Defending Against Membership Inference Attacks on Beacon Services

RAJAGOPAL VENKATESARAMANI, Washington University in St. Louis
ZHIYU WAN and BRADLEY A. MALIN, Vanderbilt University
YEVGENIY VOROBEYCHIK, Washington University in St. Louis

Large genomic datasets are created through numerous activities, including recreational genealogical investigations, biomedical research, and clinical care. At the same time, genomic data has become valuable for reuse beyond their initial point of collection, but privacy concerns often hinder access. Beacon services have emerged to broaden accessibility to such data. These services enable users to query for the presence of a particular minor allele in a dataset, and information helps care providers determine if genomic variation is spurious or has some known clinical indication. However, various studies have shown that this process can leak information regarding if individuals are members of the underlying dataset. There are various approaches to mitigate this vulnerability, but they are limited in that they (1) typically rely on heuristics to add noise to the Beacon responses; (2) offer probabilistic privacy guarantees only, neglecting data utility; and (3) assume a batch setting where all queries arrive at once. In this article, we present a novel algorithmic framework to ensure privacy in a Beacon service setting with a minimal number of query response flips. We represent this problem as one of combinatorial optimization in both the batch setting and the online setting (where queries arrive sequentially). We introduce principled algorithms with both privacy and, in some cases, worst-case utility guarantees. Moreover, through extensive experiments, we show that the proposed approaches significantly outperform the state of the art in terms of privacy and utility, using a dataset consisting of 800 individuals and 1.3 million single nucleotide variants.

CCS Concepts: • Security and privacy → Pseudonymity, anonymity and untraceability; Privacy protections; • Applied computing → Bioinformatics;

Additional Key Words and Phrases: Beacon services, privacy protection, genomic databases, access control

ACM Reference format:
Rajagopal Venkatesaramani, Zhiyu Wan, Bradley A. Malin, and Yevgeniy Vorobeychik. 2023. Defending Against Membership Inference Attacks on Beacon Services. ACM Trans. Priv. Sec. 26, 3, Article 42 (July 2023), 32 pages.
https://doi.org/10.1145/3603627

This work was supported by the National Institutes of Health (NIH) under grant RM1HG009034 and the National Science Foundation (NSF) CAREER award program under grant IIS-1905558.

Authors’ addresses: R. Venkatesaramani and Y. Vorobeychik, Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63108; emails: {rajagopal, yvorobeychik}@wustl.edu; Z. Wan and B. A. Malin, Vanderbilt University; 2201 West End Avenue, Nashville, TN 37235; emails: zhiyu.wan@vanderbilt.edu, b.malin@vumc.org.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2023 Copyright held by the owner/author(s).
2471-2566/2023/07-ART42 $15.00
https://doi.org/10.1145/3603627

ACM Transactions on Privacy and Security, Vol. 26, No. 3, Article 42. Publication date: July 2023.
1 INTRODUCTION

Genomic sequencing has become sufficiently cheap to support a wide range of services in the clinical and biomedical research domains, as well as for recreational consumers. As a result, the creation of large databases of genome sequences has become commonplace. However, not all organizations have access to the same information, thus there is a need to make such information more widely available. The sharing of genomic data has the potential to stimulate further scientific and clinical advances; however, it also introduces privacy risks. For example, healthcare organizations harboring genomic data may promise their patients that they will not disclose information that can be tied back to them. This balance between privacy and the value of shared data (commonly referred to as utility) has led to the creation of genomic data sharing services to reveal limited amounts of genomic information, for example, by sharing only summary statistics [19, 21, 31, 32].

Over the past several years, the Beacon service, which is promoted by the GA4GH (Global Alliance for Genomics and Health), has been increasingly adopted. This service enables a user to query for the presence of particular minor alleles in an underlying private genomic dataset (i.e., the non-dominant variant in a specific position in the genome). Although exposing such limited information may appear safe, it has been shown to be vulnerable to membership inference attacks because it allows users to issue queries for every region of the genome [22, 30]. These attacks assume that the attacker knows the genome of the target and leverage a statistical test, often in the form of a likelihood ratio, that couples this information with the Beacon response to a collection of queries to determine whether the target is a member of the Beacon dataset. The resulting membership inference can in turn reveal sensitive information about the individual, such as their health status, that membership in this dataset entails, or simply be in violation of the privacy promises made to the dataset constituents when the data was collected.

A common approach to mitigate privacy risks in Beacon services is to flip the values in a subset of the query responses [7, 17, 30]—for example, responding that a particular allele is absent when, in fact, it is present in the dataset. However, not all methods offer privacy guarantees, and when they do, they are often probabilistic, as is the case for those that are based on statistical perturbations, such as Differential Privacy (DP) [2, 7, 8]. Moreover, while minimizing the number of flipped queries is a standard measure of utility, no prior approaches offer formal optimality guarantees.

We introduce a novel framework for preserving privacy in the context of membership inference attacks on Beacon services. We consider both a batch Beacon setting, where all queries by a given party are specified at once, and an online Beacon setting, in which queries arrive sequentially. The former has been the primary focus of prior Beacon privacy analyses [7, 17, 22, 30], whereas the latter is more akin to the way that Beacon services are actually run in practice (e.g., https://beacon-network.org/), and to the best of our knowledge, a formal framework for the online Beacon has not been proposed before. Additionally, we consider two threat models. The first, and most common in prior literature, involves an a priori fixed threshold used by an attacker in the Likelihood Ratio Test (LRT). The second is adaptive in the sense that it takes as input Beacon query responses after the flipping strategy has been applied, along with a secondary genomic dataset, to adaptively identify a threshold that separates those in the Beacon dataset from those who are not. The former threat model captures adversaries that are more opportunistic (e.g., a parent trying to determine sensitive information about a child), whereas the latter models stronger, highly informed adversaries who are trying to systematically harvest data (e.g., to sell to others). To the best of our knowledge, ours is the first work to consider such an adaptive attacker model.

We present algorithmic approaches in each of the aforementioned privacy settings. Our strongest results are in the batch setting with fixed-threshold attacks, where we show that in the
important special case of very small sequencing rates, we can obtain both privacy and a provable worst-case approximation of the minimum number of queries to flip by drawing a connection to the set cover problem. We further provide principled algorithmic approaches for the general problems in the batch setting, for both fixed-threshold and adaptive threat models, all of which guarantee privacy under these threat models. Moreover, we present effective algorithms for preserving privacy while minimizing the number of flipped queries in the online setting. Finally, through extensive experiments, we demonstrate that the proposed approaches significantly outperform the state of the art, including those based on DP, in terms of utility (minimizing the number of flipped queries) when privacy (in the context of likelihood-ratio based membership inference attacks) can be guaranteed, and in privacy, utility, or both when complete privacy cannot be achieved for all members of the Beacon dataset. Furthermore, we show that the performance of the proposed approaches remains robust even to adaptive attacks that try to infer which queries have been flipped.

To summarize, we make the following contributions:

1. A novel mathematical framework for preserving privacy against membership inference attacks in Beacon services,
2. a formal model of an online Beacon,
3. a novel adaptive attacker model for membership inference that is more powerful than conventional likelihood ratio based attacks,
4. principled algorithmic approaches for preserving privacy in Beacon services with worst-case privacy guarantees, and
5. a comprehensive empirical evaluation of our proposed methods compared with state-of-the-art baselines against likelihood ratio based non-adaptive and adaptive attacks.

2 PRELIMINARIES

A Beacon is a web service that responds to queries about the presence/absence of a specific allele (say, nucleotide A) at a particular position (e.g., position 1,212,028) on a particular chromosome (say, chromosome 10) for any genomic records in the database [16, 26]. Such queries are only meaningful when there is variation of alleles in the overall population, and the positions that exhibit such variation are typically referred to as Single Nucleotide Variants (SNVs). Thus, we say that the Beacon service responds to queries about the existence of a particular SNV.

Now, suppose that the Beacon service (or simply the Beacon) responds to queries pertaining to a collection of \( m \) SNVs, which we index by an integer \( j \in \{1,\ldots,m\} \). An SNV for each individual \( i \) actually contains two alleles, one from each parent, but to simplify the discussion, we encode each SNV \( j \) as a binary value \( d_{ij} \) that indicates the presence (\( d_{ij} = 1 \)) or the absence (\( d_{ij} = 0 \)) of the minor allele. As such, we represent an individual \( i \) as a binary vector \( d_i = \{d_{i1},\ldots,d_{ij},\ldots,d_{im}\} \).

We say that the individuals who are a part of this dataset are in the Beacon, contrasting with those not in this dataset, who are referred to as not in the Beacon. When an SNV at position \( j \) is queried, the Beacon returns a response \( x_j = 1 \) if at least one individual \( i \) in the Beacon has the minor allele, and \( x_j = 0 \) otherwise.

Let \( B \) be the set of \( n \) individuals in the Beacon, and let \( S \) be the set of \( m \) SNV positions that can be queried in the Beacon (in other words, we can view Beacon queries simply as integer indices corresponding to SNV positions). We define \( \delta \) as the genomic sequencing error and let \( f^j \) be the Alternative (minor) Allele Frequency (AAF) in the population for each position \( j \). Let \( D^j_n \) denote the probability that no individual in the Beacon has a minor allele at position \( j \), which is calculated as

\[
D^j_n = (1 - f^j)^{2n}
\]

(recall that each SNV has two alleles for each individual, hence \( 2n \)).
We begin with the well-known membership inference attack on the Beacon service by Shringarpure and Bustamante [22], with the additional assumption that the attacker knows the AAF for each position \( j \) [30]. In this attack, the attacker first submits a collection of queries \( Q \subseteq S \) to the Beacon service, then uses these to calculate the LRT statistic for each individual in the attacker’s target set \( T \) (i.e., the set of individuals whose membership in the Beacon the attacker wishes to infer). Specifically, given an individual \( i \in T \), a set of queries \( Q \), and the vector of query responses \( x \), the LRT statistic is

\[
L_i(Q, x) = \sum_{j \in Q} d_{ij} \left( x_j \log \frac{1 - D_j^i}{1 - \delta D^i} + (1 - x_j) \log \frac{D_j^i}{\delta D^i} \right).
\]  

Finally, the attacker claims that an individual \( i \in T \) is in the Beacon \( B \) when \( L_i(Q, x) < \theta \). The choice of \( \theta \) reflects the adversary’s preferred balance between sensitivity and specificity of the membership inference attack.

3 THREAT MODELS

Our threat models for membership inference attacks on the Beacon service are anchored in the Shringarpure-Bustamante LRT attack described previously. However, the LRT attack leaves open three questions. First, what is the attacker’s target set \( T \)? Second, how does the attacker arrive at the choice of a threshold \( \theta \) to determine which Beacon membership claims are made? Third, what is the set of queries used in the attack? Since we do not know \textit{a priori} which individuals will be targeted, we make the worst-case assumption that \( B \subseteq T \); in other words, the attacker targets everyone in the Beacon, along with possibly others. Our threat modeling leads to several variants of the LRT attack along the remaining two dimensions. Figure 1 provides an overview of the threat models, categorized by inference threshold and query process.

3.1 Choosing the Inference Threshold

We investigate two approaches that an adversary may use to determine when to make a membership inference claim about an individual: fixed-threshold and adaptive attacks. In \textit{fixed-threshold} attacks, an adversary uses a predefined threshold \( \theta \), which is fixed for the inference attack. This is a common threat model in the literature [13, 30] and reflects an opportunistic attacker who initially uses a private dataset to simulate LRT attacks by splitting individuals into those in a simulated...
Defending Against Membership Inference Attacks on Beacon Services

Beacon and those who are not. These offline simulations are then used to identify the \( \theta \) that best balances precision and recall with respect to the attacker’s preferences about these. The practical consequence of assuming a fixed \( \theta \) at the time of attack is that \( \theta \) is not adjusted based on query responses.

Not considering queries in determining the threshold \( \theta \) is consequential once we consider defensive measures that modify query responses. The fixed-threshold model has some limitations. The attacker could, for instance, choose a higher value of \( \theta \) than what was used by any implemented defensive measures, which would subsequently lead to a violation of privacy. Further, the attacker might look to the distribution of LRT scores to separate the two populations. For example, if modified query responses preserve a clear separation in LRT statistics between individuals who are in and not in the Beacon, a simple clustering of the statistics would enable the attacker to effectively identify those in the Beacon. Consequently, we additionally consider a stronger adaptive threat model that sets \( \theta(x_Q) \) as a function of the responses \( x_Q \) to queries \( Q \), but with the aim of limiting the false positive rate for any membership claims to be at most \( \alpha \). This in turn forces the defensive measures to ensure that the LRT scores for individuals in the Beacon and those not in the Beacon be sufficiently well mixed. Further, limiting the false positive rate avoids having to consider unreasonable attacks such as claiming that all individuals are in the Beacon. This adaptive threat model requires that the attacker can set the threshold in precisely the right place based on actual queries \( Q \). However, this can effectively be accomplished with the aid of simulation experiments using a private dataset \( D \) (now, simulating Beacon queries that implement the defensive measures). This threat model captures highly informed attackers—for example, those who do systematic harvesting of sensitive data for profit.

### 3.2 Query Process

In this work, we distinguish between three mechanisms of query access that can be provisioned to a Beacon: (1) batch query, (2) unauthenticated online query access, and (3) authenticated online query access. In the batch setting, it is assumed that the attacker queries all \( m \) SNP positions effectively simultaneously. In a sense, this is the most favorable setting for the adversary, as it provides maximum information for making membership claim decisions. It is also the setting that has received much of the attention in prior literature [13, 30]. However, typical Beacons in practice (e.g., the service provided by the GA4GH Beacon Network) can be queried sequentially, and we thus need to ensure that privacy is guaranteed even in such settings. Consequently, we also consider two online settings. The unauthenticated online setting assumes that the attacker can submit an arbitrary subset of queries. This is because if queries are not connected to a particular identity, there is no way for the Beacon to know which queries have been made by the same individual in the past. The public Beacon Network is an example of this situation. The authenticated online setting, by contrast, assumes that we can keep track of all the past queries by each individual, including the potential adversary. This entails only allowing registered and verified users access (we assume no collusion) and allows (as we show in the experiments) privacy guarantees with significantly higher utility for such users.

We now make an observation that enables us to uniformly talk about the three variants of the preceding query process: note that the sole mathematical distinction between them is the set of queries \( Q \) that we are concerned about. In the batch setting, \( Q = S \), the set of all queries—in other words, the LRT statistics of relevance for the purposes of membership inference attack is computed with respect to the set of all possible queries. In the online authenticated online setting, \( Q \) is the set of all past queries, together with the current query \( j \), since we are concerned that an individual \( i \) may be identifiable as soon as \( L_i(Q) \) drops below \( \theta \) (or \( \theta(Q) \), in the adaptive model). Finally, in the unauthenticated online setting, since we do not know which set of queries have been asked
Finally, we note that since the privacy guarantee for flipping is monotonic in the threshold, the following proposition holds.

**Proposition 4.1.** Suppose that \( x_j = 0 \) given the Beacon dataset. Then \( y_j = 1 \) can never increase the LRT statistic for any individual \( i \in B \), provided sampling error \( \delta < \frac{D_k}{L_n} \) for all \( j \).

**Proof.** Consider the \( j \)th query. If individual \( i \) does not have an alternate allele at position \( j \), flipping the Beacon response makes no difference (refer to Equation (1); \( d_{ij} = 0 \) when the individual does not have an alternate allele at position \( j \)). When the individual does have an alternate allele
Defending Against Membership Inference Attacks on Beacon Services

42:7

at position \( j \) (i.e., \( d_{ij} = 1 \)), changing the Beacon response \( x_j \) from 0 to 1 changes the contribution of query \( j \) to the LRT score from \( \log \frac{D_n^j}{\delta D_{n-1}^j} \) to \( \log \frac{1-D_n^j}{1-\delta D_{n-1}^j} \). Given sampling error \( \delta < \frac{D_n^j}{D_{n-1}^j} \), dividing on both sides by \( \delta \), we have \( \frac{D_n^j}{\delta D_{n-1}^j} > 1 \) (as \( \delta \geq 0 \)) and, consequently, \( \log \frac{D_n^j}{\delta D_{n-1}^j} > 0 \). Since \( \delta < \frac{D_n^j}{D_{n-1}^j} \), multiplying both sides by \( D_{n-1}^j \) yields \( D_n^j > \delta D_{n-1}^j \) (since \( D_{n-1}^j \in [0, 1] \)), which implies that \( 1 - D_n^j < 1 - \delta D_{n-1}^j \). Dividing both sides by \( 1 - \delta D_{n-1}^j \), we have \( \frac{1-D_n^j}{1-\delta D_{n-1}^j} < 1 \), since \( \delta \in [0, 1] \), \( D_{n-1}^j \in [0, 1] \), which in turn implies that \( \log \frac{1-D_n^j}{1-\delta D_{n-1}^j} < 0 \). \( \square \)

Since a privacy violation means that the LRT statistic for at least one individual in the Beacon is too small, our goal is necessarily to increase these scores until privacy is guaranteed for all individuals in the Beacon. Henceforth, we assume that \( \delta < 0.25 < \frac{D_n^j}{D_{n-1}^j} = (1 - f^j)^2 \) (since \( f^j < 0.5 \) for all \( j \)). Consequently, Proposition 4.1 implies that flipping a 0 response to a 1 is counterproductive, and we need not consider it as a possible solution to Problem (3). While flipping a 0 response to a 1 may in principle help lower the scores for individuals not in the Beacon and therefore result in better-mixed LRT scores, we consider the issue of mixing more systematically as a part of the following adaptive attack model. Further, a majority (greater than 96%) of Beacon responses in our data are initially 1.

Then, without loss of generality, we can assume that our consideration set \( S \) includes only the query responses that are initially 1.

Next, we show that even in a very restricted special case, the problem of minimizing the number of flips to guarantee the privacy of all individuals in the Beacon is \( \text{NP} \)-hard.

First, we define a decision version of Problem (3), which we refer to as Beacon-Privacy-D.

Definition 4.2 (Beacon-Privacy-D). Input: A collection of individuals \( i \in B \) with genomic information \( D \) and induced Beacon query responses \( x \); a constant \( k \). Question: Can we flip a subset \( F \) of query responses where \( |F| \leq k \) and \( L_i(Q_i, x_{Q_i}(F)) - \theta(x_{Q_i}(F)) \geq 0 \)?

Theorem 4.3. Beacon-Privacy-D is \( \text{NP} \)-complete even if \( \delta = 0 \) and \( \theta(x_{Q_i}(F)) \) is a constant.

Proof. We reduce from the Set Cover problem, which we now formally define.

Definition 4.4 (Set Cover). Input: A universe \( U \) of elements, and a collection of sets \( R = \{R_1, \ldots, R_n\} \) with \( R_j \subseteq U \) and \( \bigcup_j R_j = U \); a constant \( k \). Question: Is there a subset \( T \subseteq R \) of sets such that \( U = \bigcup_{t \in T} t \) and \( |T| \leq k \).

First, note that Beacon-Privacy-D is in \( \text{NP} \), since given a set \( F \) of flips, it is straightforward to verify that the privacy constraint holds for each individual \( i \).

To prove that the problem is \( \text{NP} \)-hard, we reduce from the Set Cover problem. First, observe that in the case where \( \delta = 0 \), and by Proposition 4.1, to guarantee privacy of any individual \( i \in B \), it suffices to flip a single response \( x_j \) from 1 to 0 from all with \( d_{ij} = 1 \) (if we use the convention that division by 0 results in \( \infty \), any such flip causes \( L_i(Q, x) = \infty \)).

Now, let elements of \( U \) correspond to individuals in the Beacon (i.e., \( B = U \)). Let subsets \( R_j \) correspond to queries \( j \), where each element represents an individual \( i \) with \( d_{ij} = 1 \). Since without loss of generality we can assume that each \( R_j \) is non-empty (since we can ignore any empty subsets in both Set Cover and in the construction of Beacon-Privacy-D instance by Proposition 4.1), this also implies that the corresponding query response is \( x_j = 1 \), as at least one individual has \( d_{ij} = 1 \).

For any individual (element of \( U \)) \( i \notin R_j \), we set \( d_{ij} = 0 \). Furthermore, since \( \bigcup_j R_j = U \), each
individual has at least one \( j \) with \( d_{ij} = 1 \). Finally, the constant \( k \) is now the constraint on the size of \( F \), the subsets of queries to flip.

Suppose that we find the set \( F \) of queries to flip that guarantees privacy. Let \( T = F \)—that is, indices of subsets \( R_j \) in SET COVER. Since \(|F| \leq k\), \(|T| \leq k\), so it suffices to show that \( U = \bigcup_{t \in T} t \). Solution to BEACON-PRIVACY-D means that flips \( F \) guarantee privacy of each \( i \in B = U \). By the preceding observation that it suffices to flip any query \( j \) with \( d_{ij} = 1 \) to guarantee the privacy of \( i \), \( R_j \) is the subset of individuals for whom privacy is guaranteed, and thus \( \bigcup_{t \in F} R_j = U \), since we must guarantee privacy of all individuals. Since \( T = F \), we have covered the universe \( U \).

For the other direction, suppose that there exists a solution to SET COVER, \( T \) with \(|T| \leq k\) and \( U = \bigcup_{t \in T} t \). Set \( F = T \), and flip all queries with \( j \in F \). Since it suffices to guarantee privacy of any \( i \in B \) by flipping any query \( j \) with \( d_{ij} = 1 \), and since \( R_j \) is the collection of all individuals for whom we can guarantee privacy by flipping \( j \), and since \( \bigcup_{t \in T} t = U \), by our construction this implies that privacy is guaranteed for all \( i \in B \). \( \square \)

# 5 THE BATCH SETTING

We begin by considering the batch setting in which the adversary submits a set of queries \( Q \) all at once, where \( Q \neq \emptyset \). This provides the building blocks for all query settings in our threat model, both what we called the batch query setting with \( Q = S \) previously and for the online settings.

Recall that a binary vector \( x \) corresponds to the true query responses, whereas \( y \) represents whether responses have been flipped. Let \( Q_1 \) be the subset of queries \( Q \) with \( x_j = 1 \) and \( Q_0 \) be the subset with \( x_j = 0 \). Let \( A_j = \log \frac{1-D_n^j}{1-D_{n-1}^j} \) and \( B_j = \log \frac{D_n^j}{D_{n-1}^j} \). We can then rewrite \( L_i(Q, x) \) as follows:

\[
L_i(Q, x) = \sum_{j \in Q_1} d_{ij} A_j + \sum_{j \in Q_0} d_{ij} B_j.
\]

Moreover, by Proposition 4.1, we never flip queries in \( Q_0 \), which means that for our purposes, the preceding second term is a constant. Now, if we apply the query flip strategy \( y \), the resulting LRT statistic, which we denote by \( L_i(Q, x, y) \), becomes

\[
L_i(Q, x, y) = \sum_{j \in Q_1} d_{ij}(1 - y_j) A_j + y_j B_j + \sum_{j \in Q_0} d_{ij} B_j.
\]

Define \( \Delta_{ij} = d_{ij}(B_j - A_j) \) and \( \eta_i = \sum_{j \in Q_1} d_{ij} A_j + \sum_{j \in Q_0} d_{ij} B_j \). Then

\[
L_i(Q, x, y) = \sum_{j \in Q_1} \Delta_{ij} y_j + \eta_i.
\]  

(4)

Note that \( \eta_i \) is actually also a function of the set of queries \( Q \). For the remainder of this section, this will not be important and so we omit this dependence. However, this becomes important in the following online setting.

Next, we consider approaches to solve the BEACON-PRIVACY-PROBLEM first in the fixed-threshold and subsequently in the adaptive attacks.

## 5.1 Fixed-Threshold Attacks

We begin by presenting an Integer Linear Programming (ILP) approach for optimally solving the general BEACON-PRIVACY-PROBLEM in the batch setting with fixed-threshold attacks. This is a straightforward consequence of the problem structure that we derived earlier. First, note that we wish to minimize the number of flips, which is equivalent to minimizing the number of ones in \( y \).
Second, note that the privacy constraint is \( L_i(Q, x, y) \geq \theta \), which is linear in \( y \). Consequently, the following ILP solves the Beacon-Privacy-Problem:

\[
\min_{y \in \{0, 1\}^m} \sum_j y_j \text{ subject to } : \\
\sum_{j \in Q_1} \Delta_{ij} y_j + \eta_i \geq \theta \quad \forall \ i \in B. 
\]

This ILP has worst-case exponential running time and, as a result, will have trouble scaling to large problems that include thousands of individuals and millions of SNPs. We address the scalability in two ways: first, we identify important special cases that either enable the ILP solvers to leverage problem structure or admit an approximation algorithm with worst-case guarantees, and second, we present two greedy algorithms for solving the general variant of the Beacon-Privacy-Problem. The key property of all solutions we propose is that they satisfy the privacy constraints by construction.

5.1.1 Small Sequencing Error Rates. Genomic sequencing error rates \( \delta \) are often quite small, on the order of \( 10^{-6} \). We now show that for sufficiently small sequencing error rates (with \( \delta = 0 \) a special case), we can represent the Beacon-Privacy-Problem as a Minimum Set Cover instance. This in turn implies that we can solve our problem using a greedy algorithm with a logarithmic worst-case approximation guarantee.

Generally speaking, for a sufficiently small \( \delta \), the terms \( B_j \) will be extremely large for any query \( j \) that we may choose to flip from a 1 to a 0 and, in particular, will be much larger than \( A_j \). Thus, for every individual \( i \) in the Beacon, flipping any \( j \in Q_1 \) will result in a very large increase in \( \Delta_{ij} = d_{ij}(B_j - A_j) \). This increase will, indeed, be so large as to guarantee that \( L_i(Q, x, y) \geq \theta \). As a consequence, it will suffice to flip any query with \( d_{ij} = 1 \) for \( i \) to no longer be categorized as in the Beacon by the attack. Of course, just how small \( \delta \) needs to be for this to work will be a function of both the problem parameters \( D_{jn} \) and \( D_{jn-1} \), as well as \( \theta \). We emphasize that this line of reasoning is specific to the fixed-threshold attack model; the issue is far more subtle in adaptive attacks (Section 5.2). Next, we make this premise precise.

For each individual \( i \in B \), define \( P_i = \{ j \in Q_1 | d_{ij} = 1 \} \)—in other words, \( P_i \) is the set of all queries \( j \) for which (a) \( x_j = 1 \) and (b) the individual \( i \) actually has the associated alternate allele for query \( j \) (i.e., \( d_{ij} = 1 \)). We now provide a sufficient condition on \( \delta \) such that we can flip any query in \( P_i \) for each \( i \in B \) to guarantee privacy under fixed-threshold attacks.

**Definition 5.1.** A set of queries \( F \) that have been chosen to flip is a Beacon-Cover if \( \forall i \in B, F \cap P_i \neq \emptyset \).

In words, \( F \) is a Beacon-Cover if every individual in the Beacon is covered by some query flipped in \( F \) that is also in \( P_i \). We now define two additional components to the notation that will be useful throughout our analysis. First, define

\[
D_n = \min_{j \in Q_1} \log \left( \frac{D_{jn}}{1 - D_{jn}} \right).
\]

Second, define

\[
\eta = \min_i \left( \sum_{j \in Q_1} d_{ij} \log(1 - D_{jn}) + \sum_{j \in Q_0} d_{ij} \log \frac{D_{jn}}{0.25D_{jn-1}} \right).
\]

The following proposition presents a bound on \( \delta \) that ensures that a Beacon-Cover guarantees privacy against fixed-threshold attacks.
Theorem 5.2. Suppose that \( \delta \leq \frac{1}{1 + e^{\theta - \eta - D_n}} \). Then, if \( F \) is a Beacon-Cover, it guarantees privacy of all \( i \in B \) against fixed-threshold attacks with threshold \( \theta \).

Proof. Recall that for a fixed-threshold attack, the Beacon privacy guarantee for a given \( F \) and associated indicator vector \( y \), formalized in Equation (2), is that

\[
\forall i \in B, \quad L_i(Q, x, y) = \sum_{j \in Q_i} \Lambda_{ij} y_j + \eta_i \geq \theta.
\]

Let \( \Delta_i = \min_{j \in P_i} \Lambda_{ij} \). Then if for each \( i \), \( \Delta_i + \eta \geq \theta \), the preceding condition certainly follows as well, since \( \sum_{j \in Q_i} \Lambda_{ij} y_j \geq \Delta_i \) by definition of a Beacon-Cover, and

\[
\eta = \min_i \left( \sum_{j \in Q_i} d_{ij} \log(1 - D_h) + \sum_{j \in Q_o} d_{ij} \log \frac{D_{ij}^{j}}{0.25D_{n-1}^{j}} \right).
\]

\[
\leq \min_i \left( \sum_{j \in Q_i} d_{ij} \log \frac{1 - D_{n-1}^{j}}{1 - \delta D_{n-1}^{j}} + \sum_{j \in Q_o} d_{ij} \log \frac{D_{ij}^{j}}{\delta D_{n-1}^{j}} \right).
\]

\[
\leq \eta_i.
\]

Now, \( \Delta_i = \min_{j \in P_i} d_{ij}(B_j - A_j) \), and since \( \Delta_{ij} > 0 \) for any \( j \in P_i, d_{ij} = 1 \) for any \( j \in P_i \). Consequently,

\[
\Delta_i = \min_{j \in P_i} d_{ij}(B_j - A_j) = \min_{j \in P_i} \left( \log \frac{D_{ij}^{j}}{\delta D_{n-1}^{j}} - \log \frac{1 - D_{n-1}^{j}}{1 - \delta D_{n-1}^{j}} \right)
\]

\[
= \min_{j \in P_i} \left( \log D_h^{j} - \log \delta D_{n-1}^{j} - \log(1 - D_h^{j}) + \log(1 - \delta D_{n-1}^{j}) \right)
\]

\[
= \min_{j \in P_i} \left( \log \frac{D_h^{j}}{1 - D_h^{j}} + \log \frac{1 - \delta D_{n-1}^{j}}{\delta D_{n-1}^{j}} \right)
\]

\[
\geq \min_{j \in P_i} \left( \log \frac{D_h^{j}}{1 - D_h^{j}} + \min_{j \in P_i} \left( \log \frac{1 - \delta D_{n-1}^{j}}{\delta D_{n-1}^{j}} \right) \right)
\]

\[
\geq D_n + \min_{j \in P_i} \left( \log \left( \frac{1}{\delta D_{n-1}^{j}} - 1 \right) \right) \geq D_n + \log \left( \frac{1}{\delta} - 1 \right),
\]

where the last inequality follows since \( D_{n-1}^{j} \leq 1 \). Now, if \( \delta \leq \frac{1}{1 + e^{\theta - \eta - D_n}} \), then

\[
\log \left( \frac{1}{\delta} - 1 \right) \geq \log \left( e^{\theta - \eta - D_n} \right) = \theta - \eta - D_n.
\]

Consequently, \( \Delta_i \geq \theta - \eta \) for each \( i \in B \), which is just a rearranging of the preceding desired condition.

The benefit of Theorem 5.2 is that it suffices for \( F \) to “cover” each individual in the sense that for every individual \( i \) in the Beacon, there is at least one flipped query in \( F \) that suffices to ensure that the score \( L_i(Q, x, y) \geq \theta \) —that is, to ensure that \( i \)’s privacy is preserved under the fixed-threshold threat model in the batch setting. This in turn allows us to represent the Beacon-Privacy-Problem as a Min-Set-Cover instance. The Min-Set-Cover problem is the optimization variant of Set Cover, which we now define formally.

Definition 5.3 (Min-Set-Cover). Input: A universe \( U \) of elements, and a collection of sets \( R = \{ R_1, \ldots, R_n \} \) with \( R_j \subseteq U \) and \( \bigcup_j R_j = U \); a constant \( k \). Goal: Minimize \( |T| \) over \( T \subseteq R \) such that \( U = \bigcup_{t \in T} t \).
We now show how to represent our problem as an instance of the Min-Set-Cover problem. The key advantage of this representation will be a greedy algorithm for solving the Beacon-Privacy-Problem in this setting that yields a logarithmic worst-case approximation guarantee [24]. The representation is similar to the one used in the proof of Theorem 4.3 but, of course, is in the opposite direction. Specifically, we are given a Beacon-Privacy-Problem instance, which we now use to construct a Min-Set-Cover instance. Let \( U = B \), the set of the individuals in the Beacon, while each \( R_j \) corresponds to query \( j \), and is comprised of the individuals \( i \in B \) whose privacy will be protected if we flip \( j \). Formally, \( R_j = \{ i \in B | j \in P_i \} \).

Now, we can leverage the greedy algorithm for Min-Set-Cover to solve our problem. The greedy algorithm works as follows. The collection of subsets \( T \) is initialized to be empty. Then, in each iteration, we add a subset \( S_j \) to \( T \) that maximizes the number of elements in \( U \) it covers that are not already covered by \( T \). We stop when the entire universe \( U \) is covered. Algorithm 1, which we refer to as Greedy Min Beacon Cover (GMBC) presents a direct adaptation of this to our problem. The following is thus a direct corollary of Theorem 5.2.

**Corollary 5.4.** Suppose that \( \delta \leq \frac{1}{1+e^{\theta - \eta - D_n}} \). Then Algorithm 1 gives an \( O(\log(n)) \)-approximation to the Beacon-Privacy-Problem.

**Algorithm 1: Greedy Min Beacon Cover**

**Input:** A set \( B \) of individuals in the Beacon, a subset \( P_i \) for each individual, and a collection of queries \( Q \).

**Output:** Subsets of queries \( F \subseteq S \) to flip.

**Initialization:** \( F = \emptyset \), \( C = \emptyset \).

while \( (B \setminus C) \neq \emptyset \) do

Set \( l = 1 \), \( T = \emptyset \), \( N = -1 \).

for \( j \in (Q \setminus F) \) do

Set \( T_j = \{ i \in (B \setminus C) | j \in P_i \} \).

if \( |T_j| > N \) then

Set \( T = T_j \).

Set \( N = |T| \).

Set \( l = j \).

end

end

Set \( F = F \cup l \).

Set \( C = C \cup T \).

end

5.1.2 Alternate Allele Frequencies Drawn from the Beta Distribution. A common assumption in prior literature is that the AAFs are drawn from a beta distribution [30], replacing the \( D'_n \) and \( D'_{n-1} \) terms in the LRT score calculation with the expectation over the distribution, which we denote by \( \bar{D}_n \) and \( \bar{D}_{n-1} \), respectively. This in turn means that \( A_j \) and \( B_j \) are independent of \( j \), and we now denote them by constants \( A \) and \( B \), respectively. As a result, \( \Delta_{ij} = d_{ij}(B - A) \) and \( \eta_i = A \sum_{j \in Q_1} d_{ij} + B \sum_{j \in Q_0} d_{ij} \), and we obtain a simpler expression for the LRT statistic induced by query flips \( y \):

\[
L_i(Q, x, y) = (B - A) \sum_{j \in Q_1} d_{ij} y_j + \eta_i.
\]

Consequently, the privacy condition for each \( i \in B \) is equivalent to

\[
\sum_{j \in P_i} y_j \geq k_i, \quad \text{where} \quad k_i = \frac{\theta - \eta_i}{B - A}.
\]

Note that, under our assumption, \( \delta < 1/4, B - A > 0 \). This has two algorithmic implications. First, it yields a significantly simpler set of privacy constraints in the integer linear program (5) to obtain
the optimal solution to the Beacon-Privacy-Problem. Second, we can derive a natural greedy heuristic for this case that generalizes the preceding GMBC algorithm.

The high-level idea of the greedy heuristic is to iteratively choose a query result to flip that affects the largest number of individuals. This idea is formalized in Algorithm 2, which we term Greedy k-Cover (GKC).

**Algorithm 2: Greedy k-Cover**

**Input**: A set $B$ of individuals in the Beacon, a subset $P_i$ and a constant $k_i$ for each individual, and a collection of queries $Q$.

**Output**: Subsets of queries $F \subseteq S$ to flip.

**Initialization**: $F = \emptyset$, $C = \emptyset$.

while $(B \setminus C) \neq \emptyset$ do

  Set $l = 1$, $N = -1$.

  for $j \in (Q \setminus F)$ do

    Set $T_j = \{i \in (B \setminus C) | j \in P_i\}$.

    if $|T_j| > N$ then

      Set $N = |T_j|$.

      Set $l = j$.

    end

  end

  Set $F = F \cup l$.

  for $i \in (B \setminus C)$ do

    if $F \cap P_i \geq k_i$ then

      Set $C = C \cup i$.

    end

  end

end

5.1.3 Heuristic Approach for the General Case. Although the two special cases considered earlier are instructive, the assumptions in these do not always hold. However, the integer programming approach (5) is unlikely to scale to large problems, especially when we have millions of queries to consider. We now present a general-purpose greedy heuristic that builds on the GKC algorithm. First, observe that in the general setting, there is no longer a meaningful notion of “cover,” since each query and individual have an associated specific contribution $\Delta_{ij}$. Yet, recall that $\Delta_{ij} = d_{ij}(B_j - A_j)$, and, consequently, if $j \in P_i$, then the marginal impact of flipping query $j$ on the LRT statistic of $i$ only depends on query $j$. Define $\Delta_j = (B_j - A_j)$ so that $\Delta_{ij} = d_{ij}\Delta_j$. For any subset of individuals $P \subseteq B$ and query $j$, let $T_j = \{i \in P | j \in P_i\}$ be the set of individuals for whom $j \in P_i$. We can then define the average marginal contribution of each query $j$ and population $P$ as

$$\bar{\Delta}_j(P) = \frac{|T_j|\Delta_j}{|P|}.$$

The greedy heuristic we propose iteratively chooses a query $j$ to flip with the largest marginal contribution $\bar{\Delta}_j(P)$, where the population $P$ consists of the individuals whose privacy has yet to be guaranteed. Note that for this heuristic to work reasonably well, it is crucial that $\Delta_j > 0$. This is indeed the case as shown in the proof of Proposition 4.1 (which implies that $B_j - A_j > 0$) if $\delta < 1/4$. This means that as we flip query responses, we cannot decrease the LRT score for any individual, and, consequently, any individual $i$ whose privacy is already protected remains protected. This heuristic, which we call MI Greedy (MIG), where MI stands for Marginal Impact, is formalized as Algorithm 3.
ALGORITHM 3: MI Greedy

Input: A set $B$ of individuals in the Beacon, a subset $P_i$ and $\eta_j$ for each individual, marginal contributions $\Delta_j$ for each query, a collection of queries $Q$, and a threshold $\theta$.

Output: Subsets of queries $F \subseteq S$ to flip.

Initialization: $F = \emptyset$, $C = \emptyset$.

while $(B \setminus C) \neq \emptyset$ do
  Set $l = 1$, $N = -1$.
  for $j \in (Q \setminus F)$ do
    Set $T_j = \{i \in (B \setminus C) | j \in P_i\}$.
    Set $\bar{\Delta}_j = \frac{\Delta_j |T_j|}{|B \setminus C|}$.
    if $\bar{\Delta}_j > N$ then
      Set $N = \bar{\Delta}_j$.
      Set $l = j$.
    end
  end
  Set $F = F \cup l$.
  for $i \in (B \setminus C)$ do
    if $\sum_{j \in F} \Delta_{ij} + \eta_i \geq \theta$ then
      Set $C = C \cup i$.
    end
  end
end

5.2 Adaptive Attacks

The threat model articulated so far has assumed that the attacker fixes the decision threshold $\theta$ prior to executing any queries, and $\theta$ is independent of queries. However, since our defense involves the alteration of query responses, an adaptive attacker should make use of query responses in identifying an appropriate threshold. In other words, the adaptive attacker chooses an inference threshold based on a maximum allowable false positive rate, given the defense. Limiting the false positive rate ensures that the attack is reasonable (i.e., avoids cases like claiming that all individuals or no individuals are in the Beacon). From the perspective of privacy protection, this means that it is not sufficient to ensure that LRT statistics for all individuals exceed some predefined threshold, but we must actually ensure that the Beacon and non-Beacon populations are well mixed in terms of the respective LRT statistics as calculated based on the modified query responses. Analogously, this can be interpreted as minimizing the area under the ROC curve for attacker performance. We now formalize this idea.

Consider our encoding $y$ of which query responses to flip, and let $\bar{B}$ denote a set of individual genomes not in the Beacon (e.g., a data sample of these from the general population). The LRT statistic for each individual $i \in \bar{B}$ can be computed just as for any $i \in B$. Let us take $K$ individuals from $\bar{B}$ with the lowest LRT statistics, denoting the set of these individuals by $\bar{B}^{(K)}$. The concrete instantiation of the adaptive threat model then uses the following threshold:

$$\theta(Q) = \frac{1}{K} \sum_{k \in \bar{B}^{(K)}} \left( \sum_{j \in Q} \Delta_{kj} y_j + \eta_k \right).$$

We can interpret this as representing an attacker’s tolerance for false positives. For example, if the distribution of LRT scores is approximately symmetric around the mean, then $K/(2|\bar{B}|)$ is approximately the false positive rate. As a result, the privacy constraint for each $i \in B$ in the adaptive
attack setting becomes
\[
\sum_{j \in Q_1} \Delta_{ij} y_j + \eta_i \geq \frac{1}{K} \sum_{k \in B} \left( \sum_{j \in Q_1} \Delta_{kj} y_j + \eta_k \right) = \sum_{j \in Q_1} \sum_{k \in B} \left( \frac{\Delta_{kj}}{K} \right) y_j + \sum_{k \in B} \left( \frac{\eta_k}{K} \right).
\]
Define
\[
\Delta_j^{(K)} = \sum_{k \in B(K)} \frac{\Delta_{kj}}{K} \quad \text{and} \quad \eta^{(K)} = \sum_{k \in B(K)} \frac{\eta_k}{K}.
\]
Rewriting the preceding expression, we then obtain the following privacy condition for \(i \in B\):
\[
\sum_{j \in Q_1} \left( \Delta_{ij} - \Delta_j^{(K)} \right) y_j + \eta_i \geq \eta^{(K)}.
\] (6)
Finally, by defining \(\Delta_{ij}^{(K)} = \Delta_{ij} - \Delta_j^{(K)}\), we can rewrite this in the form identical to Equation (4) for the fixed threshold attacks:
\[
\sum_{j \in Q_1} \Delta_{ij}^{(K)} y_j + \eta_i \geq \eta^{(K)}.
\] (7)
Superficially, this suggests that we can directly apply the methods developed in Section 5.1 for privacy protection against fixed-threshold attacks. Additionally, indeed, we can directly incorporate the linear privacy constraint (7) into the linear integer program (5). However, this threat model now breaks the greedy algorithms we previously proposed. The first reason is that \(\delta\) now figures as a part of the threshold \(\theta(Q)\) and, consequently, is embedded in \(\Delta_j^{(K)}\) in two potentially conflicting ways. The second (and related) issue is that although a fixed-threshold threat model implied, for \(\delta < 0.25\), that \(\Delta_{ij} > 0\), this is clearly no longer necessarily the case for \(\Delta_{ij} - \Delta_j^{(K)}\). This has two consequences: (1) greedily adding one query \(j\) to the flip set \(F\) may actually cause privacy violation of an individual whose privacy constraint was previously satisfied, and (2) the integer linear program (5) may no longer have a feasible solution even though it is feasible for a fixed \(\theta\).
In practice, this means that the choice of \(K\) cannot be overly conservative. Moreover, to enable us to directly reuse the general-purpose greedy algorithm from Section 5.1 for privacy protection against the adaptive threat model, we only consider flipping queries \(j\) for which \(\Delta_{ij}^{(K)} \geq 0\) for all \(i \in B\).

6 THE ONLINE SETTING

Thus far, we assumed that the attacker submits all queries \(Q\) all at once, computes LRT statistics, and decides which individuals to make membership claims about. In practice, however, queries to the Beacon arrive over time, and privacy violations may arise even inadvertently if the attacker is, say, a relative of an individual in the Beacon who happens to observe that a rare collection of minor alleles that their family member possesses is also in the Beacon. Since individual queries may increase as well as decrease LRT statistics, it may well be the case that queries flipped in anticipation of a batch attack—even with \(Q = S\)—nevertheless violate privacy for some query sequences. Consequently, in the online setting, we need to assure Beacon service privacy for subsets of queries.

However, note that in practice we may not need to be concerned about arbitrary subsets of queries: since it is optimal from an attacker’s perspective to make use of all query responses they have observed, we need only guarantee privacy for the subset of queries submitted by any user...
thus far—in other words, of course, if we know which queries the user submitted. This issue of whether or not the Beacon service knows which queries have previously been submitted by a user motivates a natural distinction between two classes of online use settings that we discussed in Section 3: authenticated access, where the Beacon knows all prior queries (i.e., access requires authentication and identity is carefully verified), and unauthenticated access, where the Beacon does not have this information. Next, we formalize the online query setting, and subsequently consider in turn authenticated and unauthenticated access.

6.1 A Model of the Online Beacon

The online query setting is characterized by a sequence of \( T \) queries \( \{q_1, \ldots, q_T\} \), with \( q_t \in S \) denoting a \( t \)th query about a particular SNV (we alternatively refer to this as a query at time \( t \), with time here being equivalent to the order in the query sequence). Similarly, at each point in time, including \( t = 0 \) (i.e., before any queries), the Beacon can decide to flip a subset of query responses \( F_t \). In this setting, the set of all queries flipped is \( F = \cup_t F_t \); however, in the online setting, we need not flip them all at once. The reason we may choose to defer flipping a particular query response is that observed queries are informative, and a particular observed query sequence may warrant flipping many fewer responses than, say, a worst-case sequence or the batch of all queries \( S \). There is an additional constraint that we must impose on \( F_t \): query responses are commitments, in the sense that if at any point \( t \) we choose to honestly respond to a query \( j \), we must do so in the future; similarly, if we chose to flip the query response, we must do so in the future as well. Given this constraint, we assume without loss of generality that the query sequence is non-repeating—that is, for all \( 1 \leq t, t' \leq T, q_t \neq q_{t'} \) (since future identical queries are responded to exactly as the first time they are encountered).

At time \( 1 \leq t \leq T \), we have a collection \( Q_{t-1} \) of past queries, along with the query \( q_t \) that just arrived, resulting in the query set \( Q_t \) observed thus far. A privacy guarantee now entails that privacy of no individual \( i \in B \) is violated at any time \( t \). For a fixed-threshold threat model, this translates into the following privacy condition:

\[
L_i(Q_t, x) \geq \theta(Q_t).
\]

As we observed in Section 5.2, the condition has an analogous form for adaptive attacks. Since we are in the online setting, we can now choose subsets of queries to flip over time rather than all at once. We can encode the associated decisions \( F_t \) as binary vectors \( y_t \), resulting in the following privacy condition:

\[
L_i(Q_t, x, y_t) \equiv \sum_{j \in Q_t} \Delta_{ij} y_{j,t} + \eta_i(Q_t) \geq \theta(Q_t),
\]

where we now make it explicit that \( \eta_i \) in the modified LRT statistics depends on the query set \( Q_t \).

6.2 Authenticated Access

The key feature of authenticated access settings that we can leverage is the knowledge at any time \( t \) of the prior queries \( Q_{t-1} \) as well as the current query \( q_t \) to which the Beacon is about to respond (effectively, assuming that there is no collusion among Beacon clients, unlike in the following unauthenticated setting). The following proposition makes the intuitive observation that in the authenticated access setting, one never needs to make a decision whether to flip a query or not at time \( t \) for any \( j \neq q_t \).

**Proposition 6.1.** For any \( 1 \leq t \leq T \), there is an optimal query flip policy with \( F_t \subseteq \{q_t\} \).

This follows because if you wish to flip a particular query \( j \), you need not implement this decision until you actually observe the query.

ACM Transactions on Privacy and Security, Vol. 26, No. 3, Article 42. Publication date: July 2023.
6.2.1 Fixed-Threshold Attacks. We begin in the authenticated setting by again considering the fixed-threshold attacks. For this setting, our assumptions imply that \( \Delta_j = B_j - A_j > 0 \) for all \( j \). Consequently, if responding honestly to the query \( q_t \) would violate privacy, we would always wish to flip the response. This is captured in the following proposition.

**Proposition 6.2.** In the fixed-threshold threat model and authenticated access setting, if \( \exists i \in B \) such that \( L_i(Q_t, x, y) < \theta \), then \( F_t = \{q_t\} \).

Proposition 6.2 suggests a simple online heuristic for ensuring privacy while minimizing the number of query flips: flip \( j \) if and only if \( q_t = j \) and adding \( q_t \) violates privacy. This is formalized in Algorithm 4. Our following experiments demonstrate that this simple heuristic is remarkably effective in practice. We observe that although this heuristic may not be optimal, it does guarantee privacy in this setting under the reasonable assumption that \( \theta \leq 0 \) (otherwise, privacy is impossible, due to the fact that the constraint is violated even before any queries are made).

**ALGORITHM 4: Online Greedy**

**Input:** A set \( B \) of individuals in the Beacon, previous queries \( Q \), history of which queries have been flipped previously \( y \), the incoming query \( q \) and threshold \( \theta \).

**Output:** Decision whether or not to flip query \( q \).

if \( \exists i \in B : L_i(Q \cup q, x, y) < \theta \) then
  return Yes
end
return No

**Proposition 6.3.** Suppose that \( \theta \leq 0 \). Then the **Online Greedy** (OG) algorithm guarantees privacy against fixed-threshold attacks in the online authenticated access setting.

**Proof.** We prove this by induction. For the base case, note that privacy is guaranteed at \( t = 0 \) since \( L_i(\emptyset, x, y) = 0 \geq \theta \) for all \( i \) and \( \theta \leq 0 \). Next, suppose that \( L_i(Q_{t-1}, x, y_{t-1}) \geq \theta \) and \( L_i(Q_t, x, y_{t-1}) \geq \theta \)—that is, we need not flip the response to the current query \( q_t \)—privacy is not violated at time \( t \). Suppose that \( L_i(Q_t, x, y_{t-1}) < \theta \), which means that we flip the response to query \( q_t \) in the OG algorithm. Let \( y_{j,t} = y_{j,t-1} \) for all \( j \neq q_t \) and \( y_{j,t} = 1 \) for \( j = q_t \). Then

\[
L_i(Q_t, x, y_t) = \sum_{j \in Q_{t-1}} \Delta_{ij} y_{j,t} + \eta_i(q_t)
\]

\[
= \sum_{j \in Q_{t-1}} \Delta_{ij} y_{j,t-1} + \eta_i(Q_{t-1}) + \eta_i(q_t) + \Delta_{i,q_t}
\]

Now, \( \eta_i(q_t) = d_{i,q_t}(x_q A_{q_t} + (1 - x_{q_t}) B_{q_t}) \), whereas \( \Delta_{i,q_t} = d_{i,q_t}(B_{q_t} - A_{q_t}) \). Moreover, recall that if \( \delta < 0.25 \), \( B_j > 0 \) and \( A_j < 0 \) for all queries \( j \). Since \( L_i(Q_t, x, y_{t-1}) < \theta \), it must be that \( d_{i,q_t} = 1 \), since otherwise \( \eta_i(q_t) = 0 \), and \( x_{q_t} = 1 \), since otherwise \( \eta_i(q_t) > 0 \). Thus, \( \eta_i(q_t) = A_{q_t} \) and \( \Delta_{i,q_t} = B_{q_t} - A_{q_t} \). Consequently, \( \eta_i(q_t) + \Delta_{i,q_t} = A_{q_t} + B_{q_t} - A_{q_t} = B_{q_t} > 0 \geq \theta \). Since this holds for every individual \( i \) and time \( t \), privacy against threshold attacks is guaranteed for all individuals and query sequences. \( \square \)

6.2.2 Adaptive Attacks. As before, adaptive attacks complicate things considerably, but we can nevertheless leverage the algorithmic idea developed for fixed-threshold attacks. In the case of adaptive attacks, recall that the privacy condition for each \( i \in B \) becomes

\[
\sum_{j \in Q_{t-1}} \Delta_{ij}^{(K)} y_j + \eta_i(Q_t) \geq \eta_i^{(K)}(Q_t),
\]
where we now make the dependence of $\eta_i(Q_t)$ and $\eta^{(K)}(Q_t)$ explicit. Note that we can still apply the preceding OG algorithm, but with an important change: now, both $\eta_i(Q_t)$ and $\eta^{(K)}(Q_t)$ must be updated after receiving each query $q$. Modulo this change, the algorithm, upon observing a query $q$, checks whether $\exists i \in B$ such that $\sum_{j \in Q_t} \Delta_{ij}^{(K)} y_j + \eta_i(Q_t) < \eta^{(K)}(Q_t)$, flips $q$ if this is true, and does not otherwise.

The crucial issue, however, is that we can no longer guarantee privacy in this setting, since flipping a query $q$ may now actually cause the privacy condition for some other individual to be violated. However, in our following experiments, we show that our populations remain well mixed in terms of LRT statistics (for which the adaptive privacy condition is a proxy).

### 6.3 Unauthenticated Access

The key distinction between authenticated and unauthenticated access in our model is that in the latter case the Beacon does not know which queries have previously been made when it receives a new query $q$ at any given point in time. We therefore model this setting by assuming that the query sequence (besides $q$) is adversarial. Specifically, the privacy constraint now takes the following form:

$$\forall i, t, \min_{Q_t \subseteq Q_{t-1}} L_i(Q_t \cup q_t, x, y_t) - \theta(Q_t) \geq 0.$$  (8)

We use $Q_t$ to emphasize that since we do not know the past query sequence and wish to protect the privacy of every $i \in B$, we are assuming that the sequence of queries is independently adversarial for each $i$. We now show that in the unauthenticated setting, the temporal aspect collapses, and the optimal decision about which queries to flip can be made at time $t = 0$.

**Proposition 6.4.** In the unauthenticated access setting, if all $j \in S$ are queried by some finite time $t$ and there exists a solution to the BEACON-PRIVACY-PROBLEM, then there is an optimal solution with the property that $F = F_0$ and $F_t = \emptyset$ for all $t > 0$.

**Proof Sketch.** For a sufficiently large $t$, $Q_{t-1} = S$. As such, $Q_t \cup q_t = Q_t$. Then, it must be true that $F_t = \emptyset$ and $F = \cup_{t < t} F_t$. Since $F$ must guarantee privacy for all queries at time $t' \geq t$, it must be a minimal set of queries to do so, and we can simply identify such a set at $t = 0$. □

This means that the unauthenticated online access setting is effectively a worst-case batch setting, where the worst case set of queries is chosen independently for each $i \in B$. Note that this proposition appears to contradict Proposition 6.1, but in fact it does not, as neither claims that the optimal solution it characterizes is unique. In this case, too, we can wait to implement the flips in $F$ until the associated queries are actually observed for the first time.

The consequence of Proposition 6.4 is that we can simplify somewhat the definition of privacy in the unauthenticated setting:

$$\forall i, \min_{Q_t \subseteq S} L_i(Q_t, x, y) - \theta(Q_t) \geq 0.$$  (9)

#### 6.3.1 Fixed-Threshold Attacks

Recall that $P_i(Q) = \{j \in Q_i | d_{ij} = 1\}$. While we previously omitted the dependence of $P_0$ on $Q$, this must be explicit in the online setting. The next proposition shows that in the case of fixed-threshold attacks, the privacy condition reduces to a particularly simple form.

**Proposition 6.5.** In the unauthenticated access setting with fixed-threshold attacks, the privacy condition (9) is equivalent to

$$\forall i, \sum_{j \in P_i(S) \setminus F} A_j \geq \theta.$$  (10)
Proof. Fix $i \in B$. We begin by unpacking the LRT score resulting from a flipping strategy $y$ in Equation (9) (since $\theta$ is fixed, that is the only thing affected by the choices of queries):

$$\min_{Q_i \subseteq S} L_i(Q_i, x, y) = \sum_{j \in Q_i} \Delta_{ij} y_j + \sum_{j \in Q_i} d_{ij} A_j + \sum_{j \in Q_{i,0}} d_{ij} B_j.$$ 

First, observe that since $B_j > 0$ by our assumption that $\delta < 0.25$ and from (the proof of the) Proposition 4.1, query set $S_0$ (i.e., those with $x_j = 0$) would not be included, since they can only increase the LRT statistic. Similarly, none of the queries with $d_{ij} = 0$ will be included since these do not contribute to the LRT statistic. Consequently, $Q_i \subseteq P_i(S)$. Moreover, since $B_j > A_j$ under the same assumptions, none of the terms with $y_j = 1$ are included. Consequently, $Q_i \subseteq P_i(S) \setminus F$. Moreover, since $A_j < 0$, all queries in $P_i(S) \setminus F$ will be included.

An important implication of Proposition 6.5 is that in this setting, flipping queries is equivalent to masking them. The reason is that since flipping increases LRT statistics, the worst-case subset of queries will never include any queries that have been flipped, effectively masking all of them.

As a consequence of Proposition 6.5, we can represent the solution to the BEACON-PRIVACY-PROBLEM in this setting as the following integer linear program:

$$\min_{y} \sum_{j} y_{j} \quad \text{subject to:} \quad \sum_{j \in P_i} |A_j| y_j \geq \theta - \sum_{j \in P_i} A_j,$$

where $|A_j|$ refers to the absolute value of $A_j$. Moreover, in the special case that AAFs follow the beta distribution and we use their expectations, we can make direct use of the methods from Section 5.1.2, including the GKC algorithm (with $k_i = \theta + |P_i|$), where $|P_i|$ is the size of the set $P_i$ (slightly overloading notation). Similarly, even in the general case, we can leverage the heuristic algorithm in Section 5.1.3, replacing $\Delta_j$ with $|A_j|$. Finally, we observe that in the special case $\theta = 0$, there is only one feasible solution, which is $F = \cup_i P_i$. However, this solution is always feasible (but not necessarily optimal) if $\theta \leq 0$, and we can thus always guarantee privacy in such a setting for fixed-threshold attacks.

6.3.2 Adaptive Attacks. Recall that even in the batch setting, since $\Delta_{ij}^{(K)}$ may be negative for some $i, j$, privacy constraints may be violated for some individuals if we flip certain queries $j$. Since in the unauthenticated setting we are making decisions up front, we only consider the subset of queries $j$ for which $\Delta_{ij}^{(K)} \geq 0$ for all $i$.

Unpacking the condition in Equation (9) and rearranging terms, we obtain the following privacy condition for adaptive attacks for each individual $i \in B$:

$$\min_{Q_{i,1} \subseteq S_i} \sum_{j \in Q_{i,1}} (\Delta_{ij}^{(K)} y_j + d_{ij}^{(K)} A_j) + \min_{Q_{i,0} \subseteq S_0} d_{ij}^{(K)} B_j \geq 0,$$

where $d_{ij}^{(K)} = d_{ij} - \sum_{k \in B(K)} d_{kj}/K$. Since the second term on the left-hand side does not depend on $y$ (equivalently, $F$), we can precompute it, setting $k_i = -\min_{Q_{i,0}} d_{ij}^{(K)} B_j$. Consequently, we obtain the condition

$$\min_{Q_{i,1}} \sum_{j \in Q_{i,1}} (\Delta_{ij}^{(K)} y_j + d_{ij}^{(K)} A_j) \geq k_i.$$

We now use this expression to obtain a variant of the MIG heuristic for this setting. The key idea behind this heuristic was to choose a query $j$ to flip that has the highest average marginal impact in each iteration (omitting individuals previously “covered” in the sense that their privacy is satisfied). Since we only consider flipping queries with $\Delta_{ij}^{(K)} \geq 0$, this will not have a detrimental impact on any such “covered” individuals, as it can only increase their LRT statistics. For any $i$ not
yet covered, define $\mu_{ij}$ to be the marginal impact of flipping a query $j$. If $\Delta_{ij}^{(K)} + d_{ij}^{(K)}A_j \geq 0$, this query will be omitted as a result of the flip, and the marginal contribution is thus $\mu_{ij} = |d_{ij}^{(K)}A_j|$. If, however, $\Delta_{ij}^{(K)} + d_{ij}^{(K)}A_j < 0$, this query will remain, but its contribution will be reduced by $\Delta_{ij}^{(K)}$ and the marginal contribution is therefore $\mu_{ij} = \Delta_{ij}^{(K)}$ as in the batch setting.

7 EXPERIMENTS

7.1 Experiment Setup

7.1.1 Dataset. The dataset used in this work was originally made available by the organizers of the 2016 iDash Privacy and Security Workshop [25] as part of their Practical Protection of Genomic Data Sharing Through Beacon Services challenge. The goal of the challenge was for teams to develop computational approaches that release as many truthful responses as possible through a modified Beacon before the Shringarpure-Bustamante attack [22] could be used to re-identify an individual. In this study, we use SNVs from chromosome 10 for a subset of 400 individuals to construct the Beacon and another 400 individuals excluded from the Beacon. Unless otherwise specified, we set the genomic sequencing error rate to $\delta = 10^{-6}$, as in the iDash challenge.

7.1.2 Computational Environment. Experiments were carried out on a PC with an AMD Ryzen 7 3800x processor and 64 GB DDR4 3600 MHz-CL19 RAM running Ubuntu version 18.04.5, using Python version 3.6.12.

7.1.3 Baselines. We compare our approaches to three state-of-the-art baselines. The first is Strategic Flipping (SF) [30], which is the winning entry to the 2016 iDash Privacy Challenge and uses a combination of greedy and local search. We compare two versions of this approach: first, the version as previously implemented (SF), and second, a variant, SF-M in the adaptive settings that uses our definition of privacy instead of setting a static threshold using a maximum allowable false positive rate. Note that SF and SF-M are equivalent in the fixed-threshold setting. The second baseline is Random Flipping (RF) proposed by Raisaro et al. [17]. RF randomly flips a subset of unique alleles in the Beacon dataset by sampling from a binomial distribution. The third baseline we consider is DP as proposed for this setting by Cho et al. [7]. These baselines are configured so that in the fixed-threshold batch setting they maximize utility within their respective parameter configuration space while guaranteeing privacy, which is defined as the fraction of individuals in the Beacon for whom the privacy constraint under the respective threat model is satisfied. As privacy protection for all individuals was not always guaranteed by the baseline methods in the adaptive threshold case, we present two sets of results: (1) each baseline is tuned to be the best achievable privacy, then additional results are presented for parameters that are less conservative to understand the relative impact on utility, and (2) tuning all methods to a similar level of utility to the best extent possible and comparing the relative privacy achieved using ROC curves.

7.1.4 Additional Adaptive Attacks. Although one of our threat models is explicitly adaptive, it does not consider the full scope of attack adaptivity that can be leveraged. We therefore consider two additional adaptive attacks in our evaluation: allele inference attack, which attempts to infer which queries were flipped by considering cross-SNV correlations as in prior work [4, 28], and mimicry attack, which attempts to infer which queries were flipped by simulating our defense on synthetic data samples.

In the allele inference attack, an adversary leverages correlations among SNVs to infer which SNVs might have been flipped by our approaches. A common measure of correlation between SNVs that is leveraged in past variations of this attack is Linkage Disequilibrium (LD) [23]. A bi-allelic
SNV is a position on the genome for which two possible alleles are seen across the population. For a pair of bi-allelic SNVs (with alleles \{A,a\} and \{B,b\}, respectively), the LD is defined by

\[
LD = P(AB) - P(A)P(B),
\]

where \(P(AB)\) is the frequency with which \(A\) and \(B\) occur together, and \(P(A)\) and \(P(B)\) are the individual allele frequencies for \(A\) and \(B\), respectively. The value of \(LD\) lies between \(-0.25\) and \(0.25\), with a positive value indicating higher-than-random association of the two alleles. We note that all SNVs considered in this study are bi-allelic. The allele inference attack proceeds as follows. Upon receiving a “no” response from the Beacon for an SNV \(j\), the attacker calculates LD values for SNV \(j\) paired with \(l\) neighboring SNVs on either side of \(j\) on the genomic sequence, based on the fact that alleles close to each other on the genomic sequence are more likely to be correlated [20]. If the maximum LD thus calculated for any pair \((i,j)\) lies above a certain threshold \(t_{LD}\), and the Beacon response for SNV \(i\) is “yes,” the attacker flips the Beacon response for SNV \(j\) to “yes.”

In the mimicry attack, the attacker attempts to infer the SNVs that might have been flipped by the proposed approaches by simulating them either on (1) similar public datasets or (2) synthetic datasets constructed using knowledge of AAFs. To simulate this, we generate 10 synthetic populations and corresponding simulated Beacons using the AAFs from our original iDash dataset. Each synthetic population also consists of 400 individuals in the Beacon and 400 individuals not in the Beacon. To construct the synthetic populations, we generate genomic sequences where SNVs are considered to be independent, and a minor allele exists at position \(j\) for each individual with probability equal to the AAF for the \(j\)th position from the iDash dataset. The attack proceeds as follows. Our privacy-preserving algorithms corresponding to each threat model described previously are used on these 10 synthetic datasets to obtain a probability value \(p_j\) that SNV \(j\) will be flipped. In the batch setting, the attacker proceeds to flip Beacon responses for the top \(R\) queries in terms of \(p_j\) from “no” to “yes” if the Beacon response is indeed “no” to begin with. In our experiments, we use \(l = 5\) and \(t_{LD} = 0.2\), which limits the impact of the attack to highly correlated pairs. In the online setting, we assume that upon receiving a “no” response for a SNV \(j\), the attacker subsequently queries neighbors to calculate LD, if they have not already been queried.

In the mimetic attack, the attacker attempts to infer the SNVs that might have been flipped by the proposed approaches by simulating them either on (1) similar public datasets or (2) synthetic datasets constructed using knowledge of AAFs. To simulate this, we generate 10 synthetic populations and corresponding simulated Beacons using the AAFs from our original iDash dataset. Each synthetic population also consists of 400 individuals in the Beacon and 400 individuals not in the Beacon. To construct the synthetic populations, we generate genomic sequences where SNVs are considered to be independent, and a minor allele exists at position \(j\) for each individual with probability equal to the AAF for the \(j\)th position from the iDash dataset. The attack proceeds as follows. Our privacy-preserving algorithms corresponding to each threat model described previously are used on these 10 synthetic datasets to obtain a probability value \(p_j\) that SNV \(j\) will be flipped. In the batch setting, the attacker proceeds to flip Beacon responses for the top \(R\) queries in terms of \(p_j\) from “no” to “yes” if the Beacon response is indeed “no” to begin with. In our experiments, we evaluate the effect of flipping the top 5% and top 10% of query responses in terms of \(p_j\). In the online setting, as queries are made one at a time, computing the top \(R\) queries in terms of \(p_j\) is not possible until all SNVs are queried. Instead, the attacker flips the Beacon response for a SNV \(j\) if the original Beacon response is “no” and \(p_j\) is above a chosen threshold \(t_p\). We evaluate the effect of this attack for \(t_p\) set to 0.5, 0.7, and 0.9.

### 7.2 The Batch Setting

#### 7.2.1 Fixed-Threshold Attacks with a Small Sequencing Error

In our first set of experiments, we consider a setting where the sequencing error is negligible. In these experiments only, we set \(\delta = 10^{-240}\), which is sufficiently small that it suffices to flip a single Beacon response per individual to guarantee privacy in this setting. Although a sequencing error this small is unrealistic in modern genomic sequencing, the associated results offer an instructive starting point. In this setting, because flipping queries when the minor allele is very frequent is likely to degrade the trust in the system, we consider the effect of a restriction on the rarity of occurrence of the alternate allele on the number of flips needed to secure privacy.

Figure 2(a) compares the number of flipped queries between the proposed GMBC and the three baselines. We can see that GMBC allows us to guarantee privacy with significantly (more than an order of magnitude) fewer false Beacon responses compared to the baselines. The suboptimality of SF stems from not accounting for how many individuals a SNV affects, and only looking at the average over the population, a limitation that GMBC overcomes.
7.2.2 Fixed-Threshold Attacks with AAFs Drawn from Beta Distribution. Next, we consider the setting with a static prediction threshold $\theta$, where the AAFs are assumed to be drawn from a beta distribution. Recall that in this setting, it suffices to flip $k_i$ SNVs per individual. Once again, we present results comparing to the three baselines, now varying the value of $\theta$. From here on, we limit ourselves to experiments where there are no restrictions on how frequently an alternate allele can be present in a population, due to the much larger compute times needed to handle multiple high-precision values arising from the AAFs for more than 1.3 million SNVs. Henceforth, we also set $\delta = 10^{-6}$. Figure 2(b) presents the results comparing the proposed GKC approach to the baselines for $\theta \in [-2,000 \sim -2,000]$. Again, we see that the proposed GKC algorithm again flips orders of magnitude fewer SNVs compared to the alternatives while guaranteeing privacy.

7.2.3 General Case: True AAFs. Next, we look at the more general case with no assumptions on AAFs, a realistic $\delta = 10^{-6}$, and an adversary who computes a static threshold $\theta$ based on some prior knowledge. This setting is much more representative of a real-world attack. Figure 3(a) compares the number of queries flipped MIG to the baselines, over a range of prediction thresholds $\theta$. In this figure, also note that for the DP and RF baselines, for each value of $\theta$, we present the performance with an empirically selected parameter ($\epsilon$ and $p$, respectively) that yields the highest utility while preserving privacy for all individuals, with the corresponding parameter denoted in the plot. We can observe that MIG again outperforms both RF and DP by several orders of magnitude in terms of utility (all approaches preserve privacy of all individuals in the Beacon dataset). SF is closer to MIG but still flips considerably more queries.

7.2.4 Adaptive Attacks. Although the algorithms devised with a fixed $\theta$ in mind guarantee privacy given this assumption, adaptive attackers can defeat these approaches by taking the revised Beacon queries explicitly into account when determining the threshold. This is illustrated in Figure 3(b), which shows the LRT scores for 400 individuals in the Beacon and 400 others not in the Beacon after flipping responses using MIG for $\theta = 0$. Although the LRT scores for the individuals in the Beacon do end up above 0, they also remain below the scores computed on those not in the Beacon. Additionally, even though the attacker does not in fact know who is in the Beacon, a clustering attack can separate the two populations, albeit with a high false positive rate. Specifically, using one-dimensional $k$-means clustering of the LRT scores achieves a 100% true positive rate at the somewhat substantial cost of a 30% false positive rate, averaged over 20 runs. Next, we evaluate the effectiveness of the proposed algorithms that aim to explicitly account for this more sophisticated attack.
Fig. 3. Comparing utility and LRT scores in the fixed threshold batch setting with general AAFs.

Fig. 4. Comparing utility, privacy, and LRT scores in the adaptive attack batch setting.

Figure 4(a) shows the number of flips that need to be flipped by our adaptive attack variant of MIG, as well as by the various baselines in this setting. The values of $\epsilon$ and $p$ used for DP and RF, respectively, are shown in parentheses in the plot legend. MIG again flips orders of magnitude fewer SNVs than either RF or DP, as well as SF-M for higher values of $K$ (the size of the non-Beacon LRT comparison group discussed in Section 5.2). SF is more competitive and actually flips fewer SNVs than MIG when $K$ is higher, but, as we will see presently, it offers very poor privacy in this setting.

Figure 4(b) shows privacy (as a function of $K$ with respect to the definition of privacy in Section 5.2). Although MIG and SF-M preserve privacy of all individuals in the Beacon in this setting, both DP (for $\epsilon \in \{0.1, 0.5, 1\}$) and SF fail to achieve privacy for all individuals. Figure 4(c) illustrates that explicitly accounting for adaptive attacks, MIG yields LRT scores that are much more mixed between individuals in and not in the Beacon dataset than if we are to assume fixed-threshold attacks. Quantitatively, setting $K = 20$ increases the false positive rate for the clustering attack from 30% to greater than 50% on average.

7.2.5 ROC Curves for Adaptive Attack: Batch Setting. Figure 5(a) presents ROC curves for the adaptive attack in the batch setting. Note that in this case, a lower area under the curve is better, as the ROC curve corresponds to attack success. Also note that the maximum false positive rate
up until which the area under the ROC curve (AUC) remains zero for MIG and SF-M corresponds to the percentage of the total population for which the solution is computed (20% of individuals not in the Beacon corresponds to 10% of the total population), beyond which the AUC is non-zero for all approaches (the plot line corresponding to SF-M has a slight non-zero slope between FPR = 0.1 and FPR = 0.2 in Figure 5(a)).

Figure 5(b) presents ROC curves for the adaptive batch setting, comparing MIG to the various baselines when all methods are tuned to have similar utility in terms of the number of SNVs flipped. Parameters for RF and DP were chosen using the expected utility, computed over an average of five runs. Here, we see that MIG significantly outperforms all baselines—illustrating that to achieve a similar level of privacy to MIG, the baselines must suffer a much greater utility loss.

7.3 The Online Setting

7.3.1 Authenticated Access. Recall that in the authenticated setting, the defender has access to each user’s query history, and thus a decision about whether to flip the Beacon response for a SNV can be greedily made at runtime. Unlike previous settings where all methods were able to achieve perfect privacy for all individuals in the Beacon, this will no longer always be the case in the online setting. Consequently, we also compare our methods with the baselines in terms of privacy, defined as the fraction of the individuals in the Beacon whose privacy is not violated.

First, we consider fixed-threshold attacks. Figure 6 compares OG with the baselines when $\theta = 0$ (the worst case threshold in the online setting). Recall that OG provably achieves privacy in such settings but does flip more queries than SF (but far fewer than other baselines). In contrast, SF compromises privacy of a considerable fraction of individuals in the Beacon (as do other baselines) when few SNVs have been queried. In Figure 7, we consider adaptive attacks. When $K = 1$, OG tends to have better privacy but considerably lower utility than SF and SF-M (and dominates the other two baselines in both). For $K = 10$, however, OG has better utility than all baselines but SF but slightly lower privacy than SF-M (and better than others). Although SF achieves the highest utility, it has extremely poor privacy.

7.3.2 ROC Curves for Adaptive Attack: Authenticated Online Setting. We compare the performance of the various methods using ROC curves in the authenticated online setting with the adaptive threshold model. Figure 8(a), 8(b), and 8(c) present ROC curves when 100,000, 500,000, 1,000,000.
Fig. 6. Comparing utility and privacy in the authenticated online setting with fixed-threshold attacks; $\theta = 0$.

Fig. 7. Comparing utility and privacy in the authenticated online setting with adaptive attacks.

When $K = 1$, OG outperforms all baselines except DP with $\epsilon = 0.1$ when 100,000 SNVs are queried, and DP with $\epsilon = 0.5$ and $\epsilon = 0.1$ when the number of SNVs queried is increased to 500,000 and 1.3 million. However, DP flips a significantly larger number of SNVs when compared to MIG, as can be observed in Figure 7(a). Also note that at $K = 1$, DP does violate privacy more often than MIG, as can be seen in Figure 7(c). When $K = 10$, OG outperforms all baselines, except
DP with $\epsilon = 0.1$, although the performance of the two approaches is closer in this case when compared to the case where $K = 1$. The performance of SF-M also shows significant improvement relative to the $K = 1$ case.

In contrast, Figure 9 presents results in the authenticated online setting, where all methods are tuned to have similar utility. Here, as opposed to the results in Figure 8, we observe that for both $K = 1$ and $K = 10$, OG significantly outperforms all baselines, achieving far greater privacy for the same utility loss.

### 7.3.3 Unauthenticated Access

Finally, we compare the proposed approaches to baselines in the unauthenticated online setting. Once again, we begin with fixed-threshold attacks. As shown in Figure 10, OMIG (our algorithm variant for this setting) flips more queries than SF but does guarantee privacy, whereas SF compromises the privacy of a subset of individuals, particularly as more SNVs can be queried. The other two baselines also achieve privacy, although at a considerable loss in utility compared to OMIG.

Figure 11 presents a similar comparison for adaptive attacks. In this setting, all methods including OMIG now compromise privacy, with DP doing so the least. However, OMIG is now again orders of magnitude better than most of the baselines in terms of utility (with SF-M now performing relatively poorly in terms of both utility and privacy). Although SF performs similarly to OMIG in terms of privacy, it has slightly lower utility.

### 7.3.4 ROC Curves for Adaptive Attack: Unauthenticated Online Setting

Figure 12 presents ROC curves comparing the performance of the various baselines to our variant (OMIG) in the
unauthenticated online setting with an adaptive threshold attack, when 100,000, 500,000, and 1.3 million SNVs are queried. In all three cases, DP outperforms OMIG, and SF achieves very similar performance to OMIG. However, both DP and SF provide lower utility when compared to OMIG, as can be seen from Figure 11. In this setting, RF is seen to perform better for lower values of the probability \( p \), in contrast to other settings. Figure 13 shows that in the unauthenticated online setting, all methods perform comparably when tuned for similar utility.

Fig. 9. ROC curves in the authenticated online setting. All methods are tuned for similar utility.

Fig. 10. Comparing utility and privacy in the online unauthenticated fixed-threshold attack setting; \( \theta = -1,000 \).
7.4 Allele Inference Attack

Next, we evaluate the impact of the allele inference attack on the proposed methods as well as the different baselines. Figure 14(a) presents results in the fixed threshold batch setting, where all methods originally guaranteed privacy. It can be seen that the attack has very little impact on any
of the methods, with privacy remaining above 94% in all cases. Empirically selected best parameters (in terms of utility while guaranteeing privacy for all) used for DP and RF for the various values of \( \theta \) are highlighted in the plot in the corresponding color.

Figure 14(b) presents results for the adaptive threshold batch setting. The attack has no impact on DP and RF, and thus the plot shows only original privacy performance for these baselines. For SF, SF-M, and MIG, we present original performance as well as performance after the allele inference attack, with plot lines corresponding to the latter denoted by “(LD)” in the legend. Yet again, the impact of the allele inference attack is minimal for all methods, including the proposed approach.

Finally, we note that the allele inference attack had no noticeable impact on any of the methods in the online settings. In the case of our OG algorithm, this can be attributed to the fact that the SNVs flipped in the online setting are different from the ones flipped in the batch setting. The decision to flip each SNV depends on the query sequence in case of authenticated access. In the case of unauthenticated access, the SNVs flipped do not have enough highly correlated neighbors to impact privacy. The effect of allele inference on the different baselines is negligible when averaged over multiple random query sequences in the online settings.

7.5 Mimicry Attack

Finally, we present the impact of mimicry attacks using simulated datasets, starting with the batch setting using a fixed-threshold threat model. Figure 15(a) presents a privacy comparison in this setting, when the top 5% and 10% of query responses from the synthetic beacons in terms of flipping probability \( p_j \) are flipped to “yes.” Much like allele inference, the mimicry attack has very little impact on any of the methods, with privacy remaining above 96%.

Figure 15(b) similarly compares privacy achieved by the various methods after being subjected to the mimicry attack, when the top 5% and 10% of query responses are flipped. Only the proposed approach is affected, and although the associated reduction in privacy is now tangible, it is still quite small, with MIG remaining comparable to DP in terms of privacy, and competitive with SF-M.

Next, we present results for the online setting with authenticated access. Recall that in the online setting, the attack flips the Beacon response for SNV \( j \) to “yes” if the original response is “no” and the probability \( p_j \) of SNV \( j \) being flipped computed over the synthetic data is above threshold \( t_p \). For the fixed-threshold threat model, the attack is seen to have no impact on performance. Figure 16 presents a privacy comparison across the various methods for \( t_p \) set to 0.5, 0.7, and 0.9 for the adaptive threshold setting when \( K = 1 \) and \( K = 10 \). Comparing to the privacy originally achieved
by these methods as presented in Figure 7, we observe that the mimicry attack has a nominal impact on SF and RF, but other methods (including our approach) are essentially unaffected. The performance does not vary significantly across the various values of $t_p$. Finally, we note that the attack has no impact on performance in the unauthenticated online setting.

8  RELATED WORK

Privacy Violation of Shared Genomic Data. As genotyping and sequencing costs began falling in the early 2000s, it became evident that large amounts of genomic data would be collected in clinical and, particularly, research settings. To ensure that such information was widely
disseminated, various programs were instituted to streamline the collection and redistribution of such data. One such example is the Database of Genotypes and Phenotypes, which was established by the National Institutes of Health to support the mandate of the 2007 Genome-Wide Data Sharing Policy [1]. Although access to individual-level records required review and approval, public open access of summary statistics about SNV rates was initiated to provide insight into the information in such datasets. Shortly thereafter, it was shown by Homer et al. [14] that it was indeed possible to determine if individuals contributed to a mixture of DNA (i.e., the aggregated summary statistics)—even when the individual’s contribution to the mixture is as little as 1% by comparing allele frequencies obtained from probe intensities, to the allele frequencies of a reference population such as from the International HapMap Project [10]. This and other types of attacks on genomic data are addressed at greater length in the study by Elrich and Narayanan [9]. Since the Shringarpure-Bustamante attack on Beacon services was introduced in 2015 [22], there have been several refinements to this class of attack [4, 5, 17, 28], including methods to account for the effects of kinship (i.e., having multiple family members in the Beacon) and genomic reconstruction using inherent correlations between SNVs.

**Protecting the Privacy of Genomic Data.** Various defenses have been proposed against attacks on genomic data. These approaches typically rely on some degree of masking through noise injection or suppression of a subset of the data. The efficacy of masking a subset of shared genomic data was studied in the work of Sankararaman et al. [21]. More recently, preserving privacy was translated into a game-theoretic perspective, which considered how to account for the capabilities of an attacker [31]. Various genomic privacy-preserving methods have been summarized by Bonomi et al. [6]. Specific to Beacon services, Raisaro et al. [17] present an RF heuristic that perturbs unique alleles in the database using a binomial distribution, as well as a query-budget approach for authenticated users. A DP approach was adopted by Cho et al. [7], which forms one of the baselines in this article. The winning entry [30] to the 2016 iDash Privacy and Security Workshop challenge defines a differential discriminative power to select SNVs from the dataset for which responses are flipped.

In 2022, a new specification for genomic Beacons, named Beacon v2, was proposed [18] with a three-tiered user authentication model as a security measure. A user’s access to the Beacon’s data is limited by the tier at which it is accessed—anonymous, registered, or controlled. Data at the anonymous tier of classification may be accessed by anyone, as a guest user. Accessing data at the registered tier requires identification by means of a sign-up process using personal credentials. Accessing data at the controlled tier requires the user to specifically apply, and be granted access to do so. The specification, however, makes no note of the various membership inference attacks proposed over the years against version 1. Several reviews of various threats to privacy in the context of genomic data, technical and legal defenses currently in use, and privacy-preserving machine learning techniques were also presented in 2022 [12, 15, 29]. Another technique that is gaining popularity is the use of homomorphic encryption and blockchain technology to enable the use of cloud services or datasets distributed across several sites for genotype computations without sending unencrypted private genomic information over the network [3, 11, 27].

**9 DISCUSSION AND LIMITATIONS**

We presented a novel framework for privacy-preserving design of Beacons in the context of membership inference attacks leveraging an LRT statistic. Our framework precisely dissects the many ways in which the Beacon service can be configured and used, such as allowing queries as a batch or in a sequence, and allowing authenticated access to individuals whose identities can be verified, or simply opening the service to the public. We also considered two distinct threat models, one of which has been explicitly studied in prior literature, whereas the second involves a stronger
Defending Against Membership Inference Attacks on Beacon Services

adaptive attack and has not been formally defined or analyzed in prior work. We presented polynomial-time, highly scalable algorithms that exhibit privacy guarantees for some of these instantiations of our model, and in one special case a provable approximation of optimal utility (while guaranteeing privacy). Moreover, the proposed algorithms typically outperform prior art in privacy (against LRT-based membership inference attacks), utility, or both. Finally, we also showed that our approach is largely unaffected by allele inference and mimicry attacks that try to infer which SNVs are flipped to preserve privacy.

Our approach has several limitations. First, our privacy model is specific to the Beacon service and the LRT-based attack; it is possible that other attacks can be devised that can defeat our approach, although we are not aware of any existing attacks that do. Second, flipping query responses, although common in prior art, is not always a viable means to protect the Beacon service (e.g., it may degrade public trust in the service). An alternative framework of masking a subset of SNV queries may offer another practical solution without this limitation, but this may in turn result in an even greater degradation of the utility of the Beacon service. Finally, a far greater challenge is to determine practically plausible adversaries, to parameterize potential solutions for realistic adversarial models.

REFERENCES

[1] NIH. 2007. Not-OD-07-088: Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). NIH. Retrieved June 13, 2023 from https://grants.nih.gov/grants/guide/notice-files/not-od-07-088.html#publication.

[2] Md. Momin Al Aziz, Reza Ghasemi, Md. Waliullah, and Noman Mohammed. 2017. Aftermath of Bustamante attack on genomic Beacon service. BMC Medical Genomics 10, 2 (2017), 43–54.

[3] Mohammed Alghazwi, Fatih Turkmen, Joeri Van Der Velde, and Dimka Karastoyanova. 2022. Blockchain for genomics: A systematic literature review. Distributed Ledger Technologies: Research and Practice 1, 2 (2022), 1–28.

[4] Kerem Ayoz, Erman Ayday, and A. Ercument Cicek. 2021. Genome reconstruction attacks against genomic data-sharing beacons. Proceedings on Privacy Enhancing Technologies 3 (2021), 28–48.

[5] Kerem Ayoz, Miray Aysen, Erman Ayday, and A. Ercument Cicek. 2020. The effect of kinship in re-identification attacks against genomic data sharing beacons. Bioinformatics 36, Suppl. 2 (2020), i903–i910.

[6] Luca Bonomi, Yingxiang Huang, and Lucila Ohno-Machado. 2020. Privacy challenges and research opportunities for genomic data sharing. Nature Genetics 52, 7 (2020), 646–654.

[7] Hyunghoon Cho, Sean Simmons, Ryan Kim, and Bonnie Berger. 2020. Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. Cell Systems 10, 5 (2020), 408–416.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Theory of Cryptography Conference. 265–284.

[9] Yaniv Erlich and Arvind Narayanan. 2014. Routes for breaching and protecting genetic privacy. Nature Reviews Genetics 15, 6 (2014), 409–421.

[10] Richard A. Gibbs, John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli L. Yu, Huanming Yang, Lan-Yang Ch’ang, et al. 2003. The International HapMap Project. Nature 426 (2003), 789–796.

[11] Gamze Gürsoy, Eduardo Chielle, Charlotte M. Brannon, Michail Maniatakos, and Mark Gerstein. 2022. Privacy-preserving genotype imputation with fully homomorphic encryption. Cell Systems 13, 2 (2022), 173–182.

[12] Gamze Gürsoy, Tianxiao Li, Susanna Liu, Eric Ni, Charlotte M. Brannon, and Mark B. Gerstein. 2022. Functional genomics data: Privacy risk assessment and technological mitigation. Nature Reviews Genetics 23, 4 (2022), 245–258.

[13] Inken Hagestedt, Yang Zhang, Mathias Humbert, Pascal Berrang, Haixu Tang, XiaoFeng Wang, and Michael Backes. 2019. MBeacon: Privacy-preserving beacons for DNA methylation data. In Proceedings of the Network and Distributed Systems Security Symposium.

[14] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics 4, 8 (2008), e1000167.

[15] Wonsuk Kim and Junhee Seok. 2022. Privacy-preserving collaborative machine learning in biomedical applications. In Proceedings of the 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC’22). IEEE, Los Alamitos, CA, 179–183.

[16] Bartha M. Knoppers. 2014. International ethics harmonization and the Global Alliance for Genomics and Health. Genome Medicine 6, 2 (2014), Article 13, 3 pages.
[17] Jean Louis Raisaro, Florian Tramer, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, et al. 2017. Addressing Beacon re-identification attacks: Quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association 24, 4 (2017), 799–805.

[18] Jordi Rambla, Michael Baudis, Roberto Ariosa, Tim Beck, Lauren A. Fromont, Arcadi Navarro, Rahel Paloots, et al. 2022. Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond. Human Mutation 43, 6 (2022), 791–799.

[19] Laura L. Rodriguez, Lisa D. Brooks, Judith H. Greenberg, and Eric D. Green. 2013. The complexities of genomic identifiability. Science 339, 6117 (2013), 275–276.

[20] Sahel Shariati Samani, Zhicong Huang, Erman Ayday, Mark Elliot, Jacques Fellay, Jean-Pierre Hubaux, and Zoltán Kutalik. 2015. Quantifying genomic privacy via inference attack with high-order SNV correlations. In Proceedings of the 2015 IEEE Security and Privacy Workshops. 32–40. https://doi.org/10.1109/SPW.2015.21

[21] Sriram Sankararaman, Guillaume Obozinski, Michael I. Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. Nature Genetics 41, 9 (2009), 965–967.

[22] Suyash S. Shringarpure and Carlos D. Bustamante. 2015. Privacy risks from genomic data-sharing beacons. American Journal of Human Genetics 97, 5 (2015), 631–646.

[23] Montgomery Slatkin. 2008. Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9, 6 (2008), 477–485.

[24] Petr Slavík. 1996. A tight analysis of the greedy algorithm for set cover. In Proceedings of the ACM Symposium on Theory of Computing. 435–441.

[25] Haixu Tang, Xiaofeng Wang, Shuang Wang, and Xiaojian Jiang. 2016. iDash Privacy and Security Workshop 2016. Retrieved June 13, 2023 from http://www.humangenomeprivacy.org/2016/.

[26] María Torres-Español, Seyed Yahya Anvar, and María-Jesús Sobrido. 2016. Variations in the genome: The Mutation Detection 2015 meeting on detection, genome sequencing, and interpretation. Human Mutation 37, 10 (2016), 1106–1109.

[27] Leon Visscher, Mohammed Alghazwi, Dimka Karastoyanova, and Fatih Turkm en. 2022. Poster: Privacy-preserving genome analysis using verifiable off-chain computation. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 3475–3477.

[28] Nora Von Thenen, Erman Ayday, and A. Ercument Cicek. 2019. Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics 35, 3 (2019), 365–371.

[29] Zhiyu Wan, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A. Malin. 2022. Sociotechnical safeguards for genomic data privacy. Nature Reviews Genetics 23, 7 (2022), 429–445.

[30] Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley Malin. 2017. Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services. BMC Medical Genomics 10, 2 (2017), 87–100.

[31] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xie, Ellen Wright Clayton, Murat Kantarcioglu, and Bradley Malin. 2017. Expanding access to large-scale genomic data while promoting privacy: A game theoretic approach. American Journal of Human Genetics 100, 2 (2017), 316–322.

[32] Carol J. Weil, Leah E. Mechanic, Tiffany Green, Christopher Kinsinger, Nicole C. Lockhart, Stefanie A. Nelson, Laura L. Rodriguez, and Laura D. Buccini. 2013. NCI think tank concerning the identifiability of biospecimens and “omic” data. Genetics in Medicine 15, 12 (2013), 997–1003.

Received 15 March 2022; revised 7 April 2023; accepted 4 June 2023