacts in specific use cases (*39*, *40*), which may either sacrifice too much data utility or provide insufficient protection. To address this problem, risk assessment and mitigation based on game theoretic models have been introduced (*41*–*43*) (see note S1 for a summary of the current literature).

In our work, we show that a game theoretic model can reveal the optimal sharing strategy to data subjects. The model is such that we can conduct experiments involving protection against a multistage attack using either real-world datasets or large-scale simulated datasets. Our results demonstrate that the game theoretic model can efficiently assess and effectively mitigate privacy risks. The fine-grained sharing strategy recommended by our model can minimize the chance that a data subject will be successfully re-identified while maximizing the data utility and keeping the released dataset useful and the data sharing process fair.

## MATERIALS AND METHODS

We investigate a situation in which a data subject chooses how much of the subject's genomic data to share in a public repository, such as the 1000 Genomes Project (*23*), OpenSNP (*11*), or the Personal Genome Project (*44*). For example, the subject may share the entire sequence, a subset of single-nucleotide polymorphisms (SNPs), a subset of short tandem repeats (STRs), or nothing at all. Our goal is to determine the subject's optimal data sharing decision, balancing the monetary benefit (or some benefit that can be translated into monetary terms) of data sharing and the re-identification risk. Re-identification risk, in turn, arises from two sources: first, an adversary, or anyone who has the incentive and means to attempt to determine the identities of subjects in anonymized shared datasets; and, second, other data about the subject that are already available to the public, possibly at some cost. In Fig. 1A, we illustrate this setting in the context of a particular re-identification attack introduced by Gymrek *et al.* (*18*), which we refer to as "the Gymrek attack," in which the adversary first used the shared data together with Ysearch (another public dataset) to infer an individual's surname and subsequently used this additional information to perform a record linkage attack using a third dataset (e.g., PeopleFinders).

We initially assume that the adversary and the data subject have the same beliefs over the probability that an attack will be successful. This allows us to formalize the encounter just described as a Stackelberg, or leader-follower, game (*45*), in which the subject acts as a leader, choosing how much of their genomic data to share, and the adversary is the follower who obtains the shared data and then decides whether to execute an attack (see Fig. 1B). We then take account of the fact that the adversary, since they do not know the values of the masked attributes, can only estimate the probability of the success of an attack, while the data subject has better information.

Note that the primary focus of the encounter is the genomic dataset that results from the explicit data sharing decision by the subject; however, two additional peripheral datasets (e.g., Ysearch and PeopleFinders) play a role in the re-identification attack and are viewed as a part of the overall environment by both the subject and the adversary. Note that neither has any impact on the availability of these datasets.

We model the subject's decision about what portion of the data to share by a vector $\mathbf{s} = < s_1, \cdots, s_j, \cdots, s_m > \in \mathbb{B}^m$ of 0 and 1 values in which $m$ is the number of attributes in the record, $s_j = 0$ if the $j$th attribute is masked, and $s_j = 1$ if the $j$th attribute is shared. Thus, if the subject shares a collection of STRs indexed by $j$, then an entry of 0 in this vector implies that the subject does not share the corresponding STR, whereas an entry of 1 means that this STR is shared. Since the sharing strategy $\mathbf{s}$ of the subject involves masking (outcome-equivalent to redacting) a subset of STRs, we refer to the subject's strategies as masking strategies and the game as a masking game.

The adversary observes the subject's released set of STRs, encoded by their strategy vector $\mathbf{s}$, and decides whether to attack, which we encode as a binary choice $a \in \{0,1\}$, where $a = 1$ means that the adversary decides to attack and $a = 0$ means that they do not attack. If the adversary chooses to attack, then the actual execution of the attack follows a two-stage process, such as the one demonstrated by the Gymrek attack. Specifically, in stage I, the adversary infers the target's surname by comparing the subject's genomic record in database $D$ with genomic records retrieved from database $D_G$. In stage II, the adversary tries to re-identify the subject by linking the subject's genomic record in database $D$ with an identified record in database $D_I$ on a set of demographic attributes (namely, year of birth, gender, and state of residence) and the inferred surname. In general, the attack can have more than two stages, in which each stage infers new information, based on an additional dataset, which can be used in subsequent stages.

A crucial element of our game theoretic model of the interaction between the data subject and the adversary is the sequence of decisions. The subject (the leader) makes a decision first, choosing a masking strategy $\mathbf{s}$. The adversary (the follower) then observes the data and, consequently, what has been masked and chooses whether to attack depending on the masking strategy $\mathbf{s}$; that is, the adversary's strategy is a function of $\mathbf{s}$.

To formalize the consequences of the decisions by both the data subject and the adversary, we now introduce additional notation and assumptions. First, let $b(\mathbf{s})$ capture the monetary benefit to the subject that results from sharing data using a masking strategy $\mathbf{s}$. This benefit function is known to both the subject and the adversary. We assume that the benefit function is nondecreasing in the amount of data shared (see note S2 for a concrete example). Second, if a record is successfully re-identified by the adversary, then we assume that the subject incurs a loss of $L$, which is also the amount gained by the adversary.

The probability that an attack will be successful depends on the masking strategy used by the data subject. Since the data subject knows the values of the masked attributes, the data subject will have a better understanding of the probability of success of an attack, the re-identification risk probability, than the adversary. As noted above, in practice, the adversary can only estimate the likelihood of an attack's success because the adversary, in contrast to the subject, does not know the values of the masked attributes. Initially, to enable us to present the game situation as a standard Stackelberg game, we assume that both the data subject and the adversary assign the same probability of success to each strategy $\mathbf{s}$. Let $p(\mathbf{s})$ denote this probability. Taking $p(\mathbf{s})$ as given, both the data subject and the adversary can maximize their expected payoffs. Then, as further described below, taking expected payoffs as the actual payoffs, we have a Stackelberg game of perfect information and can study its equilibria.

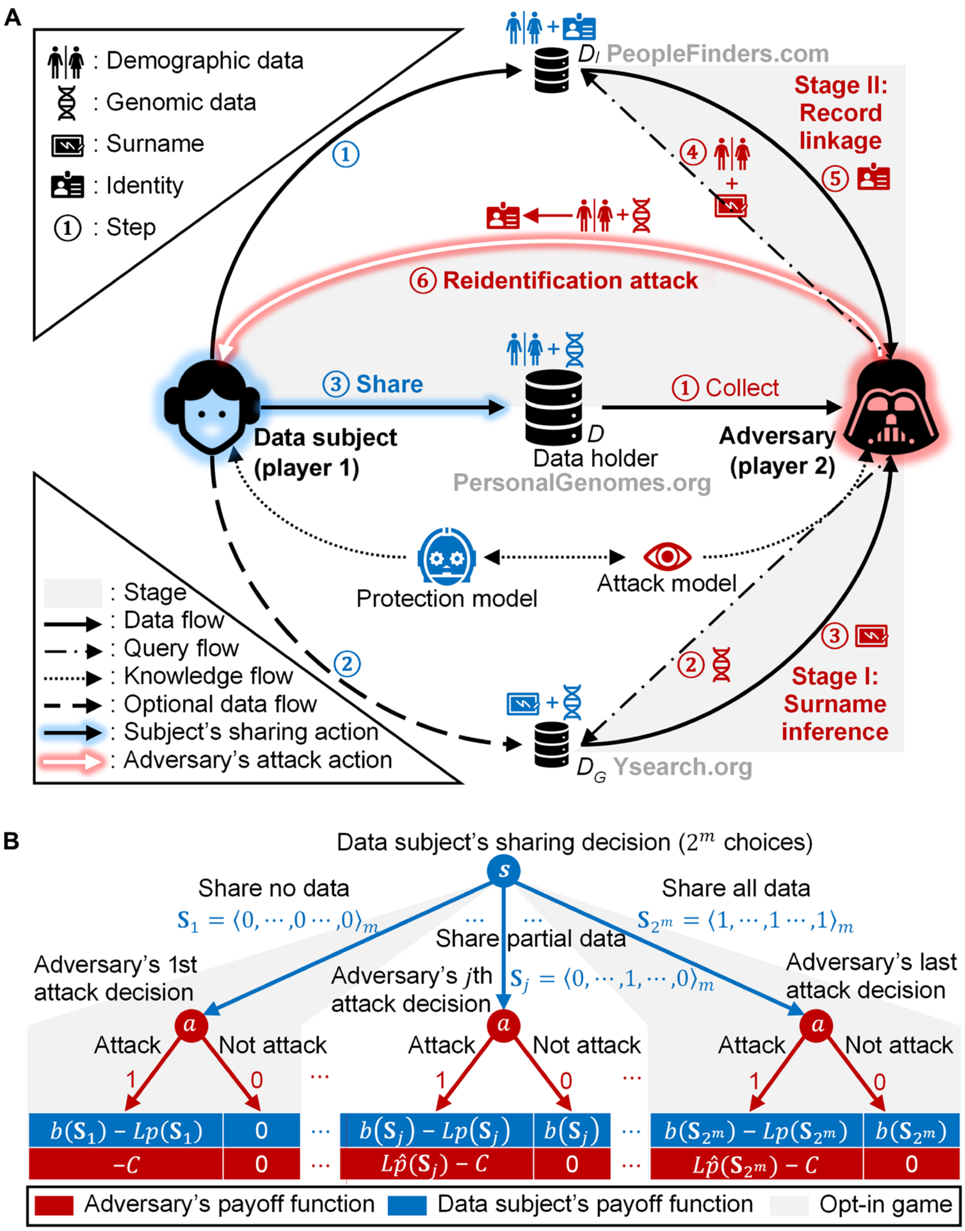
A
: Demographic data
: Genomic data
: Surname
: Identity
① : Step
$D_I$ PeopleFinders.com
Stage II: Record linkage
①
④
+
⑤
⑥ Reidentification attack
③ Share
① Collect
Data subject (player 1)
$D$
Data holder
PersonalGenomes.org
Adversary (player 2)
Protection model
Attack model
②
② ③
Stage I: Surname inference
$D_G$ Ysearch.org
: Stage
: Data flow
: Query flow
: Knowledge flow
: Optional data flow
: Subject's sharing action
: Adversary's attack action
B
Data subject's sharing decision ($2^m$ choices)
$s$
Share no data
$\mathbf{S}_1 = \langle 0, \cdots, 0 \cdots, 0\rangle_m$
Share all data
$\mathbf{S}_{2^m} = \langle 1, \cdots, 1 \cdots, 1\rangle_m$
Share partial data
$\mathbf{S}_j = \langle 0, \cdots, 1, \cdots, 0\rangle_m$
Adversary's 1st attack decision
Adversary's $j$th attack decision
Adversary's last attack decision
$a$
Attack
Not attack
1
0
$b(\mathbf{S}_1) - Lp(\mathbf{S}_1)$
0
$-C$
0
$b(\mathbf{S}_j) - Lp(\mathbf{S}_j)$
$b(\mathbf{S}_j)$
$L\hat{p}(\mathbf{S}_j) - C$
0
$b(\mathbf{S}_{2^m}) - Lp(\mathbf{S}_{2^m})$
$b(\mathbf{S}_{2^m})$
$L\hat{p}(\mathbf{S}_{2^m}) - C$
0
🟥 Adversary's payoff function
🟦 Data subject's payoff function
Opt-in game

**Fig. 1. A multistage privacy attack and its game theoretic protection.** (**A**) A system-wide perspective of a multistage re-identification attack and its protection. Person-specific data records of a subject are accessible to an adversary through three databases: a targeted genomic database ($D$), a genetic genealogy database ($D_G$), and a public identified database ($D_I$). The adversary re-identifies a genomic record by inferring surnames in stage I and linking it to a public record in stage II. The data subject selects a sharing strategy based on a game model only when sharing data in $D$. (**B**) A masking game represented in the extensive form. In a masking game, the data subject moves first, and the adversary moves next. Each terminal node is associated with both players' payoffs. $\mathbf{S}j$ is an $m$-dimensional vector of 0 and 1 values, representing the $j$th concrete action of the data subject. More denotation details are in the main text. The opt-in game is a special variation of the masking game in which the data subject only has those two strategies.

With the notation in hand, we can now define both players' expected payoff (utility) functions. The payoff of the data subject (subscript $d$ for defender) is

$$v_d(\mathbf{s}, a) = b(\mathbf{s}) - Lp(\mathbf{s})\, a \tag{1}$$

The payoff of the adversary is

$$v_a(\mathbf{s}, a) = (Lp(\mathbf{s}) - C)\, a \tag{2}$$

where $C$ is the cost paid by the adversary to execute the attack. Note that the adversary's payoff is 0 if he or she chooses not to attack (i.e., $a = 0$), no matter what the subject does. Note also that we can generalize our model to other data sharing scenarios by redefining the benefit function $b(\mathbf{s})$ and the re-identification risk probability function $p(\mathbf{s})$ as appropriate.

With the specification of strategy sets and terminal payoffs, we have defined a Stackelberg leader-follower game. Our goal is to identify the subject's strategy $\mathbf{s}$ (with some tie-breaking rule if needed) that is a part of a strong Stackelberg equilibrium (SSE; a special case of a subgame perfect Nash equilibrium) (*46*). Let $\phi(\mathbf{s}) = \{a \mid (Lp(\mathbf{s}) - C)a \geq (Lp(\mathbf{s}) - C)(1 - a)\}$; the function $\phi(\mathbf{s})$ is the set of best responses of the adversary to the subject's strategy $\mathbf{s}$. Then, SSE corresponds to the following optimization problem for the subject

$$\max_{\mathbf{s},a\in\phi(\mathbf{s})} b(\mathbf{s}) - Lp(\mathbf{s})\,a \tag{3}$$

To solve the game, the subject can find the best (i.e., SSE) strategy ($\mathbf{s}^*$) using backward induction, where the subject first computes the best responses $\phi(\mathbf{s})$ of the adversary for all possible strategies $\mathbf{s}$, and then choose the best strategy for the subject (e.g., through exhaustive search).

The key practical limitation of the discussion above is that, in practice, since the data subject has their complete dataset (which the adversary does not), they can compute the re-identification probability $p(\mathbf{s})$. This is not, however, the case for the adversary. We capture this asymmetry by allowing the adversary to use an estimated re-identification probability, which we denote by $\hat{p}(\mathbf{s})$, in computing their payoff and thus their best response function. The data subject can also use $\hat{p}(\mathbf{s})$ in computing the best response function of the adversary, but then the data subject can use the true re-identification risk probability in calculating their optimal choice.

More precisely, both the data subject and the adversary can calculate the best response function of the $\phi(\mathbf{s})$ using the re-identification risk probability estimate $\hat{p}(\mathbf{s})$, say $\hat{\phi}(\mathbf{s}) = \{a \mid (L\hat{p}(\mathbf{s}) - C)\,a \geq (L\hat{p}(\mathbf{s}) - C)(1 - a)\}$. Then, the data subject, knowing when the adversary will attack, can calculate their optimal strategy, $\max_{\mathbf{s},a\in\hat{\phi}(\mathbf{s})} b(\mathbf{s}) - Lp(\mathbf{s})\,a$, using the actual probability of success $p(\mathbf{s})$. Note that the special feature of a Stackelberg game that the leader makes only one decision enables us to use the two possibly different re-identification risk probability functions.

We now describe precisely how we compute both the true (for the data subject) and estimated (for the adversary) re-identification probabilities. Specifically, in a two-stage attack, given $\mathbf{s}$, $p_1(\mathbf{s})$ denotes the probability of success of stage I of the attack, $p_2(\mathbf{s})$ denotes the probability of success of stage II of the attack given that stage I succeeds, $\hat{p}_1(\mathbf{s})$ denotes the adversary's estimation on the probability of stage I's success, and $p'(\mathbf{s})$ denotes the probability of stage II's success given that stage I is omitted (i.e., no surname inference performed). Thus, we can represent the true and estimated probabilities of a two-stage re-identification attack's success as shown in Eq. 4

$$\begin{aligned} p(\mathbf{s}) &= p_1(\mathbf{s})\,p_2(\mathbf{s})\,a'(\mathbf{s}) + p'(\mathbf{s})\,(1 - a'(\mathbf{s}))\\ \hat{p}(\mathbf{s}) &= \hat{p}_1(\mathbf{s})\,p_2(\mathbf{s})\,a'(\mathbf{s}) + p'(\mathbf{s})\,(1 - a'(\mathbf{s}))\\ a'(\mathbf{s}) &= \begin{cases} 1, & \hat{p}_1(\mathbf{s})\,p_2(\mathbf{s}) > p'(\mathbf{s}) \\ 0, & \hat{p}_1(\mathbf{s})\,p_2(\mathbf{s}) \leq p'(\mathbf{s}) \end{cases} \end{aligned} \tag{4}$$

where $a'(\mathbf{s})$ is 1 if stage I is better executed and is 0 if stage I is better omitted. The specific settings of $p_1(\mathbf{s})$, $p_2(\mathbf{s})$, $\hat{p}_1(\mathbf{s})$, and $p'(\mathbf{s})$ depend on the attack model and datasets used in an attack.

On the basis of Eqs. 3 and 4, the optimization problem can be represented specifically for a two-stage attack, if $a'(\mathbf{s}) = 1$, as shown in Eq. 5

$$\max_{\mathbf{s},a\in\hat{\phi}(\mathbf{s})} b(\mathbf{s}) - Lp_1(\mathbf{s})\,p_2(\mathbf{s})\,a \tag{5}$$

where $\hat{\phi}(\mathbf{s}) = \{a \mid (L\hat{p}_1(\mathbf{s})\,p_2(\mathbf{s}) - C)\,a \geq (L\hat{p}_1(\mathbf{s})\,p_2(\mathbf{s}) - C)\,(1 - a)\}$. Further details about the derivations of the game model and the ways to set parameters are described in note S3. Note that all parameters can be reasonably set according to a use case. Specifically, they can be tailored to a specific dataset, the attack model, or the valuation provided by a subject. In addition, as our results show, extensive sensitivity and robustness analyses (i.e., stress tests) can help verify the sensitivity and robustness of the parameter settings against the uncertainty in an environment or a subject's knowledge.

We note that in some cases, a subject can only choose from two options: They can either opt in and share all required information or opt out and share nothing. We can view this scenario as a variation of the game we defined above where the strategy of the subject is restricted to two options: share everything ($\mathbf{s}$ = <1, ⋯,1>) and share nothing ($\mathbf{s}$ = <0, ⋯,0>). We refer to this game as the opt-in game. This game's model can be represented as a part of the masking game model, as shown in Fig. 1B (shaded in gray). We can restrict the strategy space for the defender (i.e., the data subject) to any subset of the entire set of strategies $\mathbf{s}$.

Solving a game via exhaustive search may be time consuming, especially when a complex attack model is considered. Algorithms that could be used to accelerate the search process (e.g., the greedy algorithm and the pruning technique) are described in note S4. The implemented game solver with datasets included can be accessed from https://github.com/zhywan/msrigs (archived at https://doi.org/10.5281/zenodo.5543369). The notation used throughout the main text and all supplements is summarized in table S1.

## RESULTS

### Experimental design

To demonstrate our model and evaluate the effectiveness of our methods, we conducted two sets of experiments based on genomic datasets. In one set of experiments, we used real datasets composed of STRs on the Y chromosome (Y-STRs) derived from Craig Venter's genomic record and the Ysearch dataset with 156,761 records and 100 Y-STRs, as they were used by Gymrek *et al.* (*18*). To protect the privacy of the corresponding subjects and enable replications of our investigation, we sanitized (i.e., modified for privacy protection) the original datasets without affecting the demonstration (see note S5 for details about the data sanitization process). In the other set of experiments, to evaluate the effectiveness of our methods in a larger and more controllable environment and to facilitate replications of our investigation without privacy concerns, we simulated a genetic genealogical population of 600,000 individuals (see note S6 for details about the data preparation process), from which multiple datasets were sampled. To further evaluate our methods' effectiveness under various circumstances and uncertainties, we conducted a sensitivity analysis for eight parameters and three experimental settings and conducted a robustness analysis for three parameters. The default values for parameters for the experiments are provided in table S1.

To measure the effectiveness, we calculated the average payoff for a pool of $n$ data subjects, whose records may be shared in genomic dataset $D$, as: $\overline{V} = \sum_{i=1}^{n} V_i$, where $V_i$ represents the $i$th subject's optimal payoff. We further calculated the average data utility ($\overline{U} = \sum_{i=1}^{n} U_i$) and the average privacy ($\overline{P} = \sum_{i=1}^{n} P_i$) of those subjects to show how the game model can balance these two factors. That is, we can calculate effectiveness measures (namely, the average payoff, the average data utility, and the average privacy) after obtaining the optimal payoff, the corresponding data utility, and the corresponding privacy for each subject, given the best strategy ($\mathbf{s}_i^*$), the data record, and parameter settings. For simplicity, we assumed that those subjects use the same parameter settings. More specifically, the $i$th subject's data utility ($U_i$) is defined as the benefit of sharing divided by the maximal benefit of sharing all data as follows: $U_i = b(\mathbf{s}_i^*)/B$. In addition, the $i$th subject's privacy ($P_i$) is defined as one minus the privacy risk (i.e., the probability to be successfully attacked) as $P_i = 1 - p(\mathbf{s}_i^*)\,a_i^*$, in which $a_i^* \in \phi(\mathbf{s}_i^*)$ is the adversary's best response, and $p(\mathbf{s}_i^*)$ is the probability of success of an attack given $\mathbf{s}_i^*$. Notably, the $i$th subject's optimal payoff ($V_i$) can be represented as a linear combination of the corresponding data utility and privacy: $V_i = v_d(\mathbf{s}_i^*, a_i^*) = B\,U_i - L(1 - P_i) = B\,U_i + L\,P_i - L$, in which $L$ is the loss from being re-identified. As the primary measure of effectiveness, the average payoff of those subjects is positively correlated with the metrics for utility and privacy.

Our simulated population was generated with 20 attributes, including ID, surname, year of birth, U.S. state of residence, and 16 genomic attributes. On the basis of the simulated population, we ran experiments to find the best sharing strategy for each subject in a targeted dataset in four game scenarios and four baseline scenarios. In each run of the experiments, from the simulated population, we randomly selected 1000 records for the targeted genomic dataset $D$, 20,000 records for the identified dataset $D_I$, and 20,000 records for the genetic genealogy dataset $D_G$. On the basis of these datasets, we compared eight scenarios: (i) no protection, (ii) demographics only, (iii) random opt-in, (iv) random masking, (v) opt-in game, (vi) masking game, (vii) no-attack masking game, and (viii) one-stage masking game. In all scenarios, the adversary aims to re-identify all records in dataset $D$, but he or she makes rational decisions according to their estimated payoff.

The following provides a brief description of how the recommended strategy is selected in each scenario. In the no-protection scenario, each subject in dataset $D$ shares all attributes in their data record. In the demographics-only scenario, each subject in dataset $D$ shares demographic attributes only. In the random opt-in scenario, each subject in dataset $D$ randomly decides to share the entire data record with a certain probability according to a practical setting. In the random masking scenario, each subject in dataset $D$ randomly decides to share each attribute in their data record with a fixed probability. In game scenarios, each subject in dataset $D$ makes rational decisions by playing the masking game or one of its variations. In the masking game, the subject can mask a portion of data before sharing. However, in the opt-in game, the subject can only decide to opt in to share the entire data record or to opt out. In the no-attack masking game, we assume that the subject chooses no strategy that will make the adversary attack. Whereas in the one-stage masking game, we assume that dataset $D_G$ is not available to the adversary, and thus, the attack has only one stage. Further details about the datasets used in the experiments are summarized in Table 1, while the detailed settings associated with each scenario are provided in the final section of note S3.

From the perspective of a data holder (e.g., the controller of a public data repository) who cares about the data quality and the fairness of the collected dataset from a pool of subjects, we defined two types of measures (namely, usefulness and fairness) to evaluate the downstream effects of the adopted data protection/sharing strategy in each scenario. The usefulness of a data sharing solution is based on the distance between the distributions of the shared and the unprotected data (see note S7 for details). The fairness of a data sharing solution (e.g., fairness with respect to usefulness or fairness with respect to privacy) is based on the Gini coefficient of the specific measures corresponding to each demographic group (see note S8 for details).

## Experiments based on a large-scale simulated population

We ran the experiment 100 times using the backward induction algorithm with pruning and depicted the results in Fig. 2. Figure 2A shows a violin plot of the distributions of the data subjects' average payoffs in all eight scenarios, and Fig. 2B shows a scatterplot of the data subjects' average privacy and utility in each scenario.

Several observations are worth highlighting. First, in Fig. 2A, it can be seen that the subjects' average payoff is lowest in the no-protection scenario and highest in the masking game. Second, the subjects' average payoff is improved substantially in the masking game, compared to that in the opt-in game. This observation illustrates one of the essential advantages of providing some degree of granular choice in the data sharing process. Third, the masking game works better when the adversary uses fewer data resources and, thus, keeps fewer stages in the attack. Last, a universal strategy, whether it is sharing all data or sharing demographics only, or a randomized strategy brings a negative or negligible average payoff to the subjects.

In Fig. 2B, the results representing the opt-in game are all in the plot's top left corner, which implies that the subjects' strategies in

**Table 1. A summary of the datasets used in the experiments.**

| | | Set of experiments | |
|---|---|---|---|
| | | **Large-scale evaluation** | **Real-world demonstration** |
| **Targeted genomic dataset (*D*)** | **Dataset** | Simulated genomic dataset | Craig Venter's data |
| | **Attributes** | Year of birth, U.S. state, 16 STRs | Year of birth, U.S. state, 50 STRs |
| | **Records** | 1000 | 1 |
| **Genetic genealogy dataset ($D_G$)** | **Dataset** | Simulated genetic genealogy dataset | Ysearch |
| | **Attributes** | Surname, 16 STRs | Surname, 50 STRs |
| | **Records** | 20,000 | 58,218 |
| **Public identified dataset ($D_I$)** | **Dataset** | Simulated demographic dataset | PeopleFinders |
| | **Attributes** | ID, name, year of birth, U.S. state | Name, age, U.S. state |
| | **Records** | 20,000 | ~250 million |

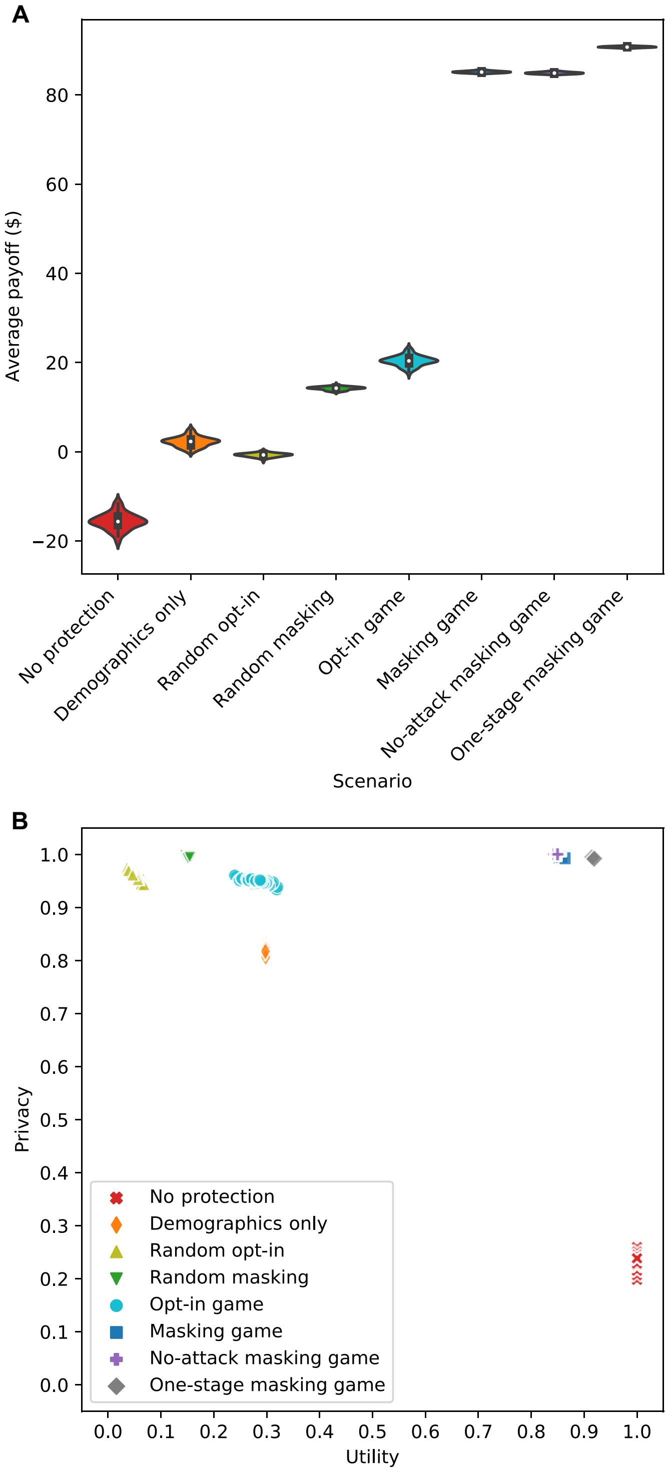

**Fig. 2. Effectiveness measures of the protection in eight scenarios against a multistage re-identification attack targeting data from 1000 subjects.** (**A**) Violin plot of eight distributions of the data subjects' average payoffs, where each distribution corresponds to one scenario. The violin plot (depicted using Seaborn) combines boxplot and kernel density estimate for showing the distribution of data subjects' payoffs in each scenario. A Gaussian kernel is used with default parameter settings. (**B**) Scatterplot of data subjects' average privacy metrics and average utility metrics, where each mark corresponds to one scenario and 1 run (of 100 runs).

this scenario tend to achieve high privacy but low utility. By contrast, the results representing the no-protection scenario are all in the plot's lower right corner, which implies that the subjects' strategies in this scenario tend to achieve high utility but low privacy. Only the results representing the masking game (and two of its variations) are in the plot's top right corner, where the subjects' strategies achieve relatively high utility and high privacy at the same time. Notably, the subjects' strategies in the no-attack masking game guarantee full privacy protection with a substantial amount of shared data. In addition, a slightly higher level of data utility is achieved when the attack has only one stage. The subjects' strategies in the remainder of the scenarios, however, are all worse than those in game scenarios. Specifically, compared to what the masking game does, the random opt-in scenario and the random masking scenario bring the subjects a similar privacy level but a much lower level of data utility. By contrast, compared to the opt-in game, the demographics-only scenario brings the subjects a similar utility level but a lower privacy level. However, compared to the outcome for the no-protection scenario, the demographics-only scenario always provides the subjects with much higher privacy level, which highlights the power of the surname inference stage in the Gymrek attack. Notably, with the game theoretic protection, the power of the surname inference stage can be reduced to the minimum, as shown by the difference between the masking game and its one-stage variation.

The effectiveness measures in the first run of the experiments are reported in Table 2 with additional statistics (i.e., SDs) of payoffs, data utilities, and privacy measures for the group of subjects. It can be seen that the subjects obtain a higher average payoff and higher average privacy when they make decisions based on game models. In addition, the subjects' average payoff in the masking game is three times greater than the one in the opt-in game. The masking game also achieves higher average utility and higher average privacy than the opt-in game does for the subjects. More specifically, about 30% of data are shared, and about 6% of the subjects are expected to be re-identified in the opt-in game. By contrast, about 86% of data are shared, and fewer than 1% of the subjects are expected to be re-identified in the masking game. Compared to all baseline scenarios and the opt-in game, the masking game achieves higher average payoff and average privacy with lower SDs for the subjects. By contrast, although the opt-in game generates a positive average payoff, a few subjects' payoffs are still negative because the corresponding SD is higher than the average. All baseline scenarios provide the subjects with a negative or relatively low average payoff. Among all scenarios, the no-protection scenario exhibits the highest SD (or variation) among subjects in terms of payoff and privacy, while the opt-in game exhibits the highest SD (or variation) among subjects in terms of data utility.

Table 2 further demonstrates the additional stage's contribution to the two-stage re-identification attack in terms of the chance of success. Note that in the demographics-only scenario, the adversary can only perform a one-stage attack. It can be seen that, in the no-protection scenario, about 76% of the subjects are expected to be re-identified, whereas, in the demographics-only scenario, about 19% of the subjects are expected to be re-identified. That is, the additional stage raised the success rate of the attack by a factor of 4. Nevertheless, such a high re-identification rate implies that most targeted subjects are uniquely identifiable from the identified dataset $D_I$. The additional stage's contribution is further demonstrated by comparing two masking games with a different number of attack stages. It can be

**Table 2. Effectiveness measures of protection scenarios against a multistage re-identification attack in the first run of the experiments.** Scenarios include (1) no protection, (2) demographics only, (3) random opt-in, (4) random masking, (5) opt-in game, (6) masking game, (7) no-attack masking game, and (8) one-stage masking game.

| Notation | Description | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Baseline | | | | Game | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $\overline{V}$ | Average payoff of data subjects | −$13.31 | −$1.92 | −$0.62 | $13.83 | $21.64 | $85.22 | $84.94 | $90.63 |
| $\sigma_V$ | SD of data subjects' payoffs | $57.65 | $33.38 | $11.96 | $12.50 | $35.85 | $7.23 | $7.22 | $5.78 |
| $\overline{U}$ | Average data utility of subjects | 1 | 0.298 | 0.043 | 0.1472 | 0.301 | 0.8599 | 0.8494 | 0.9153 |
| $\sigma_U$ | SD of subjects' data utility | 0 | 0 | 0.2029 | 0.1032 | 0.4587 | 0.0746 | 0.0722 | 0.0623 |
| $\overline{P}$ | Average privacy of data subjects | 0.2446 | 0.8141 | 0.9672 | 0.9941 | 0.9436 | 0.9948 | 1 | 0.994 |
| $\sigma_P$ | SD of data subjects' privacy | 0.3844 | 0.2225 | 0.173 | 0.0637 | 0.1271 | 0.0194 | 0 | 0.0209 |

observed that the game theoretic protection against the one-stage attack brings better averages and SDs with respect to both payoff and utility. Notably, the additional stage induces the subjects to yield only about 6% of their data utility to secure a substantially lower privacy risk, 13% fewer re-identified subjects.

Figure 3 shows the best strategies for the first 700 data subjects in the first run of the experiments in the masking scenarios. In the masking game, only a small portion of data is masked for most subjects. Notably, the no-attack masking game brings full privacy protection, and the one-stage masking game brings better average payoff, to those subjects. By contrast, in the random masking scenario, the subjects' average utility loss is substantially higher, although the average privacy risk is almost the same as those in the game scenarios.

Table 3 summarizes the computational efficiency of our implementation for each scenario, in terms of average running time across 100 runs. We solved each masking game twice using the backward induction algorithm and the greedy algorithm with pruning. The efficiency of computing the resulting strategies for all 1000 data subjects in each scenario is evaluated on a machine with a six-core 64-bit central processing unit clocked at 4.19 GHz and a 32-gigabyte random-access memory clocked at 2400 MHz. It can be seen that the implementation of the opt-in game scenario runs the fastest, with a speed that is almost the same as the fastest baseline scenario runs. In addition, the masking game solved using the greedy algorithm with pruning runs at a rate that is more than two orders of magnitude faster than the masking game solved using the backward induction algorithm with pruning while achieving approximately the same results (see note S4 for details). However, it runs about one order of magnitude slower than all of the baseline scenarios and the opt-in game scenario.

According to the results of the empirical evaluation in figs. S1 and S2, the masking game can achieve 7% higher usefulness and 22% higher fairness with respect to usefulness than the random masking approach does while keeping the privacy and the fairness with respect to privacy at a similar level. Neither can the other baseline scenario based on the *k*-anonymity protection model (*37*) outperform the masking game in terms of these two types of measures in the experiments (see notes S7 and S8 for details).

## Sensitivity and robustness analyses on parameters and settings based on simulated datasets

To test the model's sensitivity to eight parameters and three experimental settings, we compared effectiveness measures in eight scenarios across 11 sets of experiments. In each set of experiments, we changed one parameter or experimental setting and ran 20 times with different sample datasets. The results of the sensitivity analysis on eight parameters and three settings regarding the data subjects' average payoff are shown in Fig. 4.

In general, regardless of how the targeted parameter or setting varies, the subjects' average payoff in the masking game is much higher than their average payoff in other scenarios (except two variations of the masking game). An even higher average payoff is achieved in the one-stage masking game. In addition, almost the same payoff is achieved with a guarantee of full privacy protection in the no-attack masking game.

In addition, in the masking game, the subjects' average payoff is not sensitive to most parameters except the maximal benefit of sharing all data. It is attributable to the robustness of the masking game that, for most parameters, there exists at least one scenario in which the range of payoff change is more than six times of the one in the masking game. However, the sensitivity of the payoff to the maximal benefit of sharing all data is unavoidable because the payoff and the benefit are linearly and positively correlated regardless of the scenario.

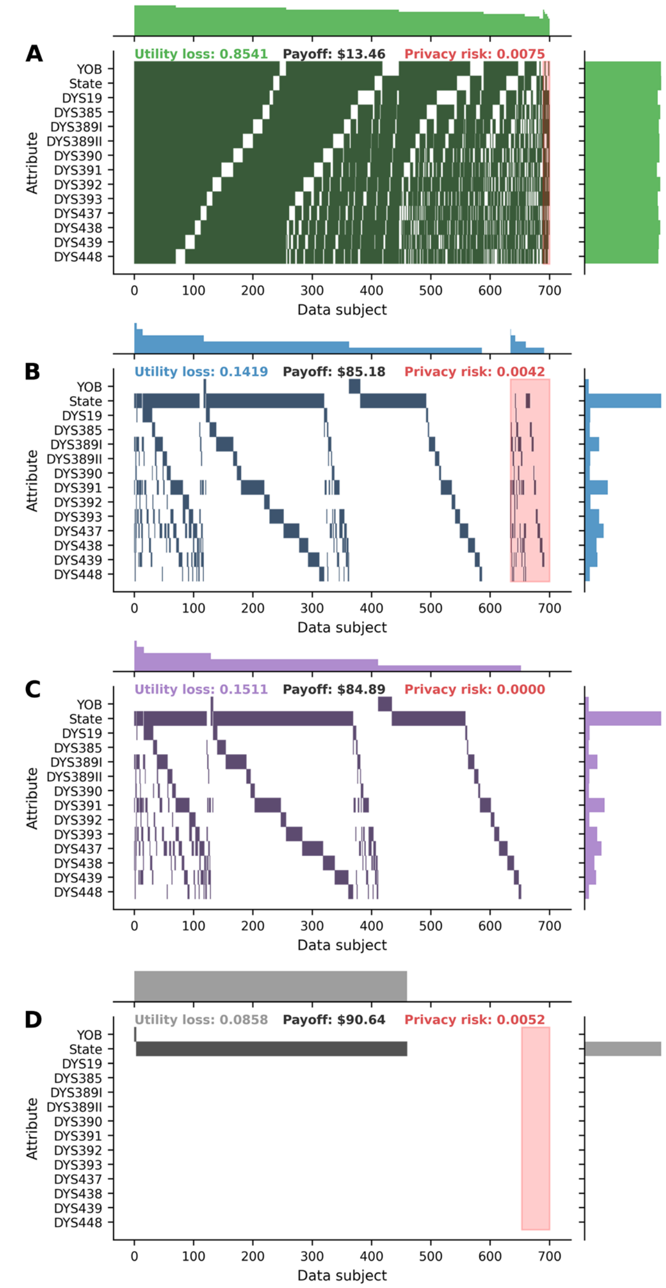

**Fig. 3. Best strategies for the first 700 data subjects in the first run of experiments in the random masking scenario and three masking game scenarios.** (**A**) Random masking scenario. (**B**) Masking game. (**C**) No-attack masking game. (**D**) One-stage masking game. Each nonwhite block indicates that a data subject masks a specific attribute. Each row represents an attribute, and each column represents a data subject. The distribution for data subjects (attributes) is summarized in the histogram on the top (on the right) with the number of bins equal to the number of data subjects (attributes). Data subjects are split into two groups: those on the left that will not be attacked and those on the right that will be attacked (shaded in red). Columns (or data subjects) within each group are sorted by the number of masked attributes in descending order. Rows (or attributes) are sorted by the order of attributes in the dataset. For each scenario, the average payoff, utility loss, and privacy risk are presented at the top center, top left corner, and top right corner, respectively. For each data subject, utility loss is defined as one minus data utility, and privacy risk is defined as one minus privacy. YOB, year of birth; DYS, DNA Y-chromosome segment.

Moreover, in most scenarios, the subjects' average payoff decreases when any one of three parameters (i.e., the number of genomic attributes, the number of records in the genetic genealogy dataset, and the loss from being re-identified) increases. By contrast, the subjects' average payoff increases as any one of five parameters (i.e., the proportion of missing genomic data, the threshold for confidence score, the number of records in the identified dataset, the maximal benefit of sharing all data, and the cost of an attack) increases in most scenarios. These trends are reasonable because, for instance, as the proportion of missing genomic data increases, less information from the genetic genealogy dataset can be used, and thus, the attack becomes less successful. By contrast, as the number of records in the genetic genealogy dataset increases, the adversary is more likely to find someone sharing similar genomes with a targeted subject, and thus, the attack becomes more successful.

By changing each one of three important experimental settings (or assumptions) from what we set in default, we found that, in game scenarios, the subjects would earn higher average payoff if (i) the attributes' weights are set further away from their information entropies, (ii) the subjects in a dataset are not required to adopt the same strategy, or (iii) an off-the-shelf machine learning approach is used for surname inference in the attack. These trends indicate that a game model's effectiveness has a potential to be improved further if some experimental setting is changed from what we set in default. The sensitivity analysis on parameters and settings regarding payoff is described in further detail in note S9.

The results of the sensitivity analysis on eight parameters regarding the data subjects' average privacy and average data utility are shown in Figs. 5 and 6, respectively. In general, the subjects' average privacy in the masking game is always much higher than their average privacy in baseline scenarios save the random opt-in and random masking scenarios. Likewise, the subjects' average data utility in the masking game is only lower than their average data utility in the no-protection scenario or in the one-stage masking game. The sensitivity analysis on parameters regarding privacy and utility is described in further detail in note S9.

To investigate how the uncertainty in the data subject's knowledge about the adversary's capabilities affects the data subject's expected payoff, we conducted a robustness analysis on uncertainty in three parameters: (i) the cost of an attack, (ii) the number of records in the genetic genealogy dataset, and (iii) the number of records in the identified dataset. The results are presented in fig. S3. In general, regardless of how the actual value of an examined parameter changes, in most cases, the error in the subjects' average expected payoff in the masking game is bounded in a range that is much smaller than the difference between the subjects' average payoff in the masking game and their average payoff in the opt-in game. In addition, the results indicate that the superiority of the masking game in terms of the data subjects' average payoff is highly robust regardless of how a data subject is uncertain about each of these examined parameters and the corresponding probability of an attack's success. The robustness analysis on uncertainty in parameters is described in further detail in note S10.

Last, we performed a sensitivity analysis on a special parameter, minority-support factor, which is an exponent in a benefit function considering both usefulness and fairness with respect to usefulness, indicating the extent to which the benefit function supports the minority groups in the shared dataset. In general, the minority groups receive exponentially higher benefits than the majority groups when

**Table 3. Computational efficiency of protection scenarios against a multistage re-identification attack averaged across 100 runs of the experiments.** Scenarios include (1) no protection, (2) demographics only, (3) random opt-in, (4) random masking, (5) opt-in game, (6) masking game, (7) no-attack masking game, and (8) one-stage masking game. BIAP, backward induction algorithm with pruning; GAP, greedy algorithm with pruning.

| | Scenario | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | | | | | Game | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | | 7 | | 8 | |
| | | | | | | BIAP | GAP | BIAP | GAP | BIAP | GAP |
| **Running time (seconds)** | 7.00 | 6.96 | 7.01 | 9.87 | 6.93 | 7935.28 | 93.63 | 7286.98 | 84.39 | 7.88 | 66.13 |

the parameter is positive and vice versa when the parameter is negative (see note S2 for details). The results, as presented in fig. S4, show that the masking game outperforms other protection approaches in terms of usefulness when the parameter is in the [0, 4] range and that the masking game outperforms other protection approaches in terms of fairness with respect to usefulness when the parameter is in the [0.5, 1.5] range (see note S11 for details).

### Experiments based on Craig Venter's data and the Ysearch dataset

In this set of experiments, we used a case study to illustrate how our model could be applied to real-world datasets. Specifically, we used Craig Venter's demographic attributes (including year of birth, state of residence, and gender) and 50 Y-STRs profiled from his genome sequence as the targeted dataset $D$. In addition, we used the Ysearch dataset as the genetic genealogy dataset $D_G$. After filtering out records with too few targeted Y-STRs, we got a dataset of 58,218 records and 50 Y-STRs with a missing proportion of approximately 26%. Last, we used Intelius.com and the 2010 U.S. Census as sources of the identified dataset $D_I$.

A query of Craig Venter's demographic attributes and surname on Intelius.com returned with two records, one of which is Craig Venter. According to the 2010 U.S. Census, 157,681 people were estimated to share the same values on the three demographic attributes with Craig Venter in the United States in 2018. This number indicates that, without the correctly inferred surname, the re-identification would be unlikely to be successful. For simplicity, we masked only genomic attributes instead of all attributes in this demonstrational case study. The parameters in both sets of experiments were set in the same way. For example, the data utility corresponding to each genomic attribute in this set of experiments was set according to its information entropy in the Ysearch dataset. Settings for all other parameters are reported in table S1. We ran the experiments using the greedy algorithm with pruning, the results of which are provided in Table 4. In addition to showing the effectiveness measures of the recommended strategies from eight scenarios, it shows effectiveness measures of all searched strategies in the masking game (with the best strategy being marked out).

From Table 4, it can be seen that the best strategy in the masking game is superior to all other strategies searched in the game, as well as all strategies recommended from all baseline scenarios and the opt-in game in terms of the resulting payoff. The underlying reason is that the first strategy during the search process with a confidence score below the threshold of 0.5 has an incorrectly inferred surname, which leads to an unsuccessful re-identification and thus successful protection.

All strategies searched in the masking game using the greedy algorithm with pruning are plotted in fig. S5, showing the relationships between three evaluation metrics (i.e., utility, payoff, and inference correctness) and the confidence score for each strategy. The optimality of a strategy depends on the correctness of the surname inference (i.e., the success of stage I of the attack), which is correlated with the threshold for confidence score in the attack model (see note S12 for details).

## DISCUSSION

The methodology described in this study enables subjects to make informed data sharing decisions in the face of complex state-of-the-art re-identification models. It enables people to answer questions such as, "Should I share my de-identified data record to an open data repository?" and "Which portion of my data record should I share to an open data repository?" Moreover, the methodology is sufficiently flexible to enable subjects to make decisions in platforms where sharing partial or modified data is allowed.

Our illustration of this methodology in the context of a known multistage attack on genomic data led to several notable findings. First, although an additional stage can substantially increase the accuracy of the re-identification when there is no protection, it makes the attack more vulnerable to our game theoretic protection because the adversary could be tricked into inferring wrong intermediate information, thus mitigating the privacy risk. Second, most subjects (acting rationally) would choose not to share data to an open data repository (e.g., the Personal Genome Project) if partial data sharing is not permitted. By contrast, most people would share most of their data if sharing partial data is permitted. This finding is intriguing because it suggests that providing subjects with options could encourage a greater degree of data sharing while avoiding re-identification. Third, subjects can choose strategies that allow for sharing a substantial amount of data with a payoff almost as high as the optimal solution, while ensuring that it is never beneficial for the data recipient to attempt re-identification, thus predicting no attack and zero risk within the context of our modeling framework. Last, our extensive sensitivity analysis illustrates how parameters of our model influence a subject's behavior differently, which can provide guidance to other stakeholders. For example, to effectively promote data sharing in general, policy-makers could increase penalties for detected privacy breaches, and data holders could increase rewards for data sharing.

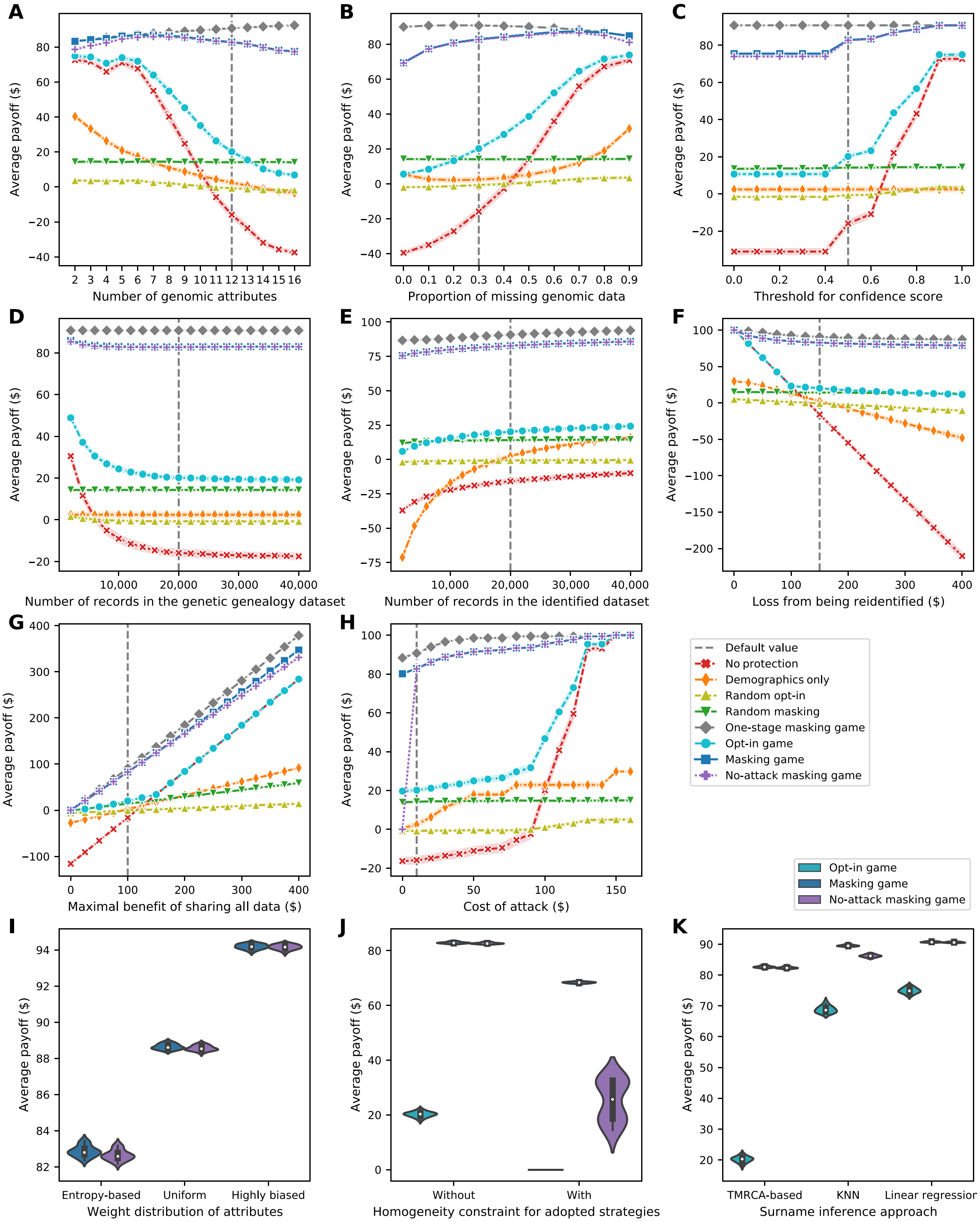

**Fig. 4. Sensitivity of the data subjects' average payoff as a function of the parameters and settings in the model.** (**A**) Line plot of the sensitivity to the number of genomic attributes. (**B**) Line plot of the sensitivity to the proportion of missing genomic data. (**C**) Line plot of the sensitivity to the threshold for confidence score. (**D**) Line plot of the sensitivity to the number of records in the genetic genealogy dataset. (**E**) Line plot of the sensitivity to the number of records in the identified dataset. (**F**) Line plot of the sensitivity to the loss from being re-identified. (**G**) Line plot of the sensitivity to the maximal benefit of sharing all data. (**H**) Line plot of the sensitivity to the cost of an attack. (**I**) Violin plot of payoff distribution's sensitivity to the strategy adoption setting. (**J**) Violin plot of the sensitivity to the surname inference approach. (**K**) Violin plot of the sensitivity to the weight distribution of attributes. Each line plot (depicted using Seaborn) shows data subjects' average payoffs, with error bars representing SDs, in eight scenarios. Each violin plot (depicted using Seaborn) combines boxplot and kernel density estimate for showing the distributions of data subjects' average payoffs in several scenarios. Gaussian kernels are used with default parameter settings. TMRCA, time to most recent common ancestor; KNN, *k*-nearest neighbors.

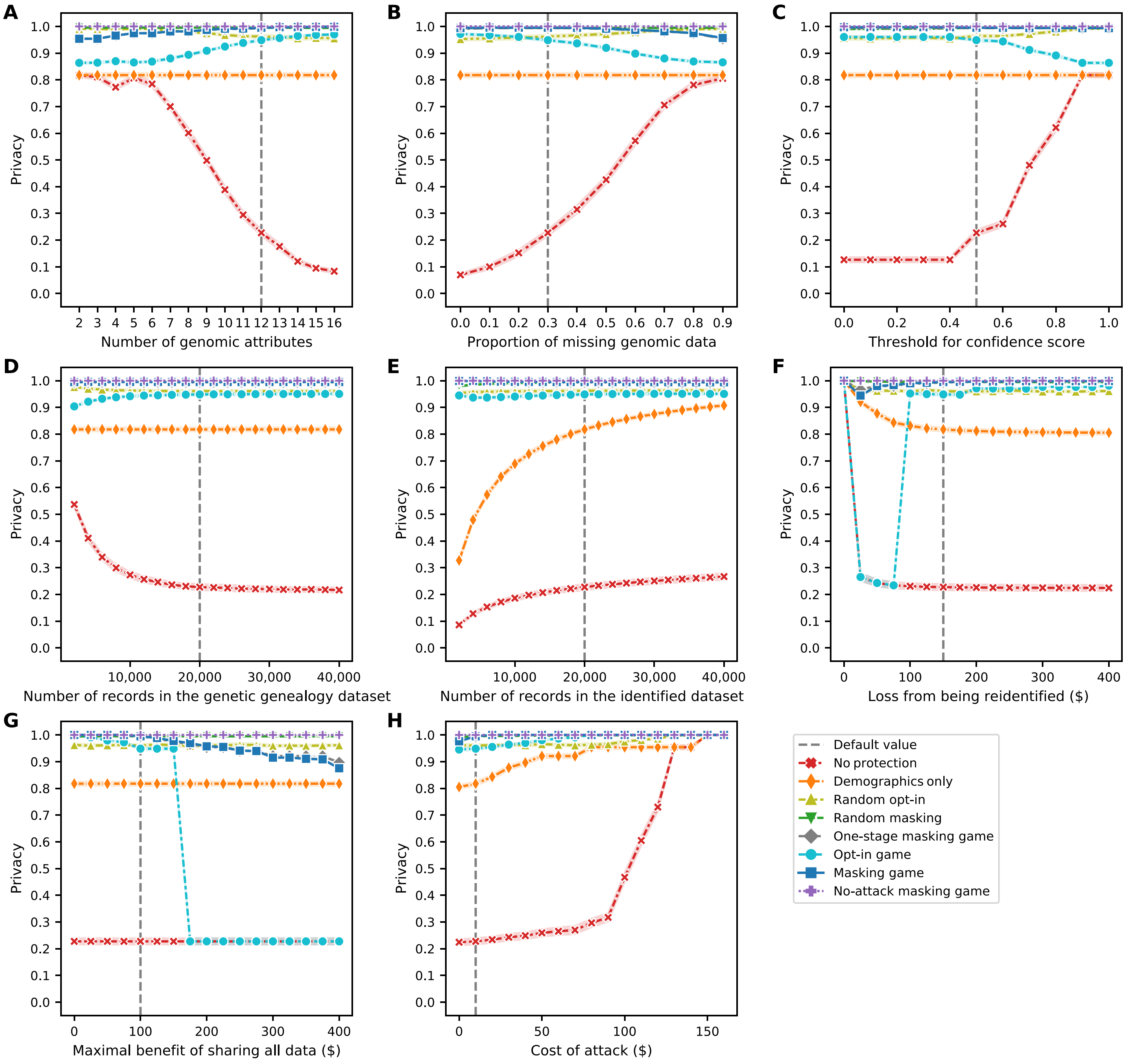

**Fig. 5. Sensitivity of the data subjects' average privacy as a function of the parameters in the model.** (**A**) Number of genomic attributes. (**B**) Proportion of missing genomic data. (**C**) Threshold for confidence score. (**D**) Number of records in the genetic genealogy dataset. (**E**) Number of records in the identified dataset. (**F**) Loss from being re-identified. (**G**) Maximal benefit of sharing all data. (**H**) Cost of an attack. Each line plot (depicted using Seaborn) shows the data subjects' average payoffs, with error bars representing SDs, in eight scenarios.

In addition, the analysis shows what weight a data subject should give to each parameter. Specifically, considering the sensitivity in the masking game, a subject should take extra care when the maximal benefit of sharing all data or the cost of an attack is low. The analysis demonstrates the robustness of the methodology that, while an adversary can push most parameters such as damages of attacks and sizes of datasets to a risky point for subjects in poorly protected scenarios, the highly effective and stable protection provided by the masking game is almost immune to these risks.

Limitations exist in our current model, which provide directions for our future work. First, we only considered one adversary and performed experiments with two-stage attacks. In the future, we will consider game theoretic models with multiple adversaries and conduct experiments with attacks that have more than two stages. For instance, in each stage, the adversary could infer a set of attributes using an external dataset or launch a linkage attack, as illustrated by attacks with far more than two stages (*20*, *47*). This is a challenging task because the chain of attack could expand over

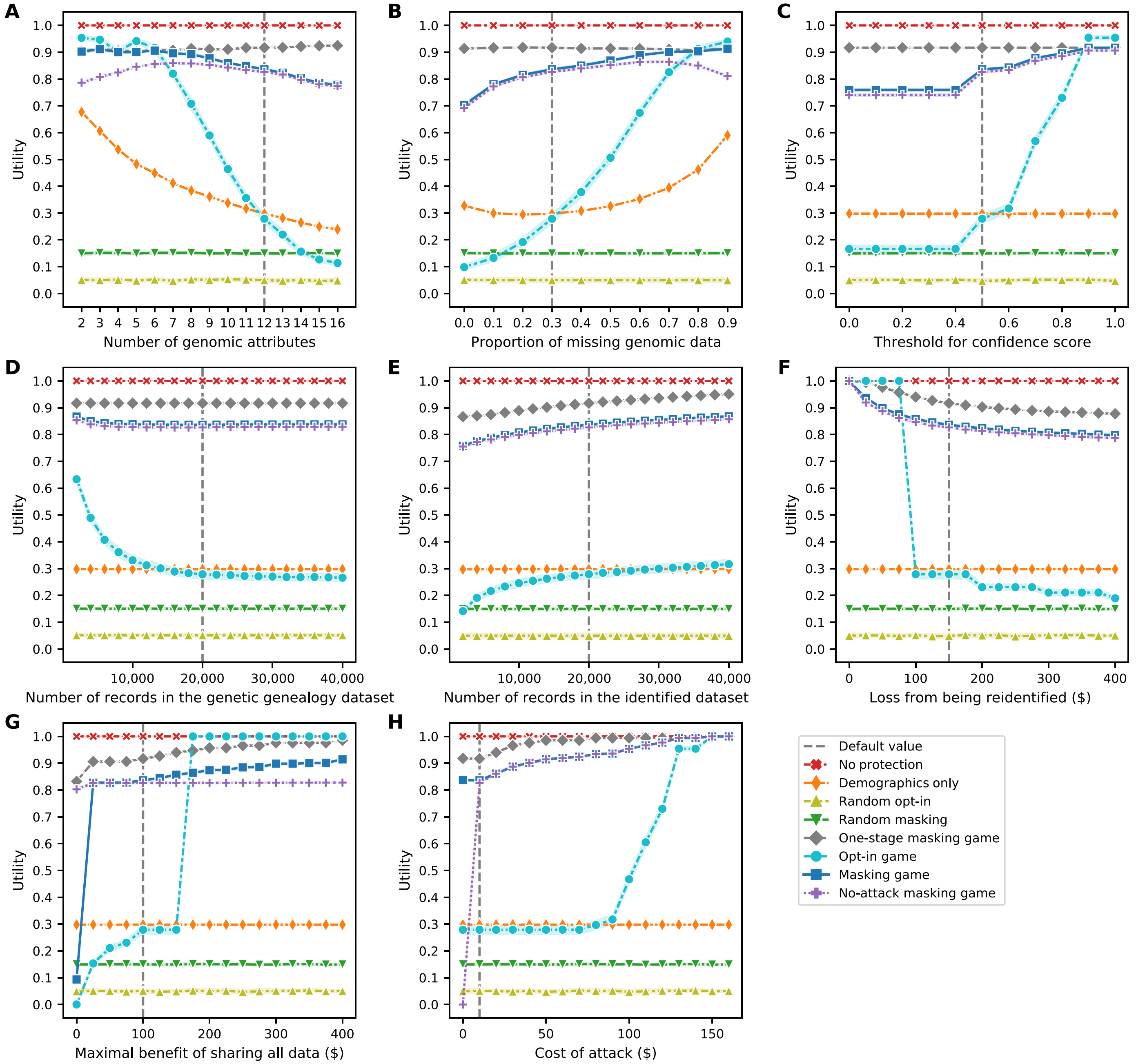

**Fig. 6. Sensitivity of the data subjects' average data utility as a function of the parameters in the model.** (**A**) Number of genomic attributes. (**B**) Proportion of missing genomic data. (**C**) Threshold for confidence score. (**D**) Number of records in the genetic genealogy dataset. (**E**) Number of records in the identified dataset. (**F**) Loss from being re-identified. (**G**) Maximal benefit of sharing all data. (**H**) Cost of an attack. Each line plot (depicted using Seaborn) shows the data subjects' average payoffs, with error bars representing SDs, in eight scenarios.

time. For example, a previously safe and trusted database may start to be targeted and attacked by adversaries if it loosens its access policies or if its vulnerabilities are found. Second, we used a simplified decision-making model in which either each subject makes the sharing decision independently, or all subjects pick the same protection strategy. In the future, we will consider interactions among related data subjects [e.g., family members (*42*)]. Third, the players modeled in our current game model do not fully reason about all the uncertainties. Both players might have incomplete and/or imperfect information. To be more realistic, more complex game theoretic models, such as the Bayesian game, can be used to model incomplete and/or imperfect information. Although our approach appears robust to uncertainty in critical parameters, further qualitative investigation [e.g., a behavioral empirical study (*48*)] would help ensure that these parameters are representative of human decision-making and that they are estimated accurately. Fourth, while we have reasoned from the perspective of a data holder (e.g., the controller of a database) to analyze and control two types

**Table 4. Results for the case study based on Craig Venter's data and the Ysearch dataset.** Scenarios include (1) no protection, (2) demographics only, (3) random opt-in, (4) random masking, (5) opt-in game, (6) masking game, (7) no-attack masking game, and (8) one-stage masking game. Only genomic attributes could be masked in this case study.

| Searched suboptimal strategy | Number of shared Y-STRs | Inferred surname | Confidence score | Utility | Privacy | Benefit | Loss | Payoff | Scenarios recommend |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 50 | Venter | 0.6688 | 1 | 0.5 | $100 | $75 | $25 | 1 |
| 1 | 49 | Venter | 0.6688 | 0.9967 | 0.5 | $99.67 | $75 | $24.67 | / |
| 2 | 48 | Venter | 0.6688 | 0.9933 | 0.5 | $99.33 | $75 | $24.33 | / |
| 3 | 47 | Venter | 0.6688 | 0.9897 | 0.5 | $98.97 | $75 | $23.97 | / |
| 4 | 46 | Venter | 0.6688 | 0.9862 | 0.5 | $98.62 | $75 | $23.62 | / |
| 5 | 45 | Venter | 0.6688 | 0.9825 | 0.5 | $98.25 | $75 | $23.25 | / |
| 6 | 44 | Venter | 0.6688 | 0.9786 | 0.5 | $97.86 | $75 | $22.86 | / |
| 7 | 43 | Venter | 0.6435 | 0.9746 | 0.5 | $97.46 | $75 | $22.46 | / |
| 8 | 42 | Venter | 0.5484 | 0.9706 | 0.5 | $97.06 | $75 | $22.06 | / |
| 9 | 41 | Venter | 0.5481 | 0.9661 | 0.5 | $96.61 | $75 | $21.61 | / |
| 10* | 40 | Karlsson | 0.4662 | 0.9593 | 1 | $95.93 | $0 | $95.93 | 6, 7 |
| 11 | 39 | Karlsson | 0.4661 | 0.9544 | 1 | $95.44 | $0 | $95.44 | / |
| / | 0 | / | / | 0.44 | 0.00 | $44 | $0.00 | $44.00 | 2 |
| / | 0 | / | / | 0 | 1 | $0 | $0 | $0 | 3, 5 |
| / | 7 | Hinze | 0.4046 | 0.5043 | 1 | $50.43 | $0 | $50.43 | 4 |
| / | 50 | / | / | 1 | 0.00 | $100.00 | $0.00 | $100.00 | 8 |

*The best strategy in the masking game, found using the greedy algorithm with pruning.

of downstream effects of data sharing (i.e., usefulness and fairness), we have not investigated the data holder's optimal strategy. One potential strategy would be to control the monetary amount paid to subjects for sharing data. However, paying subjects for the loss of privacy is a questionable practice that requires further investigations into the ethical foundations of data sharing and the corresponding ethical and societal implications. When this strategy may be permissible, future investigations could model the data holder as a player in the game, considering the interaction between a subject and a data holder and the interaction between a data holder and an adversary, to uncover the data holder's optimal strategy. Fifth, similar to most suppression-based disclosure control methods [e.g., *k*-anonymity (*49*), *l*-diversity (*50*), and *t*-closeness (*51*)], our game theoretic model might induce values that are missing not at random (MNAR) in the released dataset. In recognition of the fact that MNAR may not be desirable in certain applications, it should be noted that data could be suppressed or masked in a completely random manner. However, as our experiments illustrate, this strategy leads to sharing substantially less data.

Our game theoretic model could be applied to other multistage privacy attacks such as membership inference and genome reconstruction attacks (*52*–*55*). Since several studies have shown correlations between SNP and STR markers (*56*, *57*) and the Gymrek attack infers surnames from datasets with only Y-STR markers, it is worth examining the effectiveness of our protection model against multistage attacks involving datasets with SNP markers. For those attacks that might require greater computational effort and efficiency, we believe that the performance of the search algorithms could be improved through the assistance of distributed computing architectures or graphical processing units as they can process multiple computations simultaneously. To handle the high computational cost of solving complex game models, especially in the face of multistage attacks, a type of game-as-a-service architecture has the potential to be deployed in cloud servers to provide distributed game modeling and solving as a service to data subjects. We believe that little training effort would be required for an individual to use our game solver. Alternatively, it is possible that our solution could be integrated as a service into existing data anonymization software, such as ARX (*58*), which has already been expanded to include a general game theoretic module for risk analysis (*59*). Even if some parameter required for the game theoretic module is not set accurately, the resulting strategy might still be effective, as our sensitivity and robustness analyses have demonstrated.

## SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abe9986

View/request a protocol for this paper from *Bio-protocol*.

## REFERENCES AND NOTES

1. All of Us research program investigators, The "All of Us" research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
2. J. W. Hazel, C. Slobogin, Who knows what, and when?: A survey of the privacy policies proffered by U.S. direct-to-consumer genetic testing companies. *Cornell J. Law Public Policy* **28**, 35–66 (2018).
3. W. N. Price II, I. G. Cohen, Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
4. US Department of Health and Human Services Office for Civil Rights, Standards for privacy and individually identifiable health information; final rule. *Fed. Regist.* **67**, 53181–53273 (2002).

5. D. G. Wheeland, Final NIH genomic data sharing policy. *Fed. Regist.* **79**, 51345–51354 (2014).
6. M. Shabani, P. Borry, Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur. J. Hum. Genet.* **26**, 149–156 (2018).
7. Y. Erlich, A. Narayanan, Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014).
8. M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, X. Wang, Privacy in the genomic era. *ACM Comput. Surv.* **48**, 6 (2015).
9. N. von Thenen, E. Ayday, A. E. Cicek, Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics* **35**, 365–371 (2019).
10. M. Shabani, L. Marelli, Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep.* **20**, e48316 (2019).
11. B. Greshake, P. E. Bayer, H. Rausch, J. Reda, OpenSNP–a crowdsourced web resource for personal genomics. *PLOS ONE* **9**, e89204 (2014).
12. J. L. Roberts, S. Pereira, A. L. McGuire, Should you profit from your genome? *Nat. Biotechnol.* **35**, 18–20 (2017).
13. R. Kain, S. Kahn, D. Thompson, D. Lewis, D. Barker, C. Bustamante, C. Cabou, A. Casdin, F. Garcia, J. Paragas, A. Patrinos, A. Rajagopal, S. F. Terry, A. Van Zeeland, E. Yu, Y. Erlich, D. Barry, Database shares that transform research subjects into partners. *Nat. Biotechnol.* **37**, 1112–1115 (2019).
14. F. K. Dankar, Data privacy through participant empowerment. *Nat. Comput. Sci.* **1**, 175–176 (2021).
15. B. Malin, L. Sweeney, Determining the identifiability of DNA database entries, in *Proceedings of the AMIA 2000 Annual Symposium*, American Medical Informatics Association, Los Angeles, CA, 4 to 8 November 2000, J. M. Overhage, Ed. (Hanley & Belfus, 2000), pp. 537–541.
16. B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.* **37**, 179–192 (2004).
17. L. Sweeney, A. Abu, J. Winn, Identifying participants in the personal genome project by name (a re-identification experiment). arXiv:1304.7605 [cs.CY] (29 April 2013).
18. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
19. A. Harmanci, M. Gerstein, Quantification of private information leakage from phenotype-genotype data: Linking attacks. *Nat. Methods* **13**, 251–256 (2016).
20. I. C. Anindya, H. Roy, M. Kantarcioglu, B. Malin, Building a dossier on the cheap: Integrating distributed personal data resources under cost constraints, in *CIKM '17: Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, ACM SIGIR and ACM SIGWEB, Singapore, Singapore, 6 to 10 November 2017 (Association for Computing Machinery, 2017), pp. 1549–1558.
21. C. Lippert, R. Sabatini, M. C. Maher, E. Y. Kang, S. Lee, O. Arikan, A. Harley, A. Bernal, P. Garst, V. Lavrenko, K. Yocum, T. Wong, M. Zhu, W.-Y. Yang, C. Chang, T. Lu, C. W. H. Lee, B. Hicks, S. Ramakrishnan, H. Tang, C. Xie, J. Piper, S. Brewerton, Y. Turpaz, A. Telenti, R. K. Roby, F. J. Och, J. C. Venter, Identification of individuals by trait prediction using whole-genome sequencing data. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10166–10171 (2017).
22. D. Sero, A. Zaidi, J. Li, J. D. White, T. B. G. Zarzar, M. L. Marazita, S. M. Weinberg, P. Suetens, D. Vandermeulen, J. K. Wagner, M. D. Shriver, P. Claes, Facial recognition from DNA using face-to-DNA classifiers. *Nat. Commun.* **10**, 2557 (2019).
23. The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
24. L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, The complexities of genomic identifiability. *Science* **339**, 275–276 (2013).
25. Threlkeld, Free databases Ysearch and Mitosearch closing May 24 [Blog] (2018); https://casestone.com/threlkeld/home/latest-news/94-free-databases-ysearch-and-mitosearch-closing-may-24 [accessed 1 October 2021].
26. B. Malin, D. Karp, R. H. Scheuermann, Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J. Invest. Med.* **58**, 11–18 (2010).
27. US Equal Employment Opportunity Commission, Genetic Information Nondiscrimination Act of 2008; Final Rule. *Fed. Regist.* **81**, 31143–31159 (2016).
28. E. W. Clayton, B. J. Evans, J. W. Hazel, M. A. Rothstein, The law of genetic privacy: Applications, implications, and limitations. *J. Law Biosci.* **6**, 1–36 (2019).
29. B. A. Malin, An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J. Am. Med. Inform. Assoc.* **12**, 28–34 (2005).
30. B. Malin, G. Loukides, K. Benitez, E. W. Clayton, Identifiability in biobanks: Models, measures, and mitigation strategies. *Hum. Genet.* **130**, 383–392 (2011).
31. S. Wang, X. Jiang, S. Singh, R. Marmor, L. Bonomi, D. Fox, M. Dow, L. Ohno-Machado, Genome privacy: Challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Ann. N. Y. Acad. Sci.* **1387**, 73–83 (2017).
32. B. Berger, H. Cho, Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol.* **20**, 128 (2019).
33. D. Grishin, K. Obbad, G. M. Church, Data privacy in the age of personal genomics. *Nat. Biotechnol.* **37**, 1115–1117 (2019).
34. M. M. A. Aziz, M. N. Sadat, D. Alhadidi, S. Wang, X. Jiang, C. L. Brown, N. Mohammed, Privacy-preserving techniques of genomic data—A survey. *Brief. Bioinform.* **20**, 887–895 (2019).
35. A. Mittos, B. Malin, E. de Cristofaro, Systematizing genome privacy research: A privacy-enhancing technologies perspective, in *Proceedings on Privacy Enhancing Technologies*, PETS 2019, Stockholm, Sweden, 16 to 20 July 2019, K. Chatzikokolakis, C. Troncoso, Eds. (Sciendo, 2019), vol. 2019, pp. 87–107.
36. L. Bonomi, Y. Huang, L. Ohno-Machado, Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).
37. L. Sweeney, *k*-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz.* **10**, 557–570 (2002).
38. C. Dwork, Differential privacy, in *Lecture Notes in Computer Science: Proceedings, Part II, of the 33rd International Colloquium on Automata, Languages and Programming*, ICALP 2006, Venice, Italy, 10 to 14 July 2006, M. Bugliesi, B. Preneel, V. Sassone, I. Wegener, Eds. (Springer, 2006), vol. 4052, pp. 1–12.
39. G. Loukides, A. Gkoulalas-Divanis, B. Malin, Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7898–7903 (2010).
40. F. Tramer, Z. Huang, J.-P. Hubaux, E. Ayday, Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies, in *CCS '15: Proceedings of the 22nd ACM Conference on Computer and Communications Security*, ACM SIGSAC, Denver, CO, 12 to 16 October 2015 (Association for Computing Machinery, 2015), pp. 1286–1297.
41. Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, R. Ganta, R. Heatherly, B. A. Malin, A game theoretic framework for analyzing re-identification risk. *PLOS ONE* **10**, e0120592 (2015).
42. M. Humbert, E. Ayday, J.-P. Hubaux, A. Telenti, On non-cooperative genomic privacy, in *Lecture Notes in Computer Science: Revised Selected Papers of the 19th International Conference on Financial Cryptography and Data Security*, FC 2015, San Juan, Puerto Rico, 26 to 30 January 2015, R. Böhme, T. Okamoto, Eds. (Springer, 2015), vol. 8975, pp. 407–426.
43. Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, B. Malin, Expanding access to large-scale genomic data while promoting privacy: A game theoretic approach. *Am. J. Hum. Genet.* **100**, 316–322 (2017).
44. M. P. Ball, J. R. Bobe, M. F. Chou, T. Clegg, P. W. Estep, J. E. Lunshof, W. Vandewege, A. W. Zaranek, G. M. Church, Harvard personal genome project: Lessons from participatory public research. *Genome Med.* **6**, 10 (2014).
45. J. Watson, *Strategy: An Introduction to Game Theory* (W. W. Norton, ed. 3, 2013).
46. M. Tambe, *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned* (Cambridge Univ. Press, 2012).
47. W. Xia, M. Kantarcioglu, Z. Wan, R. Heatherly, Y. Vorobeychik, B. Malin, Process-driven data privacy, in *CIKM '15: Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, ACM SIGIR and ACM SIGWEB, Melbourne, Australia, 19 to 23 October 2015 (Association for Computing Machinery, 2015), pp. 1021–1030.
48. A. Acquisti, L. Brandimarte, G. Loewenstein, Privacy and human behavior in the age of information. *Science* **347**, 509–514 (2015).
49. L. Sweeney, Achieving *k*-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzz.* **10**, 571–588 (2002).
50. A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, *l*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* **1**, 3-es (2007).
51. N. Li, T. Li, S. Venkatasubramanian, *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity, in *ICDE '07: Proceeding of the 2007 IEEE 23rd International Conference on Data Engineering*, IEEE Computer Society, Istanbul, Turkey, 15 to 20 April 2007 (Institute of Electrical and Electronics Engineers, 2007), pp. 106–115.
52. R. Wang, Y. F. Li, X. Wang, H. Tang, X. Zhou, Learning your identity and disease from research papers: Information leaks in genome wide association study, in *CCS '09: Proceedings of the 16th ACM Conference on Computer and Communications Security*, ACM SIGSAC, Chicago, IL, 9 to 13 November 2009 (Association for Computing Machinery, 2009), pp. 534–544.
53. D. Garcia, Leaking privacy and shadow profiles in online social networks. *Sci. Adv.* **3**, e1701172 (2017).
54. I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, E. Ayday, An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 1333–1343 (2018).
55. K. Ayoz, E. Ayday, A. E. Cicek, Genome reconstruction attacks against genomic data-sharing beacons, in *Proceedings on Privacy Enhancing Technologies*, PETS 2021, The Internet, 12 to 16 July 2021, A. Johnson, F. Kerschbaum, Eds. (Sciendo, 2021), vol. 2021, pp. 28–48.

56. M. D. Edge, B. F. B. Algee-Hewitt, T. J. Pemberton, J. Z. Li, N. A. Rosenberg, Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5671–5676 (2017).
57. J. Kim, M. D. Edge, B. F. B. Algee-Hewitt, J. Z. Li, N. A. Rosenberg, Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell* **175**, 848–858.e6 (2018).
58. F. Prasser, J. Eicher, H. Spengler, R. Bild, K. A. Kuhn, Flexible data anonymization using ARX—Current status and challenges ahead. *Softw. Pract. Exp.* **50**, 1277–1304 (2020).
59. F. Prasser, J. Gaupp, Z. Wan, W. Xia, Y. Vorobeychik, M. Kantarcioglu, K. Kuhn, B. Malin, An open source tool for game theoretic health data de-identification, in *AMIA 2017 Annual Symposium Proceedings*, American Medical Informatics Association, Washington, DC, 4 to 8 November 2017 (American Medical Informatics Association, 2017), pp. 1430–1439.
60. T. Dalenius, Finding a needle in a haystack or identifying anonymous census record. *J. Off. Stat.* **2**, 329–336 (1986).
61. L. Sweeney. Simple demographics often identify people uniquely. (Technical Report LIDAP-WP3, Carnegie Mellon University, 2000); https://dataprivacylab.org/projects/identifiability/paper1.pdf [accessed 1 October 2021].
62. Y. Erlich, T. Shor, I. Pe'er, S. Carmi, Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
63. P. Ney, L. Ceze, T. Kohno, Genotype extraction and false relative attacks: Security risks to third-party genetic genealogy services beyond identity inference, in *Proceedings of the 2020 Network and Distributed System Security Symposium*, NDSS 2020, San Diego, CA, 23 to 26 February 2020 (Internet Society, 2020); https://dx.doi.org/10.14722/ndss.2020.23049 [accessed 1 October 2021].
64. M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, J.-P. Hubaux, De-anonymizing genomic databases using phenotypic traits, in *Proceedings on Privacy Enhancing Technologies*, PETS 2015, Philadelphia, PA, 30 June to 2 July 2015, A. Kapadia, S. J. Murdoch, Eds. (De Gruyter, 2015), vol. 2015, pp. 99–114.
65. R. Cai, Z. Hao, M. Winslett, X. Xiao, Y. Yang, Z. Zhang, S. Zhou, Deterministic identification of specific individuals from GWAS results. *Bioinformatics* **31**, 1701–1707 (2015).
66. L. Zhang, Q. Pan, Y. Wang, X. Wu, X. Shi, Bayesian network construction and genotype-phenotype inference using gwas statistics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 475–489 (2017).
67. B. A. Malin, Protecting genomic sequence anonymity with generalization lattices. *Methods Inf. Med.* **44**, 687–692 (2005).
68. S. Sankararaman, G. Obozinski, M. I. Jordan, E. Halperin, Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* **41**, 965–967 (2009).
69. M. Humbert, E. Ayday, J.-P. Hubaux, A. Telenti, Reconciling utility with privacy in genomics, in *WPES '14: Proceedings of the 13th Workshop on Privacy in the Electronic Society*, ACM SIGSAC, Scottsdale, AZ, 3 November 2014 (Association for Computing Machinery, 2014), pp. 11–20.
70. G. Kale, E. Ayday, O. Tastan, A utility maximizing and privacy preserving approach for protecting kinship in genomic databases. *Bioinformatics* **34**, 181–189 (2018).
71. E. Yilmaz, E. Ayday, T. Ji, P. Li, Preserving genomic privacy via selective sharing, in *WPES '20: Proceedings of the 19th Workshop on Privacy in the Electronic Society*, ACM SIGSAC, Virtual Event, USA, 9 November 2020 (Association for Computing Machinery, 2020), pp. 163–179.
72. S. Simmons, C. Sahinalp, B. Berger, Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Syst.* **3**, 54–61 (2016).
73. H. Cho, S. Simmons, R. Kim, B. Berger, Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. *Cell Syst.* **10**, 408–416.e9 (2020).
74. H. Bae, D. Jung, H. Choi, S. Yoon, AnomiGAN: Generative adversarial networks for anonymizing private medical data, in *Biocomputing 2020: Proceedings of the Pacific Symposium*, PSB 2020, Kohala Coast, Hawaii, 3 to 7 January 2020, R. B. Altman, A. K. Dunker, L. Hunter, M. D. Ritchie, T. Murray, T. E. Klein, Eds. (World Scientific, 2020), pp. 563–574.
75. B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani, F. Jay, Creating artificial human genomes using generative neural networks. *PLOS Genet.* **17**, e1009303 (2021).
76. B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, C. S. Greene, Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* **12**, e005122 (2019).
77. E. Halperin, D. A. Stephan, SNP imputation in association studies. *Nat. Biotechnol.* **27**, 349–351 (2009).
78. J. Marchini, B. Howie, Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
79. J. Kaye, L. Curren, N. Anderson, K. Edwards, S. M. Fullerton, N. Kanellopoulou, D. Lund, D. G. MacArthur, D. Mascalzoni, J. Shepherd, P. L. Taylor, S. F. Terry, S. F. Winter, From patients to partners: Participant-centric initiatives in biomedical research. *Nat. Rev. Genet.* **13**, 371–376 (2012).
80. E. W. Clayton, C. M. Halverson, N. A. Sathe, B. A. Malin, A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLOS ONE* **13**, e0204417 (2018).
81. A. L. McGuire, J. M. Oliver, M. J. Slashinski, J. L. Graves, T. Wang, P. Adam Kelly, W. Fisher, C. C. Lau, J. Goss, M. Okcu, D. Treadwell-Deering, A. M. Goldman, J. L. Noebels, S. G. Hilsenbeck, To share or not to share: A randomized trial of consent for data sharing in genome research. *Genet. Med.* **13**, 948–955 (2011).
82. J. M. Oliver, M. J. Slashinski, T. Wang, P. A. Kelly, S. G. Hilsenbeck, A. L. McGuire, Balancing the risks and benefits of genomic data sharing: Genome research participants' perspectives. *Public Health Genomics* **15**, 106–114 (2012).
83. D. Deuber, C. Egger, K. Fech, G. Malavolta, D. Schroder, S. A. K. Thyagarajan, F. Battke, C. Durand, My genome belongs to me: Controlling third party computation on genomic data, in *Proceedings on Privacy Enhancing Technologies*, PETS 2019, Stockholm, Sweden, 16 to 20 July 2019, K. Chatzikokolakis, C. Troncoso, Eds. (Sciendo, 2019), vol. 2019, pp. 108–132.
84. H. I. Ozercan, A. M. Ileri, E. Ayday, C. Alkan, Realizing the potential of blockchain technologies in genomics. *Genome Res.* **28**, 1255–1263 (2018).
85. K. Gammon, Experimenting with blockchain: Can one technology boost both data integrity and patients' pocketbooks? *Nat. Med.* **24**, 378–381 (2018).
86. D. Grishin, J. L. Raisaro, J. R. Troncoso-Pastoriza, K. Obbad, K. Quinn, M. Misbach, J. Gollhardt, J. Sa, J. Fellay, G. M. Church, J.-P. Hubaux, Citizen-centered, auditable and privacy-preserving population genomics. *Nat. Comput. Sci.* **1**, 192–198 (2021).
87. W. Xia, R. Heatherly, X. Ding, J. Li, B. A. Malin, R-U policy frontiers for health data de-identification. *J. Am. Med. Inform. Assoc.* **22**, 1029–1041 (2015).
88. M. Humbert, E. Ayday, J.-P. Hubaux, A. Telenti, Addressing the concerns of the lacks family: Quantification of kin genomic privacy, in *CCS '13: Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security*, ACM SIGSAC, Berlin, Germany, 4 to 8 November 2013 (Association for Computing Machinery, 2015), pp. 1141–1152.
89. M. Humbert, E. Ayday, J.-P. Hubaux, A. Telenti, Quantifying interdependent risks in genomic privacy. *ACM Trans. Priv. Secur.* **20**, 3 (2017).
90. D. W. Craig, R. M. Goor, Z. Wang, J. Paschall, J. Ostell, M. Feolo, S. T. Sherry, T. A. Manolio, Assessing and managing risk when sharing aggregate genetic variant data. *Nat. Rev. Genet.* **12**, 730–736 (2011).
91. Z. Wan, Y. Vorobeychik, M. Kantarcioglu, B. Malin, Controlling the signal: Practical privacy protection of genomic data sharing through beacon services. *BMC Med. Genomics* **10**, 39 (2017).
92. S. S. Shringarpure, C. D. Bustamante, Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* **97**, 631–646 (2015).
93. I. Wagner, Evaluating the strength of genomic privacy metrics. *ACM Trans. Priv. Secur.* **20**, 2 (2017).
94. M. H. Wooders, Equivalence of games and markets. *Econometrica* **62**, 1141–1160 (1994).
95. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations* (Cambridge Univ. Press, 2008).
96. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in neural information processing systems (NIPS '14): Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2*, NIPS 2014, Montreal, Canada, 8 to 13 December 2014, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, Eds. (MIT Press, 2014), pp. 2672–2680; https://dl.acm.org/doi/10.5555/2969033.2969125 [accessed 1 October 2021].
97. A. Luedtke, M. Carone, N. Simon, O. Sofrygin, Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Sci. Adv.* **6**, eaaw2140 (2020).
98. R. Nix, M. Kantarcioglu, Incentive compatible privacy-preserving distributed classification. *IEEE Trans. Dependable Secure Comput.* **9**, 451–462 (2012).
99. J. Blocki, N. Christin, A. Datta, A. D. Procaccia, A. Sinha, Audit games, in *IJCAI '13: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI 2013, Beijing, China, 3 to 9 August 2013, F. Rossi, Ed. (AAAI Press, 2013), pp. 41–47; https://dl.acm.org/doi/10.5555/2540128.2540137 [accessed 1 October 2021].
100. M. Li, D. Carrell, J. Aberdeen, L. Hirschman, J. Kirby, B. Li, Y. Vorobeychik, B. A. Malin, Optimizing annotation resources for natural language de-identification via a game theoretic framework. *J. Biomed. Inform.* **61**, 97–109 (2016).
101. R. Shokri, G. Theodorakopoulos, C. Troncoso, Privacy games along location traces: A game theoretic framework for optimizing location privacy. *ACM Trans. Priv. Secur.* **19**, 11 (2016).
102. Z. Wan, Y. Vorobeychik, E. Wright Clayton, M. Kantarcioglu, B. A. Malin, Game theory for privacy-preserving sharing of genomic data, in *Responsible Genomic Data Sharing: Challenges and Approaches*, X. Jiang, H. Tang, Eds. (Academic Press, 2020), pp. 135–160.
103. L. Sweeney, "Computational disclosure control: A primer on data privacy protection," thesis, Massachusetts Institute of Technology, Cambridge, MA (2001).
104. A. Gionis, T. Tassa, *k*-Anonymization with minimal loss of information. *IEEE Trans. Knowl. Data Eng.* **21**, 206–219 (2008).
105. J. Branson, N. Good, J. Chen, W. Monge, C. Probst, K. El Emam, Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations. *Trials* **21**, 200 (2020).

106. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).

107. M. Shabani, Blockchain-based platforms for genomic data sharing: A de-centralized approach in response to the governance problems? *J. Am. Med. Inform. Assoc.* **26**, 76–80 (2019).

108. A. Artyushina, The EU is launching a market for personal data. Here's what that means for privacy, in *MIT Technology Review* (11 August 2020); www.technologyreview.com/2020/08/11/1006555/eu-data-trust-trusts-project-privacy-policy-opinion [accessed 1 October 2021].

109. F. K. Dankar, K. El Emam, A method for evaluating marketer re-identification risk, in *EDBT '10: Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT/ICDT 2010, Lausanne, Switzerland, 22 to 26 March 2010, F. Daniel, L. Delcambre, F. Fotouhi, I. Garrigós, G. Guerrini, J.-N. Mazón, M. Mesiti, S. Müller-Feuerstein, J. Trujillo, T. M. Truta, B. Volz, E. Waller, L. Xiong, E. Zimányi, Eds. (Association for Computing Machinery, 2010), pp. 28.

110. B. Peng, M. Kimmel, simuPOP: A forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).

111. B. Peng, C. I. Amos, Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics* **24**, 1408–1409 (2008).

112. J. Comenetz, Frequently occurring surnames in the 2010 census. US Census Bureau (2016); www.census.gov/topics/population/genealogy/data/2010_surnames.html [accessed 1 October 2021].

113. US Census Bureau, Population, population change, and estimated components of population change: April 1, 2010 to July 1, 2019 (NST-EST2019-alldata). US Census Bureau (2019); www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html [accessed 1 October 2021].

114. P. Sanchez-Diz, C. Alves, E. Carvalho, M. Carvalho, R. Espinheira, O. Garcia, M. F. Pinheiro, L. Pontes, M. J. Porto, O. Santapa, C. Silva, D. Sumita, S. Valente, M. Whittle, I. Yurrebaso, A. Carracedo, A. Amorim, L. Gusmão; GEP-ISFG (The Spanish and Portuguese Working Group of the International Society for Forensic Genetics), Population and segregation data on 17 Y-STRs: Results of a GEP-ISFG collaborative study. *Int. J. Legal Med.* **122**, 529–533 (2008).

115. US Census Bureau, Source of income in 2015- people 15 years old and over, by income of specified type in 2015, age, race, Hispanic origin, and sex, in *Current Population Survey, 2016 Annual Social and Economic Supplement*, US Census Bureau (2016); www2.census.gov/programs-surveys/cps/tables/pinc-08/2016/pinc08_1_1_1.xls [accessed 1 October 2021].

116. A. Sraders, What is the middle class? Income and range. *TheStreet*, 21 January 2019 [updated 2 April 2020]; www.thestreet.com/personal-finance/what-is-middle-class-14833259 [accessed 1 October 2021].

117. M. Kimura, T. Ohta, Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2868–2872 (1978).

118. J. A. Martin, B. E. Hamilton, M. J. K. Osterman, A. K. Driscoll, Births: Final data for 2018 [Table 3]. *Natl. Vital Stat. Rep.* **68**, 13 (2019); www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_13-508.pdf [accessed 1 October 2021].

119. K. El Emam, Seven ways to evaluate the utility of synthetic data. *IEEE Secur. Priv.* **18**, 56–59 (2020).

**Acknowledgments:** We thank D. Fabbri, D. Fisher, L. Davis, J. E. Below, R. Venkatesaramani, Z. Samad, C. Yan, and F. Prasser for discussions. We thank S. Nyemba for technical support. We thank anonymous reviewers for valuable comments on the manuscript. **Funding:** This research was funded, in part, by the following grants: NIH grants RM1HG009034 (to E.W.C. and B.A.M.), R01HG006844 (to B.A.M.), and R01LM009989 (to B.A.M.). **Author contributions:** Z.W., Y.V., M.K., and B.A.M. conceived and designed the study. Z.W., Y.V., M.W., M.K., and B.A.M. created models and developed the methodology. Z.W. prepared data and performed the experiments. Z.W., W.X., Y.L., and J.G. provisioned materials/analysis tools. Z.W., Y.V., W.X., Z.Y., E.W.C., and B.A.M. analyzed, interpreted, and presented the results. Z.W. wrote the original draft. All authors wrote the manuscript. All authors read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All source code and all relevant datasets are available at https://github.com/zhywan/msrigs (archived at https://doi.org/10.5281/zenodo.5543369).

Submitted 14 December 2020
Accepted 25 October 2021
Published 10 December 2021
10.1126/sciadv.abe9986