What If, But Privately: Private Counterfactual Retrieval

Shreya Meel, *Graduate Student Member, IEEE*, Mohamed Nomeir, *Graduate Student Member, IEEE*, Pasan Dissanayake, *Graduate Student Member, IEEE*, Sanghamitra Dutta, *Member, IEEE*, and Sennur Ulukus, *Fellow, IEEE*

***Abstract*—Transparency and explainability are two important aspects to be considered when employing black-box machine learning models in high-stake applications. Providing counterfactual explanations is one way of catering this requirement. However, this also poses a threat to the privacy of the institution that is providing the explanation, as well as the user who is requesting it. In this work, we are primarily concerned with the user's privacy who wants to retrieve a counterfactual instance, without revealing their feature vector to the institution. Our framework retrieves the exact nearest neighbor counterfactual explanation from a database of accepted points while achieving perfect, information-theoretic privacy for the user. First, we introduce the problem of private counterfactual retrieval (PCR) and propose a baseline PCR scheme that keeps the user's feature vector information-theoretically private from the institution. Building on this, we propose two other schemes that reduce the amount of information leaked about the institution database to the user, compared to the baseline scheme. Second, we relax the assumption of mutability of all features, and consider the setting of immutable PCR (I-PCR). Here, the user retrieves the nearest counterfactual without altering a private subset of their features, which constitutes the immutable set, while keeping their feature vector and immutable set private from the institution. For this, we propose two schemes that preserve the user's privacy information-theoretically, but ensure varying degrees of database privacy. Third, we extend our PCR and I-PCR schemes to incorporate user's preference on transforming their attributes, so that a more actionable explanation can be received. Finally, we present numerical results to support our theoretical findings, and compare the database leakage of the proposed schemes.**

***Index Terms*—Counterfactual explanations, information-theoretic privacy, private information retrieval.**

I. Introduction

THE right to explanations mandates that any black-box machine learning model, making crucial decisions in high-stakes applications should provide the user, i.e., the applicant, with a suitable explanation for its decision [1]. In this regard, counterfactual explanations have grown as an effective means to deliver the minimum perturbation required at the user's end to alter the model's decision to a favorable outcome [2]. For instance, in the case of a bank loan rejection, a user might receive a counterfactual recommendation to increase their income by 10K to get accepted. Numerous works have focused on generating counterfactuals with different properties, namely, proximity to the user's input [3], robustness to model changes [4], [5], [6], [7], feasibility under user's constraints [8], [9], sparsity in the attributes to be changed [3], [9], and diversity of the counterfactuals [9]. We refer the reader to [10], [11], and [12] for a comprehensive survey on different methods. In this work, we introduce a framework that provides information-theoretic guarantees.

Providing an appropriate counterfactual poses serious privacy concerns, both for the user asking for a counterfactual, and for the institution delivering it. Existing works such as [13], [14], [15], and [16] focus on preserving data privacy from the institution's side, while [17], [18], [19] focus on the extraction of the model by querying for multiple counterfactual explanations. However, preserving privacy from the user's side has rarely received attention. We are particularly interested in the scenario where a user would like to obtain a counterfactual explanation without revealing their input feature vector to the institution. The user may be reluctant to share their feature vector with the institution for several reasons, e.g., if they have a limited number of attempts to apply, or if they wish to preserve their data privacy until they improve their chances of acceptance.

A. Related Works

Since the initial formulation in [2], numerous works have been focusing on generating counterfactual explanations with different properties. Proximity to the query instance [3], robustness [4], [5], actionability [8], [9], sparsity in change [3], [9], and diversity [9] are some such properties; we refer the reader to [11] and [20] for a comprehensive survey on different methods. In this work, we focus on both proximity to the original instance and actionability. Further, we introduce immutability as a constraint where the user may keep some of their feature values fixed, while obtaining a counterfactual. Moreover, we use nearest-neighbor counterfactuals as the explanation method considered. Existing works on privacy within the context of counterfactual explanations mainly focus on the institution's end. In this regard, [13] analyzes inferring

Received 7 July 2025; revised 3 January 2026; accepted 22 March 2026. Date of publication 8 April 2026; date of current version 27 October 2026. An earlier version of this paper was presented in part at the 2025 IEEE International Symposium on Information Theory and in part at the 2025 Annual Allerton Conference [DOI: 10.1109/ISIT63088.2025.11195552]. *(Corresponding author: Sennur Ulukus.)*

The authors are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: smeel@umd.edu; mnomeir@umd.edu; pasand@umd.edu; sanghamd@umd.edu; ulukus@umd.edu).

Communicated by P. Sadeghi, Associate Editor for Coding Techniques.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2026.3681883.

Digital Object Identifier 10.1109/TIT.2026.3681883

0018-9448 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

the membership of an explanation in the training set of the model while [14] provides differentially private counterfactuals. References [17] and [18] present two ways of utilizing counterfactuals to extract the model when counterfactuals are provided for any query. Reference [19] presents a model extraction strategy that specifically leverages the fact that the counterfactuals are close to the decision boundary. Explanation linkage attacks that try to extract private attributes of a nearest-neighbor counterfactual explanation are discussed in [15]. All of these works focus on the privacy of either the model or the data stored in the institution's database. In contrast, we are interested in the privacy of the applicant who is asking for an explanation.

Closely related to our problem is the nearest neighbor search problem, where the user needs to retrieve the indices of the vectors in a database, that are closest to their vector according to some similarity metric. In this regard, [21] proposes algorithms that guarantee computational privacy, both to the user and the database, while the user retrieves an approximate nearest neighbor. Reference [22] proposes an information-theoretically private clustering-based solution based on the dot-product metric. This work, however, does not consider database privacy, and the user retrieves only an approximate nearest neighbor. The work leverages techniques from private information retrieval (PIR) [23], which is a subject of independent interest, to obtain the index of the nearest counterfactual.

In PIR, a user wishes to retrieve a message out of K replicated messages in N servers without leaking any information about their required message index. The capacity of PIR, i.e., the maximum ratio between the number of the required message symbols and the number of total downloaded symbols, is found in [24]. In symmetric PIR (SPIR), an extra requirement is considered where the user cannot get any information about the other messages aside from their required message. The capacity for such a model is found in [25]. Other important variants can be found in [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], and [48]. Reference [39], proposes a cross subspace alignment (CSA) approach as a unifying framework for PIR and SPIR with additional requirements, such as security against the storing databases. These schemes are capacity-achieving in some cases, for instance, when the number of messages is large. We refer the reader to [49] for a comprehensive survey on the PIR and SPIR literature.

B. Our Contributions

We summarize our contributions as follows.

- First, we introduce the novel problem of private counterfactual retrieval (PCR), along with a baseline scheme to achieve user privacy to retrieve the index of the closest counterfactual using the ℓ_2 distance metric (Theorem 2). Then, we develop two different PCR schemes that we call Diff-PCR and Mask-PCR to provide the institution with better privacy for their database compared to the baseline scheme while achieving perfect information-theoretic privacy for the applicant (Theorem 3 and Theorem 4).
- Next, we extend our PCR formulation to a more practical setting, where we allow the user to fix a subset of their features as *immutable*, i.e., their values should not be altered in their corresponding counterfactual. The immutable set is user-specific and should be kept private from the institution. Towards this, we propose two immutable PCR (I-PCR) schemes (Theorem 5 and Theorem 6), where the user retrieves their counterfactual index from the institution under the immutability constraint.
- Further, we account for *user actionability* which guarantees preferential change of certain attributes compared to the rest. For both PCR and I-PCR, we demonstrate that our schemes can be modified to include private user actionability on the mutable attributes without compromising the privacy of the immutable attributes. We call these extended schemes PCR+ (Theorem 7) and I-PCR+ (Theorem 8).
- We quantify the database leakage for the proposed PCR and I-PCR schemes using the mutual information measure, compare the leakage between the schemes analytically, and verify the comparative leakage numerically. To provide an operational meaning to the leakage values, we numerically evaluate the leakage using both the mutual information leakage and the maximal leakage measures.

Our framework of privately retrieving the index of the closest neighbor guarantees that the institution does not learn any information about the user's input. Further, the institution does not learn the counterfactual explanation corresponding to the user either. This is achieved in two steps: First, we perform the PCR scheme to retrieve the counterfactual index from a set of non-colluding servers, which store the set of accepted samples through replication [50], [51]. Next, we use this index to perform symmetric private information retrieval (SPIR) [25] on the same indexed dataset. In SPIR, the user obtains their counterfactual explanation, without revealing its index to the server, while the server incurs no further leakage on the stored data other than the required record. The SPIR scheme applies seamlessly in the existing system, due to the replicated and non-colluding databases.

C. Paper Organization

The rest of the paper is organized as follows. Section II formally introduces the problem. Section III and Section IV describe the proposed schemes for the PCR and I-PCR settings, respectively. In Section V, we extend our schemes to include private user actionability. In Section VI, we present our numerical results. Finally, Section VII concludes our paper.

D. Notation

Throughout the paper, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. For all integers, $m < n$ we use $[m : n]$ to denote the set $\{m, m+1, \ldots, n-1, n\}$. Given a set A, we represent its cardinality by $|A|$. For a vector a, we write a^T to denote its transpose. We use $\mathbb{F}_q$ to represent a finite field, where q is a prime or a prime power. The functions $H(\cdot)$ and $I(\cdot;\cdot)$ represent the Shannon entropy and mutual information, respectively. Further, we use bold, italicized font $\boldsymbol{A}$ to denote a matrix.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

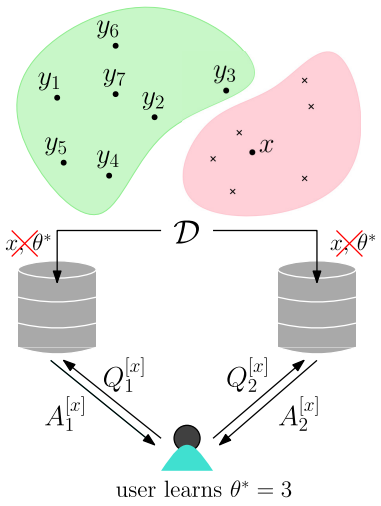


Fig. 1. PCR system model: The green and red regions represent the accepted and rejected samples, respectively. The user learns that $\theta^* = 3$ is their counterfactual index in $\mathcal{D}$.

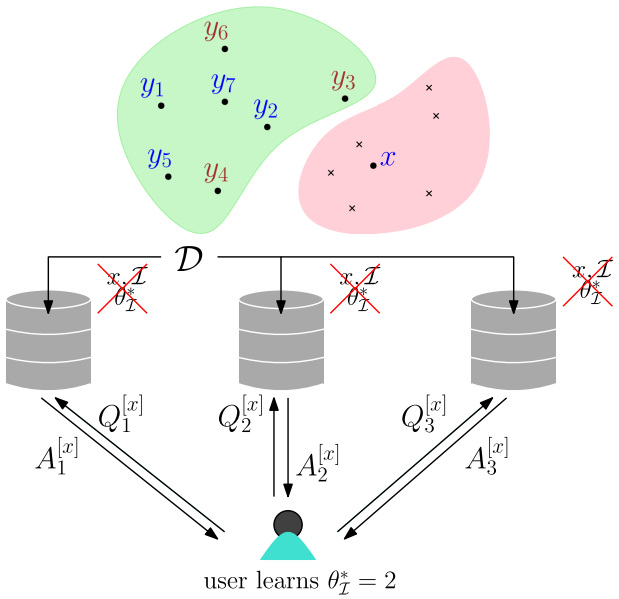


Fig. 2. I-PCR system model: Here, y_1, y_2, y_5, y_7 have the same values for the immutable features as x. The user learns that $\theta^*_{\mathcal{I}} = 2$ is their counterfactual index in $\mathcal{D}$.

II. Problem Formulation

The institution has a pre-trained binary classification model that takes a d-dimensional feature vector as input and classifies it into its target class, e.g., accepted or rejected. A user who is rejected by this model wishes to privately retrieve a valid counterfactual sample corresponding to their data sample. The user, who does not have access to the model, relies on a database $\mathcal{D}$ containing the feature vectors of a set of samples accepted by the model.

We assume that each attribute (feature) of the samples (feature vectors) is an integer in $[0 : R]$. The samples in $\mathcal{D}$ are stored in N non-colluding and non-communicating servers in a replicated manner, and are indexed as $y_1, y_2, \ldots, y_M$ where $M = |\mathcal{D}|$. Under this setup, the user retrieves the index of the accepted data sample in the database closest to their own feature vector, x, in terms of a distance metric $d(\cdot, \cdot)$. In all sections, except Section V, we assume the metric to be the ℓ_2 norm, i.e., $d(x, y) = \|y - x\|^2$. We refer to the problem without and with the immutability constraint as the private counterfactual retrieval (PCR) setting, and private counterfactual retrieval with immutable features (I-PCR) setting, respectively, which we formally describe next.

A. PCR Setting

Here, we assume that the user can change all their feature values to obtain their counterfactual, and the counterfactual index θ^* is given by

$$\theta^* = \arg\min_{i \in [M]} d(x, y_i), \tag{1}$$

as depicted in Fig. 1.

To this end, the user sends the query $Q_n^{[x]}$ to server $n \in [N]$. Upon receiving the queries, each server computes their answer $A_n^{[x]}$ using their storage, their queries and shared common randomness Z', i.e.,

$$H(A_n^{[x]} | \mathcal{D}, Q_n^{[x]}, Z') = 0. \tag{2}$$

Sharing of common randomness among the non-colluding servers can be realized by a trusted third-party who distributes this randomness or a random seed to all servers. For the latter, a synchronized pseudo-random number generator is implemented by the servers to produce the required randomness. Using the responses from all the servers, the user determines the index θ^* of their corresponding counterfactual, i.e.,

$$\text{[PCR Decodability]} \quad H(\theta^* | Q_{[N]}^{[x]}, A_{[N]}^{[x]}, x) = 0, \tag{3}$$

where $Q_{[N]}^{[x]}$ and $A_{[N]}^{[x]}$ denote all the sent queries, and all the received answers, respectively. Since user privacy is our primary focus, we impose that each server must not acquire any information on the user's sample or the index of their counterfactual, i.e., $\forall n \in [N]$,

$$\text{[PCR User Privacy]} \quad I(x, \theta^*; Q_n^{[x]} | \mathcal{D}) = 0, \tag{4}$$

where $I(\cdot; \cdot | \cdot)$ is the conditional mutual information. To quantify privacy for the servers, we consider a mutual information based metric that defines the amount of information leakage about the samples in $\mathcal{D}$ to the user, upon receiving the answers, as follows

$$\text{[PCR Database Leakage]} \quad I(y_1, \ldots, y_M; Q_{[N]}^{[x]}, A_{[N]}^{[x]} | x). \tag{5}$$

B. I-PCR Setting

In PCR with the immutability constraint, the goal of the user is to retrieve the index of the nearest sample which should differ from x in at most a subset $\mathcal{M} \in [d]$ of features. Alternatively, the user fixes their counterfactual's feature values in $\mathcal{I} = [d] \backslash \mathcal{M}$, which we call the *immutable set*. In particular, let $\mathcal{I} = \{j_1, j_2, \ldots, j_{|\mathcal{I}|}\}$ with $j_1 < j_2 < \ldots < j_{|\mathcal{I}|}$, then

$$x_{\mathcal{I}} = \left[x_{j_1}, x_{j_2}, \ldots, x_{j_{|\mathcal{I}|}}\right]^T, \tag{6}$$

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

TABLE I

SUMMARY OF THE PERFORMANCES OF THE PROPOSED SCHEME. WE SHOW COMMUNICATION COST, MINIMUM REQUIRED FIELD SIZE AND NUMBER OF SERVERS REQUIRED, TO REALIZE OUR PROPOSED SCHEMES. R IS THE MAXIMUM VALUE THAT EACH FEATURE IS QUANTIZED TO, d IS THE DIMENSION OF THE FEATURE VECTOR, M IS THE DATABASE SIZE, F IS THE MAXIMUM NUMBER OF IMMUTABLE FEATURES, L IS THE SCALING FACTOR TO ENFORCE IMMUTABILITY (SEE SECTION IV) AND L_1 IS THE LARGEST INTEGER OF THE WEIGHT VECTOR FOR ACTIONABILITY (SEE SECTION V)

Scheme	Communication cost	$\mathbb{F}_q$ size requirement
Baseline PCR ($N=2$)	$2d+2M$	$q > R^2 d$
Diff-PCR ($N=2$)	$2d+2M-2$	$q > 2R^2 d$
Mask-PCR ($N=2$)	$2d+2M$	$q > R^2 d$
Two-Phase I-PCR ($N=3$)	$9d+9M$	$q > R^2 d$
Single-Phase I-PCR ($N=3$)	$6d+3M$	$q > F(L-1)R^2 + R^2 d : L > R^2 d$
Baseline PCR+ ($N=3$)	$6d+3M$	$q > R^2 L_1 d$
Diff-PCR+ ($N=3$)	$6d+3M-3$	$q > 2R^2 L_1 d$
Mask-PCR+ ($N=3$)	$6d+3M$	$q > R^2 L_1 d$
Two-Phase I-PCR+ ($N=4$)	$14d+11M$	$q > R^2 L_1 d$
Single-Phase I-PCR+ ($N=3$)	$6d+3M$	$q > F(L-1)R^2 + R^2 L_1 d : L > R^2 L_1 d$

and $x_{\mathcal{M}}$ is defined in a similar way. Therefore, the goal of the user is to find the counterfactual index $\theta^*_{\mathcal{I}}$ given by

$$\theta^*_{\mathcal{I}} = \arg\min_{i\in\Theta} d(x, y_i) \quad \text{with} \quad \Theta = \{i \in [M] : y_{i,\mathcal{I}} = x_{\mathcal{I}}\}, \tag{7}$$

where the knowledge of $\mathcal{I}$, hence that of Θ, is restricted to the user and no information on x is revealed to the server. The system model is illustrated in Fig. 2. Since it is clear from the context, we use the same notations $Q_n^{[x]}$ and $A_n^{[x]}$ to denote the query and answer corresponding to the nth server for the I-PCR setting. The answers are deterministic functions of the database, queries and shared randomness as given in (2). The queries and answers need to satisfy the decodability, user privacy and database leakage constraints, as follows

$$\text{[I-PCR Decodability]} \; H(\theta^*_{\mathcal{I}} | Q_{[N]}^{[x]}, A_{[N]}^{[x]}, x, \mathcal{I}) = 0, \tag{8}$$

$$\text{[I-PCR User Privacy]} \; I(x, \theta^*_{\mathcal{I}}, \mathcal{I}; Q_n^{[x]}, A_n^{[x]} | \mathcal{D}) = 0, \tag{9}$$

$$\text{[I-PCR Database Leakage]} \; I(y_{[M]}; Q_{[N]}^{[x]}, A_{[N]}^{[x]} | x, \mathcal{I}). \tag{10}$$

Our goal is to develop schemes that reduce this database leakage, while maintaining perfect user privacy for both PCR and I-PCR settings. The performances of our proposed schemes, which will be analyzed in detail subsequently, are listed in Table I for convenience.

Before describing the schemes, we state the definition of a Vandermonde matrix, which is central to all our proposed schemes.

Definition 1 (Vandermonde Matrix): Let $\alpha_1, \ldots, \alpha_n$ be distinct elements of $\mathbb{F}_q \setminus \{0\}$. Then, a Vandermonde matrix of order n on $\mathbb{F}_q$ is denoted by $\boldsymbol{V}_n$, whose (i,j)th entry is given by

$$\boldsymbol{V}_n(i,j) = \alpha_i^{j-1}, \qquad j \in [n], \quad i \in [n]. \tag{11}$$

Since $\alpha_1, \ldots, \alpha_n$ are non-zero and distinct, $\boldsymbol{V}_n$ is invertible in $\mathbb{F}_q$.

In addition, for completeness, Shannon's one-time pad theorem, which we use extensively, is stated below as in [52].

Theorem 1 (One-Time Pad): For any random variable $X \in \mathbb{F}_q$, and any uniform random variable Z on $\mathbb{F}_q$, that is independent of X, the following holds

$$I(X; X+Z) = 0, \tag{12}$$

where addition is according to $\mathbb{F}_q$ arithmetic.

III. PCR SCHEMES

In this section, we propose schemes that provide feasible solutions to minimizing (5) subject to (3) and (4). For ease of exposition, we write

$$d_i(x) = \|y_i - x\|^2, \quad \forall i \in [M]. \tag{13}$$

First, we present the baseline scheme, followed by two approaches, difference of distances (Diff-PCR) and masking (Mask-PCR) that leak less information on $\mathcal{D}$, compared to the baseline, in addition to achieving user privacy. All PCR schemes require $N = 2$ replicated servers.

A. Baseline PCR

Theorem 2: There exists a scheme that can retrieve the exact closest counterfactual index from $N = 2$ servers with perfect user privacy, with the communication cost of $2(d+M)$ symbols of $\mathbb{F}_q$ where $q > R^2 d$ is prime.

To prove the achievability of Theorem 2, we present our baseline PCR scheme. Let the operating field be $\mathbb{F}_q$ where q is a prime satisfying $q > R^2 d$. Each server stores the M accepted samples, $y_1, \ldots, y_M$, each being a d dimensional vector. Let α_1, α_2 be two distinct elements of $\mathbb{F}_q$ globally known to the user and the servers.[1] The user privately generates a random vector Z uniformly from $\mathbb{F}_q^d$ and then sends the query $Q_n^{[x]}$ to the nth server based on x as

$$Q_n^{[x]} = x + \alpha_n Z, \quad n = 1, 2. \tag{14}$$

[1]The assumption that these parameters and the field size are known apriori to the user and servers is standard in the schemes for private information retrieval [39], computation [53], read-update-write [54], [55], and nearest-neighbor search [22].

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

Since x is one-time padded with $\alpha_n Z$ which is uniform on $\mathbb{F}_q^d$, each server learns no information about x, i.e., $I(x; Q_n^{[x]}|y_{[M]}) = 0$. Let the servers share M independent random variables $Z'(i), i \in [M]$ picked uniformly from $\mathbb{F}_q$. Given a query, the servers compute one answer for each $y_i \in \mathcal{D}$ as follows

$$A_n^{[x]}(i) = \|y_i - Q_n^{[x]}\|^2 + \alpha_n Z'(i) \tag{15}$$

$$= d_i(x) + \alpha_n \left(-2(y_i - x)^T Z + Z'(i)\right) + \alpha_n^2 \|Z\|^2. \tag{16}$$

Note that $A_n^{[x]}(i)$ is a second degree polynomial of α_n with $\alpha_n^2\|Z\|^2$ already known to the user. Therefore, the user cancels $\alpha_1^2\|Z\|^2$ and $\alpha_2^2\|Z\|^2$ from $A_1^{[x]}(i)$ and $A_2^{[x]}(i)$, respectively, to obtain

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \end{bmatrix} = \begin{bmatrix} A_1^{[x]}(i) - \alpha_1^2\|Z\|^2 \\ A_2^{[x]}(i) - \alpha_2^2\|Z\|^2 \end{bmatrix} \tag{17}$$

$$= \boldsymbol{V}_2 \begin{bmatrix} d_i(x) \\ -2(y_i - x)^T Z + Z'(i) \end{bmatrix}, \tag{18}$$

where $\boldsymbol{V}_2$ is the two-dimensional Vandermonde matrix given in this case as $\boldsymbol{V}_2 = \begin{bmatrix} 1 & \alpha_1 \\ 1 & \alpha_2 \end{bmatrix}$.

After multiplying (18) with $\boldsymbol{V}_2^{-1}$, the user obtains $d_i(x)$ for all $i \in [M]$. The user then compares the values of $d_i(x)$, for all $i \in [M]$ and assigns θ^* to the i for which this is minimum.

1) Communication Cost: This scheme requires two d-dimensional vectors of $\mathbb{F}_q$ to be sent, one to each server. This entails an upload cost of $2d$ symbols. Each server responds with M symbols from $\mathbb{F}_q$, thereby, incurring a download cost of $2M$. The communications for exchanging the global parameters (e.g., field size, α_1 and α_2) are not accounted for in the overall communication cost. This is because, i) they do not scale with the dimensionality d or number M of the accepted samples, and ii) they can be re-used over multiple instantiations of the scheme.

2) Computation Complexity: At the user side, the computation for the queries involves two scalar multiplications and additions in d dimensions, which is $O(d)$. To find θ^*, the user decodes $d_i(x)$ using (18) for each $i \in [M]$ and compares them to find the minimum, both using $O(M)$ operations. Therefore, the user-side computation complexity is $O(d+M)$. Each server generates M answers, where each answer involves subtraction, squaring, and addition of d elements of each vector. This results in $O(Md)$ operations at each server.

Remark 1: Note that the distance metric in the system model and all proposed schemes use the ℓ_2 norm. The system model and the schemes can be extended to any ℓ_k norm, where k is even by modifying the answer generation accordingly. We demonstrate this for the baseline PCR scheme in Appendix A.

B. Diff-PCR

Theorem 3: There exists a scheme employing $N = 2$ servers with a lower value of database leakage (5) compared to the baseline PCR, with the communication cost of $2(d + M - 1)$ symbols of $\mathbb{F}_q$ where $q > 2R^2d$ is prime.

The main idea of the scheme is to improve database privacy by revealing only the difference of the norms, as opposed to all the norms, while maintaining the user's privacy. We show that this can be accomplished with $N = 2$ replicated databases as in the baseline PCR. The field of operation is a prime $q > 2R^2d$. This is required because for all $i, j \in [M]$,

$$0 \le |d_i(x) - d_j(x)| \le R^2 d. \tag{19}$$

Therefore,

$$d_i(x) - d_j(x) \in \begin{cases} [0 : R^2 d], & d_i(x) \ge d_j(x), \\ [R^2 d + 1 : q - 1], & d_i(x) < d_j(x). \end{cases} \tag{20}$$

The servers share a common random vector $Z'_{[M-1]} = [Z'(1), Z'(2), \ldots, Z'(M-1)]^T$, where each entry is picked uniformly and independently from $\mathbb{F}_q$. As described in Section III-A, the user sends the query given in (14) to server $n = 1, 2$. Then, server n computes the following answer for each $i \in [M-1]$,

$$A_n^{[x]}(i) = \|y_i - Q_n^{[x]}\|^2 - \|y_{i+1} - Q_n^{[x]}\|^2 + \alpha_n Z'(i) \tag{21}$$

$$= \|y_i - x\|^2 - \|y_{i+1} - x\|^2 + \alpha_n \left(2(y_{i+1} - x)^T Z - 2(y_i - x)^T Z + Z'(i)\right) \tag{22}$$

$$= d_i(x) - d_{i+1}(x) + \alpha_n I'(i), \tag{23}$$

where $I'(i) = 2(y_{i+1} - y_i)^T Z + Z'(i)$. Using $A_1^{[x]}(i)$ and $A_2^{[x]}(i)$, the user exactly recovers $\|y_i - x\|^2 - \|y_{i+1} - x\|^2$, because

$$\begin{bmatrix} A_1^{[x]}(i) \\ A_2^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_2 \begin{bmatrix} d_i(x) - d_{i+1}(x) \\ I'(i) \end{bmatrix}. \tag{24}$$

From here, the user recovers the $M - 1$ differences along with $M - 1$ interference terms. In each $I'(i)$, the one-time-padding with $Z'(i)$ makes sure that no information on $y_{i+1} - y_i$ is revealed. The user finds the index θ^* of their counterfactual using Algorithm 1.

Algorithm 1 Algorithm to Compute θ^*

Input: $d_i(x) - d_{i+1}(x), \quad i \in [M-1]$
Output: θ^*
$\theta^* = 1$
for $i \in [M-1]$ **do**
  $r(i) = d_i(x) - d_{i+1}(x)$
  $d_{\theta^*}(x) - d_{i+1}(x) = \sum_{j=\theta^*}^{i} r(j)$
  if $d_{\theta^*}(x) - d_{i+1}(x) \in [R^2 d + 1 : q - 1]$ **then**
    $\theta^* \leftarrow \theta^*$
  else
    $\theta^* \leftarrow i + 1$
return θ^*

1) Communication Cost: The upload cost incurred in this scheme is $2d$ and the download cost is $2(M-1)$.

2) Computation Complexity: The user side query generation is $O(d)$ as in the baseline scheme. To find θ^*, the user finds $d_i(x) - d_{i+1}(x)$ using (24) and evaluates θ^* using Algorithm 1, which has a complexity of $O(M)$. Thus, the user's computation complexity is $O(d+M)$. The server side complexity is $O(Md)$, same as the baseline.

C. Mask-PCR

Theorem 4: There exists a scheme employing $N = 2$ servers that has a lower database leakage in terms of (5) compared to

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

the baseline PCR with the communication cost of $2(d+M)$ symbols of $\mathbb{F}_q$ where $q > R^2 d$ is prime.

In this approach, we require the servers to have access to the set of rejected samples, $\mathcal{D}_r = \{x_1, \ldots, x_K\}$ (this restriction is removed during the experimental analysis). Now, define the closure of $\mathcal{D}_r$ as follows

$$\mathcal{D}_c = \texttt{clo}(\mathcal{D}_r) \tag{25}$$

$$= \{x \in \mathbb{F}_q^d : |d_i(x) - d_j(x)| - |d_i(x_k) - d_j(x_k)| \geq 0, \quad \forall\ i, j, k\} \setminus \mathcal{D}. \tag{26}$$

In addition, we define the following metrics

$$d_k = \min_{i,j \in [M]} |d_i(x_k) - d_j(x_k)|, \qquad d_{\min} = \min_{k \in [K]} d_k. \tag{27}$$

Let μ be a random variable with support $\{0, \ldots, d_{\min} - 1\}$. Now, a user who wishes to know the closest accepted sample to their rejected sample $x \in \mathcal{D}_c$ sends the query in (14) to the nth server. Upon receiving the query, each server computes its answer for each $i \in [M]$ as follows

$$A_n^{[x]}(i) = \|y_i - Q_n^{[x]}\|^2 + \mu(i) + \alpha_n Z'(i), \tag{28}$$

where $\mu(i)$ has the same distribution as μ and $Z'(i)$ is a uniform random variable in $\mathbb{F}_q$. As the user receives the answers from the servers, they are reprocessed according to (17) to decode the distances as follows

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_2 \begin{bmatrix} d_i(x) + \mu(i) \\ I(i) + Z'(i) \end{bmatrix}. \tag{29}$$

Upon getting the values of the masked distances, $d_i(x) + \mu(i)$, $i \in [M]$, the user utilizes them to decide which is closest based on their numerical value, i.e., decode the index of the closest accepted sample. To show that the user can correctly decode with the masked distances, we need the following lemma.

Lemma 1: If $x \in \mathcal{D}_c$, then the user is able to decode the index of the closest acceptable sample.

Here, in contrast to the Diff-PCR approach, the field size does not need to be expanded. The main reason for this is that the difference calculations in this case are done at the user side and therefore the field size restriction can be dropped. The proof of Lemma 1 is provided in Appendix B.

Lemma 1 states that all the vectors in $\mathcal{D}_c$ preserve the relative distances among the accepted samples, i.e., the ordering of the distances among the points in $\mathcal{D}_c$ and accepted samples is maintained. In addition, note that the user does not know the exact value of $d_{\min}$ since they do not have any prior knowledge of the accepted samples.

1) Communication Cost: The upload cost is $2d$ and the download cost is $2M$.

2) Computation Complexity: The computation complexity at the user and the server is equal to that of the baseline PCR, i.e., $O(d+M)$ and $O(Md)$, respectively.

Remark 2: It is important to note that (27) is evaluated prior to the scheme initiation and done once only till the dataset itself changes. This is a part of pre-processing for Mask-PCR and is not a part of the main scheme, which begins with the user sending their query. This is why it is not included in the calculation of the computation complexity.

Remark 3: We believe that the requirement $x \in \mathcal{D}_c$ is a bit strong, which is why we dropped this requirement in our first experiment (see Sec VI-A) and we find $d_{\min}$ empirically instead. However, it is required for theoretical analysis to guarantee decodability.

3) Effect of Field Size in Masking: The field size can have a significant role in Mask-PCR. Recall that the minimum requirement for the field size is $R^2 d$ as explained at the beginning of this section. Let q_1 be the field size satisfying $q_1 > R^2 d$. If another field size q_2 is chosen such that $q_2 > q_1$, we can embed the samples $y_1, \ldots, y_M$ in $\mathbb{F}_{q_2}$ using a relative distance preserving transform $T : \mathbb{F}_{q_1} \to \mathbb{F}_{q_2}$ such that the following condition is satisfied for $k \in [K]$, $\quad i, j \in [M]$,

$$|d_i(x_k) - d_j(x_k)| \leq |\big(\|x_k - T(y_i)\|^2 - \|x_k - T(y_j)\|^2\big)|. \tag{30}$$

The support of the random variable used in masking μ is larger compared to when the field size is q_1 since $d_{\min} \leq |d_i(x_k) - d_j(x_k)|$. Thus, the estimation error of the user for the exact values of the samples can increase. In addition, note that this transform can be kept hidden from the user since it does not affect the result for ordering of the samples by construction.

4) Illustrative Example: This example demonstrates the masking approach and the field size expansion. Let the rejected samples, for example, be $\{[1,2]^T, [2,1]^T\}$, and the set of accepted samples be $\{[20,0]^T, [0,20]^T\}$. Let the field size be $q_1 = 809$. Thus, after simple calculations, we see that the range of the masking random variable is $\{0, \ldots, 39\}$. Let the user choose $x = [1,2]^T$. Thus, the queries sent to the two servers are given by (14) and after the servers reply, the user reprocesses the received answers according to (17), to obtain

$$\begin{bmatrix} \hat{A}_1^{[x]}(1) \\ \hat{A}_2^{[x]}(1) \end{bmatrix} = \boldsymbol{V}_2 \begin{bmatrix} 365 + \mu(1) \\ I(1) + Z'(1) \end{bmatrix}, \tag{31}$$

and

$$\begin{bmatrix} \hat{A}_1^{[x]}(2) \\ \hat{A}_2^{[x]}(2) \end{bmatrix} = \boldsymbol{V}_2 \begin{bmatrix} 325 + \mu(2) \\ I(2) + Z'(2) \end{bmatrix}. \tag{32}$$

It is clear that for any choice of $\mu(1)$, and $\mu(2)$, the user would be able to know that y_2 is closest to x.

Now, let us choose a larger field size, for example, $q_2 = 40009$ and apply the transform $T : a \to 10a$. Then, with simple calculations, we see that the range of the masking random variable is $\{0, \ldots, 399\}$ which is larger than the previous range and still maintains the relative distance. We encourage the reader to refer to Section VI-A for the effect of field size and $d_{\min}$ on the accuracy of the derived counterfactual.

D. Database Leakage

Now, we analyze the database leakage of the Diff-PCR and Mask-PCR schemes, and show that both schemes incur lower leakage than the baseline PCR scheme. The database leakage in (5) can be rewritten as follows

$$I(y_{[M]}; Q_{[N]}^{[x]}, A_{[N]}^{[x]} | x) = I(y_{[M]}; A_{[N]}^{[x]} | x, Q_{[N]}^{[x]}) + I(y_{[M]}; Q_{[N]}^{[x]} | x) \tag{33}$$

$$= I(y_{[M]}; A_{[N]}^{[x]} | x, Q_{[N]}^{[x]}) \tag{34}$$

$$= I(y_{[M]}; A_{[N]}^{[x]} | x). \tag{35}$$

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

where (34) follows since $y_{[M]} = (y_1, \ldots, y_M)$ are independent of the queries $Q_{[N]}^{[x]}$ given x. Also, we can remove $Q_{[N]}^{[x]}$ from the conditioning terms as in (35), simplifying the leakage expression. This is because, the randomness symbols in $Q_{[N]}^{[x]}$ added for the privacy of x, upon answer generation, are protected by the randomness shared among the servers. Additionally, they align on a subspace different from the database information subspace, leading to no leakage from the interference terms. The database leakage for the baseline scheme is therefore,

$$I(d_1(x), \ldots, d_M(x); y_{[M]}|x). \tag{36}$$

Let $r(i) = d_{i+1}(x) - d_i(x)$. Then, for the Diff-PCR scheme, it holds that

$$I(r(1), r(2), \ldots, r(M-1); y_{[M]}|x) \leq I(d_1(x), \ldots, d_M(x); y_{[M]}|x), \tag{37}$$

due to the data-processing inequality. This implies that the database leakage is lower than that of the baseline scheme.

To compare the leakages of the masking and the baseline schemes, we proceed as

$$\begin{aligned}
&I(y_{[M]}; d_i(x) + \mu(i), i \in [M]|x) \\
&\quad = H(d_1(x) + \mu(1), \ldots, d_M(x) + \mu(M)|x) \\
&\qquad - H(d_i(x) + \mu(i), i \in [M]|y_{[M]}, x) && (38)\\
&\quad \leq H(d_1(x), \mu(1), \ldots, d_M(x), \mu(M)|x) \\
&\qquad - H(\mu(i), i \in [M]|y_{[M]}, x) && (39)\\
&\quad = H(d_1(x), \ldots, d_M(x)|x) \\
&\qquad + H(\mu(1), \ldots, \mu(M)|x, d_1(x), \ldots, d_M(x)) \\
&\qquad - H(\mu(1), \ldots, \mu(M)|y_{[M]}, x) && (40)\\
&\quad = H(d_1(x), \ldots, d_M(x)|x) \\
&\qquad - H(d_1(x), \ldots, d_M(x)|x, y_{[M]}) && (41)\\
&\quad = I(y_{[M]}; d_1(x), \ldots, d_M(x)|x), && (42)
\end{aligned}$$

where (41) follows since the last two terms in (40) are equal, and $H(d_1(x), \ldots, d_M(x)|x, y_{[M]})$ is zero.

Remark 4: Observe that the two schemes are not analytically comparable. This is because of the parameter $d_{\min}$ in the Mask-PCR scheme, which affects the amount of leakage, since the masking random variable μ depends on it. Moreover, $d_{\min}$ depends on $\mathcal{D}$, which is arbitrarily distributed. To resolve this, we numerically evaluate the leakage for different values of $d_{\min}$ assuming that $d_{\min}$ is known globally and compare this with the Diff-PCR scheme (see Section VI-B). The numerical results show that Mask-PCR incurs a lower leakage than Diff-PCR, even when $d_{\min}$ is globally known. In reality, the leakage depends on the user's guess $\hat{d}_{\min}$ of $d_{\min}$, and the worst-case leakage occurs if $\hat{d}_{\min} = d_{\min}$. Moreover, since μ is uniform over $\hat{d}_{\min}$, the user guesses $\|y_{i+1} - x\|^2 - \|y_i - x\|^2$ for each $i \in [M-1]$ correctly with a probability less than 1, while this is known deterministically in the case of Diff-PCR. Thus, the worst case leakage in Mask-PCR is lower than that in Diff-PCR.

IV. I-PCR Schemes

In this section, we propose schemes that provide feasible solutions to minimizing (10) subject to (8) and (9). Further, let $d_i(x) = \|y_i - x\|^2, i \in \Theta$. First, we propose a scheme that operates in two phases, where the user starts by privately retrieving Θ, and uses this to prepare queries for the next phase. Next, we propose a single-phase scheme, which incurs a lower communication cost than the two-phase scheme, albeit requires a larger operating field. Finally, we show that the database leakage in the two-phase scheme is lower than that in the single-phase scheme. All I-PCR schemes require $N = 3$ replicated servers.

A. Two-Phase I-PCR

Theorem 5: There exists a two-phase scheme that can retrieve the exact closest counterfactual index, under the immutability constraint, from $N = 3$ servers with perfect user privacy, with the communication cost of $9(d + M)$ symbols of $\mathbb{F}_q$ where $q > R^2 d$ is prime.

The scheme proceeds in two phases. In the first phase, the user retrieves Θ, i.e., the indices i of y_i for which $x_{\mathcal{I}} = y_{i,\mathcal{I}}$. In the second phase, the user retrieves the squared ℓ_2 distance of the samples indexed by Θ and compares them to determine the counterfactual index $\theta^*_{\mathcal{I}} \in \Theta$. The scheme operates in a field $\mathbb{F}_q$ where $q > R^2 d$ is prime. Further, the set of queries and answers for the nth server over two phases are $Q_n^{[x]} = [Q_n^{[x_{\mathcal{I}}]}, Q_n^{[x_{\mathcal{M}}]}]$ and $A_n^{[x]} = [A_n^{[x_{\mathcal{I}}]}, A_n^{[x_{\mathcal{M}}]}]$, respectively.

1) Phase 1: Based on the user's choice of $\mathcal{I}$, the user constructs the binary vector $h_1 \in \{0,1\}^d$ such that its kth entry is

$$h_1(k) = \mathbb{1}[k \in \mathcal{I}], \quad k \in [d]. \tag{43}$$

Then, the user sends the following query tuple to server n

$$\begin{aligned}
Q_n^{[x_{\mathcal{I}}]} &= \left[Q_n^{[x_{\mathcal{I}}]}(1),\, Q_n^{[x_{\mathcal{I}}]}(2)\right] && (44)\\
&= \left[h_1 + \alpha_n Z_1,\, x \circ h_1 + \alpha_n Z_2\right], && (45)
\end{aligned}$$

where $\circ$ denotes element-wise product, and Z_1, Z_2 are chosen uniformly and independently from $\mathbb{F}_q^d$ by the user. Upon receiving the query, server n responds with the following answer corresponding to each $i \in [M]$

$$\begin{aligned}
A_n^{[x_{\mathcal{I}}]}(i) = \rho_i \|Q_n^{[x_{\mathcal{I}}]}(1) \circ y_i - Q_n^{[x_{\mathcal{I}}]}(2)\|^2 \\
+ \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i), && (46)
\end{aligned}$$

where ρ_i is chosen uniformly at random from $\mathbb{F}_q \setminus \{0\}$ by the servers, $Z_1'(i)$ and $Z_2'(i)$ are uniform random variables chosen independently from $\mathbb{F}_q$. Thus,

$$\begin{aligned}
A_n^{[x_{\mathcal{I}}]}(i) &= \rho_i \|(h_1 + \alpha_n Z_1) \circ y_i - (x \circ h_1 + \alpha_n Z_2)\|^2 \\
&\quad + \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) && (47)\\
&= \rho_i \|h_1 \circ (y_i - x)\|^2 \\
&\quad + \alpha_n \big(Z_1'(i) + 2\rho_i (h_1 \circ (y_i - x))^T (Z_1 \circ y_i - Z_2)\big) \\
&\quad + \alpha_n^2 \big(\rho_i \|Z_1 \circ y_i - Z_2\|^2 + Z_2'(i)\big). && (48)
\end{aligned}$$

From the answers of the three servers, the user obtains the answer vector $A^{[x_{\mathcal{I}}]}(i)$ as

$$\begin{bmatrix} A_1^{[x_{\mathcal{I}}]}(i) \\ A_2^{[x_{\mathcal{I}}]}(i) \\ A_3^{[x_{\mathcal{I}}]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} \rho_i \|h_1 \circ (y_i - x)\|^2 \\ I_{11}(i) \\ I_{12}(i) \end{bmatrix}, \tag{49}$$

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

where $I_{11}(i)$, $I_{12}(i)$ are interference terms that reveal no information on $\mathcal{D}$ to the user. For each i, the user concludes that, $i \in \Theta$ iff

$$\rho_i \|h_1 \circ (y_i - x)\|^2 = 0. \tag{50}$$

A non-zero value of $\rho_i\|h_1 \circ (y_i - x)\|^2$ implies that $y_{i,\mathcal{I}} \neq x_{\mathcal{I}}$ since ρ_i is a non-zero element of $\mathbb{F}_q$. Further, since ρ_is are independent random variables, the user does not learn which samples in $\mathcal{D}$ share some immutable attributes, unless they are identical to $x_{\mathcal{I}}$. If $\Theta = \emptyset$, then the user learns that there are no valid counterfactuals in $\mathcal{D}$ satisfying their current requirement, and the scheme ends here. Further, if $|\Theta| = 1$, the user does not need the second phase.

2) Phase 2: The user now compares the distances in a subset of $\mathcal{D}$, i.e., $\{y_i : i \in \Theta\}$ to find their counterfactual index θ^* without revealing Θ to the servers. Towards this, the user prepares the following binary vector $h_2 \in \{0,1\}^M$

$$h_2(i) = \begin{cases} 1, & \text{if } i \in \Theta, \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

Then, the user sends the following query tuple to server n

$$Q_n^{[x_{\mathcal{M}}]} = \left[Q_n^{[x_{\mathcal{M}}]}(1),\, Q_n^{[x_{\mathcal{M}}]}(2)\right] \tag{52}$$

$$= \left[h_2 + \alpha_n Z_3,\, x + \alpha_n Z_4\right], \tag{53}$$

where Z_3 and Z_4 are uniform vectors from $\mathbb{F}_q^M$ and $\mathbb{F}_q^d$, respectively. Each server constructs their masked database $\tilde{S}_n$ as follows

$$\tilde{S}_n = \operatorname{diag}(Q_n^{[x_{\mathcal{M}}]}(1)) \cdot \left[y_1,\, y_2,\, \ldots,\, y_M\right]^T \tag{54}$$

$$= \begin{bmatrix} (h_2(1) + \alpha_n Z_3(1)) \cdot y_1^T \\ (h_2(2) + \alpha_n Z_3(2)) \cdot y_2^T \\ \vdots \\ (h_2(M) + \alpha_n Z_3(M)) \cdot y_M^T \end{bmatrix}. \tag{55}$$

Let $\tilde{S}_n(i)$ denote the ith row of the masked database at server n. Then, their corresponding answer is given by

$$A_n^{[x_{\mathcal{M}}]}(i) = \|\tilde{S}_n(i)^T - Q_n^{[x_{\mathcal{M}}]}(2)\|^2 + \alpha_n Z_3'(i) + \alpha_n^2 Z_4'(i) \tag{56}$$

$$= \|(h_2(i) + \alpha_n Z_3(i))y_i - (x + \alpha_n Z_4)\|^2 + \alpha_n Z_3'(i) + \alpha_n^2 Z_4'(i) \tag{57}$$

$$= \|h_2(i)y_i - x\|^2 + \alpha_n\big(2(h_2(i)y_i - x)^T(Z_3(i)y_i - Z_4) + Z_3'(i)\big) + \alpha_n^2\big(\|Z_3(i)y_i - Z_4\|^2 + Z_4'(i)\big), \tag{58}$$

where $Z_3'(i)$ and $Z_4'(i)$ are uniform random variables from $\mathbb{F}_q$. From the answers of all three servers, the user computes the answer vector

$$\begin{bmatrix} A_1^{[x_{\mathcal{M}}]}(i) \\ A_2^{[x_{\mathcal{M}}]}(i) \\ A_3^{[x_{\mathcal{M}}]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} \|h_2(i)y_i - x\|^2 \\ I_{21}(i) \\ I_{22}(i) \end{bmatrix}, \tag{59}$$

where the first term reduces to

$$\|h_2(i)y_i - x\|^2 = \begin{cases} \|y_i - x\|^2 = d_i(x), & i \in \Theta, \\ \|x\|^2, & \text{otherwise.} \end{cases} \tag{60}$$

In other words, if the values of the immutable features do not match, the user does not learn any information (not even their distances) on the samples y_i, $i \in [M] \setminus \Theta$.

3) Communication Cost: The above scheme incurs an upload cost of $9d+3M$ and a download cost of $6M$. Thus, the worst case communication cost, i.e., when $|\Theta| > 1$ is $9d+9M$.

Remark 5: Consider an alternative two-phase scheme. Instead of sending $Q_n^{[x_{\mathcal{I}}]}$ and $Q_n^{[x_{\mathcal{M}}]}(2)$, the user can ask the servers to compute $x \circ h_1$ by securely sending x and h_1 in the query tuple, $Q_n^{[x_{\mathcal{I}}]} = \left[h_1 + \alpha_n Z_1,\, x + \alpha_n Z_2\right]$. Server n computes the product $(h_1 + \alpha_n Z_1) \circ (x + \alpha_n Z_2) = x \circ h_1 + \alpha_n \hat{Z}_1 + \alpha_n^2 \hat{Z}_2$ where $\hat{Z}_1$ and $\hat{Z}_2$ are the resulting interference terms. Moreover, the first term in the answer (47) is $\rho_i\|(h_1 + \alpha_n Z_1) \circ y_1 - (x \circ h_1 + \alpha_n \hat{Z}_1 + \alpha_n^2 \hat{Z}_2)\|^2$ (added to $\alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) + \alpha_n^3 Z_3'(i)$) leading to a polynomial of degree 4 in α_n. However, since $\alpha_n^4\|\hat{Z}_2\|^2, n \in [4]$ is known to the user, it can be canceled from the answers resulting in a reduced degree of 3. This requires $N = 4$ (instead of 3) replicated servers to retrieve Θ and the communication cost in the first phase is $8d+4M$. In the second phase, the user sends a single query $Q_n^{[x_{\mathcal{M}}]} = Q_n^{[x_{\mathcal{M}}]}(1) = h_2 + \alpha_n Z_3$ to any 3 out of the 4 servers who then compute their answers as described in the aforementioned scheme. This results in an upload cost of $3M$ and a download cost of $3M$. The total communication cost is, therefore, $8d + 10M$. Therefore, although the scheme requires an additional server, its communication cost is lower than the previous scheme, if $d > M$.

4) Computation Complexity: At the user side, the computation of query vectors in the first phase is $O(d)$ and that in the second phase is $O(M)$. The decoding in both phases requires $O(M)$ operations. For each server, the answer generation requires $O(Md)$ computations in both phases.

B. Single-Phase I-PCR

Theorem 6: There exists a single-phase scheme that can retrieve the exact closest counterfactual index from $N = 3$ servers with perfect user privacy, with the communication cost of $6d + 3M$ symbols of $\mathbb{F}_q$ where $q > F(L-1)R^2 + R^2 d$ is prime. Here, F is an upper bound on $|\mathcal{I}|$ and L is a globally known parameter.

For this scheme, we assume that the user assigns at most F features as immutable, where F is globally known. Thus, $\mathcal{I} \subset [d]$ where $|\mathcal{I}| \leq F$. To this end, the servers and the user agree on $\mathbb{F}_q$ with $q > F(L-1)R^2 + R^2 d$ and prime. The knowledge of F is necessary only to reduce the field size required for the scheme design. The user chooses an integer $L > R^2 d$ as a scaling factor for the immutable attributes, where the lower bound on L is crucial for the resolvability of the sets Θ and $[M]\setminus\Theta$ at the time of decoding. To save on the communication cost, we let $L = R^2 d + 1$. Next, the user designs a vector $h \in \{1, L\}^d$

$$h(k) = \begin{cases} L, & \text{if } k \in \mathcal{I}, \\ 1, & \text{otherwise.} \end{cases} \tag{61}$$

To server n, the user sends the query tuple

$$Q_n^{[x]} = \left[Q_n^{[x]}(1),\, Q_n^{[x]}(2),\right] \tag{62}$$

$$= \left[x + \alpha_n Z_1,\, h + \alpha_n Z_2\right], \tag{63}$$

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

where Z_1 and Z_2 are independent uniform random vectors in $\mathbb{F}_q^d$ that are private to the user. In response, server n computes the following answer corresponding to y_i

$$A_n^{[x]}(i) = (y_i - Q_n^{[x]}(1))^T \big((y_i - Q_n^{[x]}(1)) \circ Q_n^{[x]}(2)\big) + \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) \tag{64}$$

$$= (y_i - x)^T \big((y_i - x) \circ h\big) + \alpha_n I_1(i) + \alpha_n^2 I_2(i) + \alpha_n^3 Z_1^T (Z_1 \circ Z_2), \tag{65}$$

where

$$I_1(i) = (y_i - x)^T \big((y_i - x) \circ Z_2\big) - 2(y_i - x)^T (Z_1 \circ h\big) + Z_1'(i) \tag{66}$$

$$I_2(i) = Z_1^T \big(Z_1 \circ h\big) - 2Z_1^T \big((y_i - x) \circ Z_2\big) + Z_2'(i). \tag{67}$$

Upon receiving the answers from the 3 servers, the user cancels the term of the third degree, since it is known to them, i.e.,

$$\hat{A}_n^{[x]}(i) = A_n^{[x]}(i) - \alpha_n^3 Z_1^T (Z_1 \circ Z_2), \tag{68}$$

which gives

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \\ \hat{A}_3^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} (y_i - x)^T \big((y_i - x) \circ h\big) \\ I_1(i) \\ I_2(i) \end{bmatrix}. \tag{69}$$

Now, to determine Θ, the user checks the range of $(y_i - x)^T \big((y_i - x) \circ h\big)$. If it lies in $[0 : (d - |\mathcal{I}|)R^2]$, all immutable features match in value and $i \in \Theta$, satisfying

$$(y_i - x)^T \big((y_i - x) \circ h\big) = d_i(x). \tag{70}$$

Among the indices in Θ, the user selects the index with the minimum value of $d_i(x)$ as $\theta^*_{\mathcal{I}}$. On the other hand, if at least one immutable feature does not match, the resulting range of $(y_i - x)^T \big((y_i - x) \circ h\big)$ is $[L : L|\mathcal{I}|R^2 + (d - |\mathcal{I}|)R^2]$. Since L is chosen such that $L > R^2 d$, these two sets are disjoint and the user is able to distinguish between these samples.

1) Communication Cost: The upload cost is $2d$ symbols to each server, while the download cost is M symbols from each server. This results in a communication cost of $6d + 3M$.

2) Computation Complexity: The number of computations is $O(M + d)$ at the user side, and $O(Md)$ at each server side.

C. Database Leakage

Now, we compare the database leakage incurred by both the schemes. The database leakage in (10) can be rewritten as follows

$$I(y_{[M]}; Q_{[N]}^{[x]}, A_{[N]}^{[x]} | x, \mathcal{I}) = I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}, Q_{[N]}^{[x]}) + I(y_{[M]}; Q_{[N]}^{[x]} | x, \mathcal{I}) \tag{71}$$

$$= I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}, Q_{[N]}^{[x]}) \tag{72}$$

$$= I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}), \tag{73}$$

where (72) follows since $y_{[M]} = (y_1, \ldots, y_M)$ are independent of the queries $Q_{[N]}^{[x]}$ given x and $\mathcal{I}$, and (73) holds for both schemes because the randomness symbols in $Q_{[N]}^{[x]}$ added for the privacy of $x, \mathcal{I}$, upon answer generation are protected by the randomness shared among the servers. Additionally, they align on a subspace different from the database information subspace, leading to no leakage from the interference terms. In particular, the database leakage for the two-phase scheme originates from $\rho_i \|h_1 \circ (y_i - x)\|^2$ in (49) and $\|h_2 \circ y_i - x\|^2$ in (60). Similarly, $(y_i - x)^T \big((y_i - x) \circ h\big)$ in (69) is responsible for leakage in the single-phase scheme.

For comparing the database leakage between the two schemes, define for each $i \in [M]$,

$$E_i = \mathbb{1}\left[x_{\mathcal{I}} = y_{i,\mathcal{I}}\right]. \tag{74}$$

Thus, the answers for both schemes satisfy

$$H(E_{[M]} | A_{[N]}^{[x]}, x, \mathcal{I}) = 0, \tag{75}$$

where $E_{[M]} = (E_1, \ldots, E_M)$. The database leakage is

$$I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}) = I(E_{[M]}, y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}), \tag{76}$$

since $I(E_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}, y_{[M]}) = 0$. The right hand side of (76) can be expanded as

$$I(E_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}) + I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}, E_{[M]}) = H(E_{[M]} | x, \mathcal{I}) + I(y_{[M]}; A_{[N]}^{[x]} | x, \mathcal{I}, E_{[M]}), \tag{77}$$

where (77) follows from (75). The first term in the leakage appears for both schemes since Θ is inevitably learned by the user while computing θ^*. The second term, however, is different for the two schemes.

For the two-phase scheme, the second term in (77) is the information that the user infers from the squared norm values $\|y_i - x\|^2 = \|y_{i,\mathcal{M}} - x_{\mathcal{M}}\|^2$ only for $E_i = 1$, i.e., $i \in \Theta$. On the other hand, for the single-phase scheme, the second term contains additional information about the immutable features of $y_i, i \notin \Theta$. In particular, if one immutable feature does not match, then $(y_i - x)^T \big((y_i - x) \circ h\big) \in \mathcal{J}_1 = [L : LR^2 + R^2(d - |\mathcal{I}|)]$ whereas if two immutable features do not match, then $(y_i - x)^T \big((y_i - x) \circ h\big) \in \mathcal{J}_2 = [2L : 2LR^2 + R^2(d - |\mathcal{I}|)]$ and so on. In general $\mathcal{J}_k = [kL : kL + R^2(d - |\mathcal{I}|)]$ is the set where $(y_i - x)^T \big((y_i - x) \circ h\big)$ lies if k out of $|\mathcal{I}|$ features do not agree. These sets, though overlapping, leak non-zero amount of information on $y_i, i \notin \Theta$. Another way to see that the leakage in the single-phase scheme is higher, is because, the user observes the values of $(y_i - x)^T \big((y_i - x) \circ h\big)$ for $i \in [M] \setminus \Theta$, in addition to learning $E_{[M]}$. Whereas, for the two-phase scheme, no additional information about y_i is learnt by the user if $i \in [M] \setminus \Theta$.

Remark 6: From PCR to I-PCR, the number of replicated servers that participate for our schemes increases from $N = 2$ to $N = 3$. This is due to the added constraint of hiding the immutable set $\mathcal{I}$ from the institution, which necessitates an additional dimension to be resolved, requiring an extra server.

V. Incorporating User-Actionability

In this section, we implement user-actionability, i.e., the user's preference to modify some features over others (among the ones they are willing to change), to obtain a more actionable counterfactual, while keeping their preference private from the institution. To model this, we assume that the user assigns different *weights* to their attributes based on their preference to change those attributes to attain a counterfactual. The more reluctant the user is to change a given attribute, the

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

higher is the weight of that attribute. Accordingly, the distance metric is now the weighted ℓ_2 norm whose coordinates are scaled by the weight vector. In order to ensure that the user's preference is also private, the weight vector should be kept private from the servers.

A. PCR+ Schemes

Let $w \in [L_1]^d$ denote the user's weight vector, where L_1 is a positive integer. The modified distance metric is given by

$$d(x, y) = (y - x)^T ((y - x) \circ w). \tag{78}$$

The goal is same as (1) with $d(\cdot, \cdot)$ modified according to (78). The decodability (3), user-privacy (4) and database leakage (5) definitions are modified to include w with the random variable x in the arguments of $H(\cdot)$ and $I(\cdot; \cdot)$. Then, we have the following theorem.

Theorem 7: There exist schemes derived from Baseline PCR, Diff-PCR and Mask-PCR, that can retrieve the exact closest counterfactual with user's actionability weights using $N = 3$ servers while keeping the user's sample and actionability weights hidden from each server. The communication costs of the respective schemes are:

1) Baseline PCR+: $6d + 3M$ symbols of $\mathbb{F}_q$, $q > R^2 L_1 d$;
2) Diff PCR+: $6d + 3(M - 1)$ symbols of $\mathbb{F}_q$, $q > 2R^2 L_1 d$;
3) Mask PCR+: $6d + 3M$ symbols of $\mathbb{F}_q$, $q > R^2 L_1 d$;

where q is prime.

To prove this, we describe the extended schemes next. Since it is clear from the context, we use the same notations as the PCR setting to denote the queries and answers.

1) Baseline PCR+: For this scheme, the required minimum field size is $q > R^2 L_1 d$. Now, the user sends the query tuple $Q_n^{[x]} = [Q_n^{[x]}(1), Q_n^{[x]}(2)]$ to server n, where

$$Q_n^{[x]}(1) = x + \alpha_n Z_1, \quad Q_n^{[x]}(2) = w + \alpha_n Z_2, \tag{79}$$

and Z_1 and Z_2 are uniform independent random vectors in $\mathbb{F}_q^d$. Upon receiving the queries, the answers are generated as follows,

$$\begin{aligned} A_n^{[x]}(i) &= (y_i - Q_n^{[x]}(1))^T \operatorname{diag}(Q_n^{[x]}(2))(y_i - Q_n^{[x]}(1)) \\ &\quad + \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) \qquad (80) \\ &= d(y_i, x) + \alpha_n \Big((y_i - x)^T ((y_i - x) \circ Z_2) \\ &\quad - 2(y_i - x)^T (Z_1 \circ w) + Z_1'(i) \Big) \\ &\quad + \alpha_n^2 \Big(Z_1^T (Z_1 \circ w) - 2(y_i - x)^T (Z_1 \circ Z_2) \\ &\quad + Z_2'(i) \Big) + \alpha_n^3 Z_1^T (Z_1 \circ Z_2), \qquad (81) \end{aligned}$$

where recall that, $\circ$ denotes element-wise multiplication, and $Z_2'(i)$ are uniform random variables shared by the servers and used to invoke the one-time pad theorem.

Upon receiving the answers from $N = 3$ servers, the user decodes them $i \in [M], n \in [3]$ as follows

$$\hat{A}_n^{[x]}(i) = A_n^{[x]}(i) - \alpha_n^3 Z_1^T (Z_1 \circ Z_2), \tag{82}$$

since Z_2 is available to the user. Next,

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \\ \hat{A}_3^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} d(y_i, x) \\ I_1(i) \\ I_2(i) \end{bmatrix}, \tag{83}$$

where

$$I_1(i) = (y_i - x)^T ((y_i - x) \circ Z_2) - 2(y_i - x)^T (Z_1 \circ w) + Z_1'(i) \tag{84}$$

$$I_2(i) = Z_1^T (Z_1 \circ w) - 2(y_i - x)^T (Z_1 \circ Z_2) + Z_2'(i). \tag{85}$$

The upload cost in this scheme is $6d$, while the download cost is $3M$.

2) Diff-PCR+: The queries are given by (79). Similar to Diff-PCR, the servers construct their answers so that the user decodes only the difference of distances, instead of the exact distances. This time, the field of operation is a prime $q > 2R^2 L_1 d$, since, for all $i, j \in [M]$, the absolute difference $|d(y_i, x) - d(y_j, x)| = |(y_i - x)^T ((y_i - x) \circ w) - (y_j - x)^T ((y_j - x) \circ w)|$ is at most $R^2 L_1 d$. Therefore, in $\mathbb{F}_q$

$$d(y_i, x) - d(y_j, x) \in \begin{cases} \mathcal{H}, & y_j \text{ is closer to } x, \\ \mathcal{H}^c, & y_i \text{ is closer to } x, \end{cases} \tag{86}$$

where $\mathcal{H} = [0 : R^2 L_1 d]$ and $\mathcal{H}^c = [R^2 L_1 d + 1 : q - 1]$. The user sends the same queries as in (79). Let the servers share two common randomness vectors $Z_1' = [Z_1'(1), \ldots, Z_1'(M-1)]^T$ and $Z_2' = [Z_2'(1), \ldots, Z_2'(M-1)]^T$ each of length $M - 1$ where each entry is a uniform random variable from $\mathbb{F}_q$. Server $n \in [3]$ constructs the following answer for each $i \in [M - 1]$

$$\begin{aligned} A_n^{[x]}(i) &= (y_i - Q_n^{[x]}(1))^T \operatorname{diag}(Q_n^{[x]}(2))(y_i - Q_n^{[x]}(1)) \\ &\quad - (y_{i+1} - Q_n^{[x]}(1))^T \operatorname{diag}(Q_n^{[x]}(2))(y_{i+1} - Q_n^{[x]}(1)) \\ &\quad + \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) \qquad (87) \\ &= d(y_i, x) - d(y_{i+1}, x) + \alpha_n I_1'(i) + \alpha_n^2 I_2'(i) \qquad (88) \end{aligned}$$

where $I_1'(i) = (y_i - x)^T ((y_i - x) \circ Z_2) - (y_{i+1} - x)^T ((y_{i+1} - x) \circ Z_2) + 2Z_1^T ((y_i - y_{i+1}) \circ w) + Z_1'(i)$ and $I_2'(i) = 2(y_{i+1} - y_i)^T (Z_1 \circ Z_2) + Z_2'(i)$ are the interference terms. The answers $A_n^{[x]}(i)$ of the three servers can be written as

$$\begin{bmatrix} A_1^{[x]}(i) \\ A_2^{[x]}(i) \\ A_3^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} d(y_i, x) - d(y_{i+1}, x) \\ I_1'(i) \\ I_2'(i) \end{bmatrix}. \tag{89}$$

Therefore, the user recovers the $M - 1$ differences, while discarding the interference terms. Now, with the $M - 1$ differences, the user evaluates their counterfactual index θ^* following sequential comparisons similar to Algorithm 1, the only difference being the range in line 5 is replaced by $\mathcal{H}^c$. The upload cost in this scheme is $6d$, while the download cost is $3(M - 1)$.

3) Mask-PCR+: The operating field size is the same as that of Baseline PCR+, i.e., $q > R^2 L_1 d$. Upon receiving the queries, as given in (79), each server applies the following on each sample y_i and sends to the user

$$\begin{aligned} A_n^{[x]}(i) &= \left(y_i - Q_n^{[x]}(1)\right)^T \operatorname{diag}\left(Q_n^{[x]}(2)\right)\left(y_i - Q_n^{[x]}(1)\right) \\ &\quad + \mu(i) + \alpha_n Z_1'(i) + \alpha_n^2 Z_2'(i) \qquad (90) \\ &= d(y_i, x) + \mu(i) + \alpha_n I_1(i) + \alpha_n^2 I_2(i) \\ &\quad + \alpha_n^3 Z_1^T (Z_1 \circ Z_2), \qquad (91) \end{aligned}$$

where $\mu(i)$ is as defined in Section III-C, $Z_1'(i)$, and $Z_2'(i)$ are uniform random variables and $I_1(i)$ and $I_2(i)$ are as given in (84). Using the answers of the $N = 3$ servers and applying the

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

decoding approach given in Section V-A.1, the user obtains the masked weighted distance corresponding to y_i as follows

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \\ \hat{A}_3^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_3 \begin{bmatrix} d(y_i, x) + \mu(i) \\ I_1(i) \\ I_2(i) \end{bmatrix}. \tag{92}$$

To show that there is no need to change the range of the masking random variable μ, under the modified distance metric, we provide the following lemma.

Lemma 2: The range of the random variable designed to mask the exact distance in this case is the same as in the non-weighted case.

The proof proceeds similarly as that of Lemma 1 and is relegated to Appendix C. The upload cost in this scheme is $6d$, and the download cost is $3M$.

B. I-PCR+ Schemes

Under the immutability constraint, the weight vector is restricted to the mutable features, i.e., those in $\mathcal{M}$. To realize this, the user assigns an integer $w'(k) \in [L_1]$ for each $k \in \mathcal{M}$, where a higher value indicates more reluctance to change that attribute. Then, the user's goal is to find the index i that minimizes

$$(y_{i,\mathcal{M}} - x_{\mathcal{M}})^T \left((y_{i,\mathcal{M}} - x_{\mathcal{M}}) \circ [w'(k)]_{k \in \mathcal{M}} \right) \tag{93}$$

over all $y_i \in \mathcal{D}$ such that $y_{i,\mathcal{I}} = x_{\mathcal{I}}$, i.e., $i \in \Theta$. The decodability (8), user privacy (9) and (10) are modified to include w as an additional argument in $H(\cdot)$ and $I(\cdot;\cdot)$, along with x and $\mathcal{I}$. Then, the I-PCR schemes can be modified to achieve user-actionability as stated in the following theorem.

Theorem 8: There exist schemes derived from Two-Phase I-PCR and Single-Phase I-PCR that can retrieve the exact closest counterfactual under the immutability constraint with user's actionability, while keeping the user's sample and actionability weights hidden from each server. The communication costs of the respective schemes are:

1) Two-Phase I-PCR+: $14d + 11M$ symbols of $\mathbb{F}_q$, $q > R^2 L_1 d$
2) Single-Phase I-PCR+: $6d+3M$ symbols of $\mathbb{F}_q$, $q > F(L-1)R^2 + R^2 L_1 d$

where q is prime.

1) Two-Phase I-PCR+: For the two-phase I-PCR scheme, the field size requirement is now $\mathbb{F}_q$ with $q > L_1 R^2 d$. The first phase proceeds in the same manner as that of the Two-Phase I-PCR scheme (Section IV-A), and the user obtains Θ. In the second phase, the user creates a vector w whose kth entry is given by

$$w(k) = \begin{cases} 1, & k \in \mathcal{I}, \\ w'(k), & k \in \mathcal{M}. \end{cases} \tag{94}$$

Then, the user sends the following query to server n

$$Q_n^{[x,\mathcal{M}]} = \left[Q_n^{[x,\mathcal{M}]}(1),\, Q_n^{[x,\mathcal{M}]}(2),\, w + \alpha_n Z_5 \right] \tag{95}$$

$$= \left[h_2 + \alpha_n Z_3,\, x + \alpha_n Z_4,\, w + \alpha_n Z_5 \right], \tag{96}$$

where h_2 is defined in (51), Z_3, Z_4 and Z_5 are uniform vectors, chosen independently from $\mathbb{F}_q^M$, $\mathbb{F}_q^d$ and $\mathbb{F}_q^d$, respectively. Correspondingly, the answer returned by server n for each $i \in [M]$ is

$$\begin{aligned} A_n^{[x,\mathcal{M}]}(i) &= \left(\tilde{S}_n(i)^T - Q_n^{[x,\mathcal{M}]}(2) \right)^T \\ &\quad \cdot \left(\tilde{S}_n(i)^T - Q_n^{[x,\mathcal{M}]}(2) \circ (w + \alpha_n Z_5) \right) + \alpha_n Z_3'(i) \\ &\quad + \alpha_n^2 Z_4'(i) + \alpha_n^3 Z_5'(i) \qquad (97) \\ &= (h_2(i) y_i - x)^T \left((h_2(i) y_i - x) \circ w \right) + \alpha_n I_{31}(i) \\ &\quad + \alpha_n^2 I_{32}(i) + \alpha_n^3 I_{33}(i), \qquad (98) \end{aligned}$$

where $\tilde{S}_n(i)$ is the ith row of the masked database defined in (55), $Z_3'(i)$, $Z_4'(i)$ and $Z_5'(i)$ are uniform and independent random variables from $\mathbb{F}_q$. Since (98) is a polynomial of degree 3 in α_n, the user requires answers from $N = 4$ servers to perform decoding as follows

$$\begin{bmatrix} A_1^{[x,\mathcal{M}]}(i) \\ A_2^{[x,\mathcal{M}]}(i) \\ A_3^{[x,\mathcal{M}]}(i) \\ A_4^{[x,\mathcal{M}]}(i) \end{bmatrix} = \boldsymbol{V}_4 \begin{bmatrix} (h_2(i) y_i - x)^T \left((h_2(i) y_i - x) \circ w \right) \\ I_{31}(i) \\ I_{32}(i) \\ I_{33}(i) \end{bmatrix}, \tag{99}$$

where the first term reduces to

$$\begin{aligned} &(h_2(i) y_i - x)^T \left((h_2(i) y_i - x) \circ w \right) \\ &\quad = \begin{cases} (y_i - x)^T \left((y_i - x) \circ w \right), & i \in \Theta, \\ x^T (x \circ w), & \text{otherwise.} \end{cases} \end{aligned} \tag{100}$$

The upload and download costs in the first phase are $6d$ and $3M$, respectively, while those in the second phase are $8d+4M$ and $4M$, respectively. This results in the total communication cost of $14d + 11M$ symbols.

2) Single-Phase I-PCR+: With the single-phase scheme, actionability can be directly implemented by modifying (61) as follows

$$h(k) = \begin{cases} L, & k \in \mathcal{I}, \\ w'(k), & k \in \mathcal{M}, \end{cases} \tag{101}$$

where $L > L_1 R^2 d$. The rest of the scheme follows the description in Section IV-B with $N = 3$ servers. This leads to the same communication cost of $6d + 3M$ symbols, as that of the single-phase scheme, however, in a larger field $\mathbb{F}_q$, with $q > F(L-1)R^2 + L_1 R^2 d$.

Remark 7: Extensions to PCR+ from PCR, and to two-phase I-PCR+ from I-PCR, require an additional dimension to preserve the privacy of the user's weight vector. This is in addition to the requirement of a larger field size. Only in the single-phase I-PCR+ scheme, the number of servers, and the communication cost are not affected by actionability. However, this comes at the expense of a much higher field size.

VI. Experiments

In this section, we describe the experiments performed on both synthetic and real-world datasets. For each experiment, we describe the dataset being used, and how they are processed before implementation. The first experiment demonstrates the dependence of the accuracy of the Baseline PCR scheme on the operating field size, which is decided by the number of quantization levels R. Additionally, for the Mask-PCR

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

scheme, we empirically determine $d_{\min}$ to lift the restriction that $x \in \mathcal{D}_c$. In our next two experiments, we compute the database leakage of our schemes using two measures: i) mutual information (MI), as in our theoretical analysis, and ii) maximal (Max) leakage [56]. The reason for including Max leakage is due to the operational meaning it holds in terms of the worst-case leakage guarantee to the system designer (here, the institution/servers) on adopting a scheme. We revisit the definition of Max leakage in our setting, starting with the conditional version. Let X denote the sensitive random variable the adversary wants to guess, Y denote the observations of the scheme, and Z denote the auxiliary random variable already known to the adversary. Then, given a realization z of Z, the conditional maximal leakage denoted by $\mathcal{L}(X \to Y|Z = z)$ measures the *maximum* multiplicative increase in correctly guessing X, upon observing Y, given $Z = z$. When maximized over all realizations of z, denoted by the set $\mathcal{Z}$, we obtain the maximal leakage $\mathcal{L}(X \to Y)$ as

$$\mathcal{L}(X \to Y) = \max_{z \in \mathcal{Z}} \mathcal{L}(X \to Y|Z = z). \tag{102}$$

We know from [56] that $\mathcal{L}(X \to Y)$ in terms of X, Y and Z is:

$$\mathcal{L}(X \to Y) = \max_{z \in \mathcal{Z}} \log_q \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}: \mathbb{P}(X=x)>0} \mathbb{P}(Y = y|X = x, Z = z). \tag{103}$$

In our setting, $X = (y_1, \ldots, y_M)$, Y is the received answers $A_{[N]}^{[x]}$, and Z corresponds to the rejected instance x, in PCR and includes $\mathcal{I}$ in I-PCR schemes. We calculate the leakage values for the Baseline, Diff- and Mask-PCR schemes on a synthetic and real dataset, for various values of $d_{\min}$. Finally, we numerically evaluate the database leakage for the Two-Phase and Single-Phase I-PCR schemes, for various sizes of $\mathcal{I}$.

A. Accuracy-Quantization Trade-Off in PCR

As detailed in Section III-A, the proposed schemes operate on finite fields with size q. However, a real-world application might require some of the features to be real-valued. To circumvent this, we quantize each feature of the counterfactual instances as well as the user sample before applying the schemes to align with our system model. However, the quantization of each feature to integers in $[0 : R]$ will lead to a loss of some information that might result in an error in finding the closest counterfactual. The error can be made arbitrarily small by increasing the number of quantization levels. The greater the value of R, the larger is the field size q for absolute accuracy.

Moreover, in the case of Mask-PCR, we lift the restriction $x \in \mathcal{D}_c$ off by empirically deciding the parameter $d_{\min}$ to be used such that the scheme works for most of the points in a sample set of x. As pointed out in Section III-C, a larger field size q is needed to accommodate a larger value for $d_{\min}$.

1) Wine Quality Dataset: The Wine Quality Dataset [57] contains 4898 instances, each with 11 real-valued features. The original target variable "Quality" is categorical, taking integer values from 0 to 10. We convert the target into a binary variable by defining $t = \mathbb{1}$["Quality" ≥ 5]. This gives us 3788 instances with $t = 1$ and 183 instances with $t = 0$ (which are the potential rejected instances).

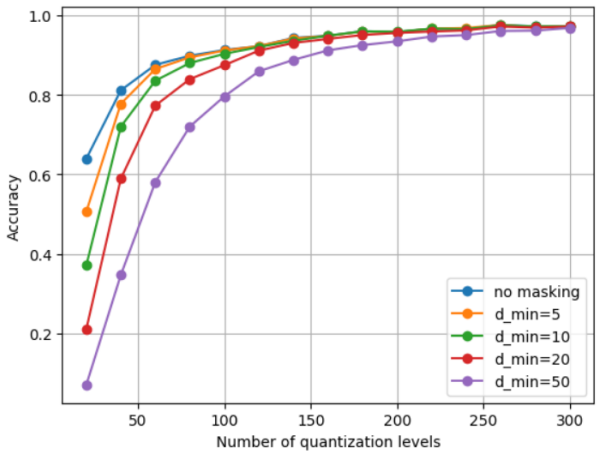


Fig. 3. Accuracy-quantization trade-off.

We consider all the instances with $t = 1$ to be the accepted (potential counterfactual) instances. Let this set be $\mathcal{S}_1$. The set $\mathcal{S}_0$ consists of all the rejected instances, i.e., with $t = 0$. In each round, we pick a subset $\mathcal{D}$, with $|\mathcal{D}| = M$, of $\mathcal{S}_1$ as the database and a subset $\mathcal{S}_x$ of $\mathcal{S}_0$ as the rejected instances. Then, we apply the retrieval schemes with varying parameters to obtain the nearest neighbor counterfactuals. The accuracy of the retrieved counterfactuals is computed as follows: Denote the counterfactual of x retrieved using the scheme by $\hat{y}(x)$ and the actual (i.e., the exact nearest neighbor) counterfactual by $y(x)$. Note that the scheme provides the index θ^* of the counterfactual, hence, we can retrieve the exact unquantized counterfactual irrespective of the quantization applied. The accuracy of the scheme is

$$\text{Accuracy} = \frac{1}{|\mathcal{S}_x|} \sum_{x_i \in \mathcal{S}_x} \mathbb{1}\Big[\|x_i - \hat{y}(x_i)\| \leq \|x_i - y(x_i)\|\Big]. \tag{104}$$

We average the accuracy over several rounds with uniformly sampled $\mathcal{D} \subset \mathcal{S}_1$ and $\mathcal{S}_x \subset \mathcal{S}_0$. We set the database size M to be 500 and the number of queries per round (i.e., $|\mathcal{S}_x|$) to be 50. We repeat the experiment for 100 rounds.

Fig. 3 shows how the accuracy improves with an increasing number of quantization levels. Further, for a given number of quantization levels, the accuracy degrades as $d_{\min}$ increases. Note that, when $d_{\min} = 1$, the masking scheme reduces to the Baseline PCR scheme, represented by the top blue curve.

B. Database Leakage in PCR

To observe the success of Diff-PCR and Mask-PCR in mitigating the leakage, and evaluate their leakage values, we proceed by following the steps in Appendix D. For computing the MI leakage, we make the following assumptions on the probability mass functions of x, and of $y_1, \ldots, y_M$ given x.

1) Synthetic Dataset: The samples are assumed to be uniformly distributed over $[0 : R]^d$ space. For each rejected instance x, the counterfactuals $y_1, \ldots, y_M$ are assumed to be uniformly distributed over $[0 : R]^d \backslash \{x\}$. We consider $R = 3$ and $d = 3$ with $M = 3$. We assume the queries are equi-probable

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

TABLE II
LEAKAGE RESULTS FOR PCR SCHEMES TO BASE $q = 757$ – SYNTHETIC DATASET

Scheme	MI Leakage	Max Leakage
Baseline PCR	1.143	1.304
Diff-PCR	0.940	1.126
Mask-PCR ($d_{\min} = 2$)	0.911	1.102
Mask-PCR ($d_{\min} = 3$)	0.777	0.989

TABLE III
LEAKAGE RESULTS FOR PCR SCHEMES TO BASE $q = 2003$ – COMPAS DATASET

Scheme	MI Leakage	Max Leakage
Baseline PCR	1.51437	1.5143801
Diff-PCR	1.51436	1.5143801
Mask-PCR ($d_{\min} = 2$)	1.51427	1.5143791

over the $[0 : R]^d$ space. Table II lists the computed leakage values under these assumptions.

2) COMPAS Dataset: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset [58] includes 6172 instances and 20 features (after one-hot encoding the categorical features) about individuals, which are used to predict the likelihood of reoffending of a convicted person. The target variable t is "is_recid". Class-wise counts are 3182 and 2990 for $t = 0$ and $t = 1$, respectively. We normalize all the features to the interval $[0, 1]$ and quantize to $R + 1$ levels with $R = 10$ during pre-processing, leaving 560 unique instances with $t = 0$ and 717 unique instances with $t = 1$. As our schemes are independent of the classifier, we use these labels as the classifications of the instances.

In order to compute the leakage values, we assume the following: 1) Queries x are equi-probable over the $t = 0$ (rejected) instances. Therefore, $\mathbb{P}_X(x) = \frac{1}{560}$ (denoted by ρ_X). 2) Since computing the histogram $\mathcal{C}(d(x), x)$ in (139) over all possible M-tuples (with $M = 5$ in this particular experiment) is too computationally intensive, we sample 100000 of the possible permutations and evaluate the leakage over those. Accordingly, given the query x, the M-tuples $y = (y_1, \ldots, y_M)$ are assumed to be equi-probable over a set of 100000 M-tuples that can be generated using the $t = 1$ (accepted) instances. Therefore, $\mathbb{P}_{Y|X}(y|x) = \frac{1}{100000}$ (denoted by $\rho_{Y|X}$). Table III presents the results. Similar to the values obtained with the synthetic dataset, we observe that Diff-PCR and Mask-PCR schemes reduce leakage compared to the baseline scheme.

C. Database Leakage in I-PCR

We compute the MI leakage value for the single-phase scheme as

$$I(y_{[M]}; (y_i - x)^T\big((y_i - x) \circ h\big), i \in [M]|x, \mathcal{I}), \tag{105}$$

where $|\mathcal{I}|$ is fixed. To evaluate Max leakage, we compute

$$\max_{x\in[0:R]^d} \max_{\mathcal{I}:|\mathcal{I}| \text{ fixed}} \mathcal{L}\big(y_{[M]} \to (y_i - x)^T\big((y_i - x) \circ h\big), i \in [M]|x, \mathcal{I}\big), \tag{106}$$

where

$$\begin{aligned}&\mathcal{L}\left(y_{[M]} \to (y_i - x)^T\big((y_i - x) \circ h\big), i \in [M]|x, \mathcal{I}\right)\\ &= \log_q\big[\text{number of unique combinations of}\\ &\qquad (y_i - x)^T\big((y_i - x) \circ h\big), i \in [M] \text{ given } x, \mathcal{I}\big].\end{aligned} \tag{107}$$

Note that, in (106) x and $\mathcal{I}$ denote the realizations of user's sample and immutable set, respectively, across which the leakage quantity is maximized. On the other hand, in (105), x and $\mathcal{I}$ denote the random variables over which the leakage computation is averaged. For the two-phase scheme, we simplify the database leakage expression using (76) and (77), to obtain,

$$I(y_{[M]}; A_{[N]}^{[x]}|x, \mathcal{I}) = H(E_{[M]}|x, \mathcal{I}) + I(y_{[M]}; A_{[N]}^{[x]}|x, \mathcal{I}, E_{[M]}). \tag{108}$$

Expanding the latter term gives,

$$\begin{aligned}&I(y_{[M]}; A_{[N]}^{[x_\mathcal{I}]}, A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]})\\ &= I(y_{[M]}; A_{[N]}^{[x_\mathcal{I}]}|x, \mathcal{I}, E_{[M]}) + I(y_{[M]}; A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]}, A_{[N]}^{[x_\mathcal{I}]}),\end{aligned} \tag{109}$$

where the first term on the right hand side is 0 since the answers after the first phase reveal no more information on $y_{[M]}$ beyond $E_{[M]}$. Now, since $H(A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]}, A_{[N]}^{[x_\mathcal{I}]}, y_{[M]}) = 0$, we re-write the second term in (109) as

$$\begin{aligned}&I(y_{[M]}; A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]}, A_{[N]}^{[x_\mathcal{I}]})\\ &\qquad = H(A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]}, A_{[N]}^{[x_\mathcal{I}]}) &(110)\\ &\qquad = H(A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, E_{[M]}). &(111)\end{aligned}$$

Here, (111) follows because, for each $i \in [M]$, after decoding, $A_{[N]}^{[x_\mathcal{I}]}(i)$ can be represented as

$$\rho_i \|y_{i,\mathcal{I}} - x_\mathcal{I}\|^2 = \nu_i E_i, \tag{112}$$

for some ν_i. Further, $I(A_{[N]}^{[x_\mathcal{M}]}; \nu_i E_i, i \in [M]|x, \mathcal{I}, E_{[M]}) = 0$ since ρ_i is chosen at random from $[1 : q - 1]$ independently, and thus ν_i is also independent from $y_{[M]}$, x, and $Q_n^{[x_\mathcal{M}]}$, making it independent of $A_{[N]}^{[x_\mathcal{M}]}$.

In case of maximal leakage, the first phase of the scheme can be viewed as a means to retrieve the auxiliary variable $E_{[M]}$, or equivalently Θ. Given this, the maximal leakage on $y_{[M]}$ from the second phase is

$$\begin{aligned}&\max_{x\in[0:R]^d} \max_{\mathcal{I}:|\mathcal{I}| \text{ fixed}} \mathcal{L}(y_{[M]} \to A_{[N]}^{[x_\mathcal{M}]}|x, \mathcal{I}, A_{[N]}^{[x_\mathcal{I}]})\\ &= \max_{x\in[0:R]^d} \max_{\mathcal{I}:|\mathcal{I}| \text{ fixed}} \log_q\big[\text{number of unique combinations of}\\ &\qquad (\|y_{i,\mathcal{M}} - x_\mathcal{M}\|^2, i \in \Theta)\big].\end{aligned} \tag{113}$$

Using the above simplification, we proceed to compute the exact MI and Max leakage values on the synthetic dataset, setting $R = 3$, $d = 3$, and the database size $M = 3$. Moreover, the set of immutable features $\mathcal{I}$ is assumed to be uniformly distributed over $[d]$, while having a fixed cardinality. For the single-phase scheme, $F = d$, $L = R^2 d + 1 = 28$ and q is chosen as 757 which is the smallest prime greater than $d(L-1)R^2 + R^2 d = 756$. Table IV presents the computed leakage values under these conditions, with varying $|\mathcal{I}|$ and logarithms taken to base $q = 757$ for both schemes.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
LEAKAGE RESULTS FOR I-PCR SCHEMES WITH VARYING $|\mathcal{I}|$ TO THE BASE $q = 757$

$\lvert\mathcal{I}\rvert$	Single-phase I-PCR		Two-phase I-PCR	
	MI Leakage	Max Leakage	MI Leakage	Max Leakage
0	1.1432	1.3039	0.9422	1.1721
1	1.4492	1.6534	0.4200	0.9728
2	1.4492	1.6534	0.0977	0.4519
3	1.1432	1.3039	0.0000	0.0000

As shown in Table IV, the leakage for the two-phase scheme is always lower than that for the single-phase scheme. For the single-phase scheme, the leakage values are equal for $|\mathcal{I}|$ and $d - |\mathcal{I}|$. As opposed to the single-phase scheme, the leakage value of the two-phase scheme decreases as $|\mathcal{I}|$ increases. This is because, the expected size of Θ reduces, revealing less information on $\mathcal{D}$. The case of $|\mathcal{I}| = d$ is interesting since the user is guaranteed to not find a counterfactual. For the two-phase scheme, the leakage is 0 since the user knows that $\Theta = \emptyset$ since x is rejected, hence $x \notin \mathcal{D}$. However, this is non-zero for the single-phase scheme.

VII. CONCLUSION

In this work, we investigated the problem of counterfactual retrieval, with information-theoretic privacy of the applicant against an institution. We started with formulating basic PCR, where the user is flexible to change any of their features to obtain the closest accepted sample. We proposed three schemes, which ensure varying levels of protection for the institution's database. We modified the PCR formulation to include the immutability constraint which fixes the values of certain features to not change in the counterfactual. Under this setup, we proposed two schemes with perfect user privacy, and varying levels of database privacy. Finally, we accounted for user actionability where the user can have different preference to change different attributes, and showed that our schemes can be extended to accommodate this setting. In addition to perfect user privacy, our schemes guarantee database privacy against the user, to some degree. Our experimental results support our findings on the comparative leakage values. Depending on the privacy budget, the institution can choose to follow one scheme over another.

APPENDIX A
BASELINE PCR WITH OTHER METRICS

If the metric for closeness is the ℓ_k norm with k even, i.e., $d(x, y) = \|y - x\|^k$, then the chosen $\mathbb{F}_q$ grows to have $q > R^k d$. The user sends the same query as (14). Upon receiving the query, the servers compute their answer for $y_i \in \mathcal{D}$ as

$$A_n^{[x]}(i) = \|y_i - (x + \alpha_n Z)\|^k + \sum_{\ell=1}^{k-1} \alpha_n^\ell Z'_\ell. \tag{114}$$

Correspondingly, the user decodes the answers to first compute

$$\hat{A}_n^{[x]}(i) = A_n^{[x]}(i) - \alpha_n^k Z^k \mathbf{1}_k \tag{115}$$

where 1_k is the all-ones column vector, and obtain $d(y_i, x)$ since

$$\begin{bmatrix} \hat{A}_1^{[x]}(i) \\ \hat{A}_2^{[x]}(i) \\ \vdots \\ \hat{A}_k^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_k \begin{bmatrix} d(y_i, x) \\ I_1(i) \\ \vdots \\ I_{k-1}(i) \end{bmatrix} \tag{116}$$

and $\boldsymbol{V}_k$ is invertible. Here, $I_\ell(i)$, $\ell \in [k-1], i \in [M]$ is the interference term

$$I_\ell(i) = (-1)^{\ell-1} \binom{k}{\ell} (y_i - x)^{k-\ell} Z^\ell + Z'_\ell. \tag{117}$$

The number of servers required for decodability is $N = k$. For the dot product metric, i.e., $d(x, y) = y^T x$, our scheme requires $N = 2$ and $q > R^2 d$. The answer returned by server n is

$$A_n^{[x]}(i) = y_i^T (x + \alpha_n Z) + \alpha_n Z', \tag{118}$$

and the decodability at the user follows since

$$\begin{bmatrix} A_1^{[x]}(i) \\ A_2^{[x]}(i) \end{bmatrix} = \boldsymbol{V}_2 \begin{bmatrix} y_i^T x \\ y_i^T Z + Z' \end{bmatrix}. \tag{119}$$

The user assigns θ^* to the $i \in [M]$ that maximizes $y_i^T x$.

APPENDIX B
PROOF OF LEMMA 1

We want to show that if $x \in \mathcal{D}_c$, then, decodability can be ensured from the masked distances. First, assume that $x \in \{x_1, \ldots, x_K\}$. To prove the required result, it suffices to prove that the random variable used preserves the order of the distances between the samples, i.e., $d_i(x) - d_j(x) > 0$ if y_j is closer to x and $d_i(x) - d_j(x) < 0$ if y_i is closer to x. Let $|d_i(x) - d_j(x)| = \ell$. Then,

$$d_i(x) - d_j(x) = \begin{cases} \ell, & y_j \text{ is closer to } x, \\ -\ell, & y_i \text{ is closer to } x. \end{cases} \tag{120}$$

Now, assume y_j is closer, thus

$$\begin{aligned} d_i(x) - d_j(x) + \mu(i) - \mu(j) &= \ell + \mu(i) - \mu(j) && (121)\\ &\geq \ell - (d_{\min} - 1) && (122)\\ &> 0, && (123) \end{aligned}$$

and if y_i is closer, we have

$$\begin{aligned} d_i(x) - d_j(x) + \mu(i) - \mu(j) &= -\ell + \mu(i) - \mu(j) && (124)\\ &\leq -\ell + (d_{\min} - 1) && (125)\\ &< 0. && (126) \end{aligned}$$

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

Now, assume $x \in \mathcal{D}_c$. Using the same approach as above, and the definition of $x \in \mathcal{D}_c \setminus \{x_1, \ldots, x_K\}$, we have $|d_i(x) - d_j(x)| \geq d_{\min}$, which proves the required result.

Appendix C
Proof of Lemma 2

Note that

$$\begin{aligned}
&|(y_i - x_k)^T ((y_i - x_k) \circ w) - (y_j - x_k)^T \left((y_j - x_k) \circ w\right)| \\
&= \left| \sum_{\ell} w_\ell (y_i(\ell) - x_k(\ell))^2 - w_\ell (y_j(\ell) - x_k(\ell))^2 \right| && (127)\\
&\geq \left| \sum_{\ell} (y_i(\ell) - x_k(\ell))^2 - (y_j(\ell) - x_k(\ell))^2 \right| && (128)\\
&= \left| d(y_i, x_k) - d(y_j, x_k) \right|. && (129)
\end{aligned}$$

Thus, $\min_k \min_{i,j} |(y_i - x_k)^T ((y_i - x_k) \circ w) - (y_j - x_k)^T \left((y_j - x_k) \circ w\right)| \geq d_{\min}$.

Appendix D
Leakage Computation

For both the MI and Max leakage, the interference terms, being independent of $y_1, \ldots, y_M$ by one-time padding, do not contribute to leakage. To compute the MI leakage values in Table II we make the following assumptions: 1) Rejected instance x are equi-probable over the $[0 : R]^d$ space. Therefore, $\mathbb{P}_X(x) = \frac{1}{(R+1)^d}$ (denote by ρ_X). 2) Given the instance x, the M-tuples $y = (y_1, \ldots, y_M)$ are equi-probable over the $[0 : R]^d \setminus \{x\}$ space. Therefore, $\mathbb{P}_{Y|X}(y|x) = \frac{1}{(R+1)^d - 1 P_M}$ (denote by $\rho_{Y|X}$) where ${}_aP_b = \frac{a!}{(a-b)!}$ is the number of permutations.

Here, we present the approach we have taken to compute the leakage of the Baseline PCR scheme. A similar approach can be used to compute the values for Diff-PCR and Mask-PCR. Let $d(x) = (d_1(x), \ldots, d_M(x))$. With the first assumption above, we simplify the MI leakage as follows:

$$\begin{aligned}
&I(y_1, \ldots, y_M; d_1(x), \ldots, d_M(x)|x) \\
&= H(d_1(x), \ldots, d_M(x)|x) \\
&\quad - \underbrace{H(d_1(x), \ldots, d_M(x)|x, y_1, \ldots, y_M)}_{=0} && (130)\\
&= H(d_1(x), \ldots, d_M(x)|x) && (131)\\
&= -\sum_{x \in \mathcal{X}} \sum_{d(x) \in \mathcal{D}(x)} \mathbb{P}_{D(X),X}(d(x), x) \log(\mathbb{P}_{D(X)|X}(d(x)|x)) && (132)\\
&= -\sum_{x \in \mathcal{X}} \sum_{d(x) \in \mathcal{D}(x)} \mathbb{P}_{D(X)|X}(d(x)|x)\, \mathbb{P}_X(x) \log(\mathbb{P}_{D(X)|X}(d(x)|x)) && (133)\\
&= -\rho_X \sum_{x \in \mathcal{X}} \sum_{d(x) \in \mathcal{D}(x)} \mathbb{P}_{D(X)|X}(d(x)|x) \log(\mathbb{P}_{D(X)|X}(d(x)|x)). && (134)
\end{aligned}$$

Now, we focus on computing the conditional pmf $\mathbb{P}_{D(X)|X}(d(x)|x)$. Note that

$$\begin{aligned}
&\mathbb{P}_{D(X)|X}(d(x)|x) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}_{D(X),Y|X}(d(x), y|x) && (135)\\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}_{D(X)|Y,X}(d(x)|y, x)\, \mathbb{P}_{Y|X}(y|x) && (136)\\
&= \rho_{Y|X} \sum_{y \in \mathcal{Y}} \mathbb{P}_{D(X)|Y,X}(d(x)|y, x) && (137)\\
&= \rho_{Y|X} \sum_{y \in \mathcal{Y}} \mathbb{1}\left[d(x) = (\|y_1 - x\|^2, \ldots, \|y_M - x\|^2)\right], && (138)
\end{aligned}$$

where (137) is due to the second assumption above. Let us define $\mathcal{C}(d(x), x) := \sum_{y \in \mathcal{Y}} \mathbb{1}[d(x) = (\|y_1 - x\|^2, \ldots, \|y_M - x\|^2)]$, as the value of the histogram of $(D(X)|X = x)$ at $d(x)$. We compute this by counting the M-tuples $(\|y_1 - x\|^2, \ldots, \|y_M - x\|^2)$ while iterating over all possible M-tuples $y = (y_1, \ldots, y_M)$ for the given x. Next, we substitute this result back into (138) to obtain

$$\begin{aligned}
&I(y_1, \ldots, y_M; d_1(x), \ldots, d_M(x)|x) \\
&= -\rho_X \rho_{Y|X} \sum_{x \in \mathcal{X}} \sum_{d(x) \in \mathcal{D}(x)} \mathcal{C}(d(x), x) \log(\rho_{Y|X} \mathcal{C}(d(x), x)). && (139)
\end{aligned}$$

For Max leakage, we do not require assumptions on the probability mass functions of x and $y_1, \ldots, y_M$. The leakage for Baseline-PCR is given by

$$\begin{aligned}
&\mathcal{L}\big((y_1, \ldots, y_M) \to (d_1(x), \ldots, d_M(x))\big) \\
&= \mathcal{L}(y \to d(x)) && (140)\\
&= \max_{x \in \mathcal{X}} \left(\log_q \left[\sum_{d(x) \in \mathcal{D}(x)} \max_{y \in \mathcal{Y}} \mathbb{P}_{D(X)|X,Y}(d(x)|x, y) \right] \right) && (141)\\
&= \max_{x \in \mathcal{X}} \left(\log_q \left[\sum_{d(x) \in \mathcal{D}(x)} \max_{y \in \mathcal{Y}} \mathbb{1}\left[d(x) = (\|y_1 - x\|^2, \ldots, \|y_M - x\|^2)\right] \right] \right) && (142)\\
&= \max_{x \in \mathcal{X}} \left(\log_q \left[\sum_{\substack{\text{unique combinations of} \\ (\|y_1 - x\|^2, \ldots, \|y_M - x\|^2)}} 1 \right] \right) && (143)\\
&= \max_{x \in \mathcal{X}} \left(\log_q \left[\text{number of unique combinations of } (\|y_1 - x\|^2, \ldots, \|y_M - x\|^2) \right] \right). && (144)
\end{aligned}$$

References

[1] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," in *A Practical Guide*, vol. 10, 1st ed., Cham, Switzerland: Springer, 2017, pp. 10–5555.

[2] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3995299

[3] D. Brughmans, P. Leyman, and D. Martens, "NICE: An algorithm for nearest instance counterfactual explanations," *Data Mining Knowl. Discovery*, vol. 38, no. 5, pp. 2665–2703, Apr. 2023.

[4] S. Upadhyay, S. Joshi, and H. Lakkaraju, "Towards robust and reliable algorithmic recourse," in *Proc. Adv. Neural Inf. Process Syst.*, Dec. 2021, pp. 16926–16937.

[5] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust counterfactual explanations for neural networks with probabilistic guarantees," in *Proc. ICML*, Jul. 2023, pp. 12351–12367.

[6] S. Dutta, J. Long, S. Mishra, C. Tilli, and D. Magazzeni, "Robust counterfactual explanations for tree-based ensembles," in *Proc. ICML*, Jun. 2022, pp. 5742–5756.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

[7] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust algorithmic recourse under model multiplicity with probabilistic guarantees," *IEEE J. Sel. Areas Inf. Theory*, vol. 5, pp. 357–368, 2024.

[8] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 344–350.

[9] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 607–617.

[10] S. Verma, J. P. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *ACM Comput. Surveys*, pp. 1–42, Oct. 2020.

[11] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining Knowl. Discovery*, vol. 38, no. 5, pp. 2770–2824, Sep. 2024.

[12] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," 2021, *arXiv:2111.00358*.

[13] M. Pawelczyk, H. Lakkaraju, and S. Neel, "On the privacy risks of algorithmic recourse," in *Proc. AISTATS*, Apr. 2023, pp. 9680–9696.

[14] F. Yang, Q. Feng, K. Zhou, J. Chen, and X. Hu, "Differentially private counterfactuals via functional mechanism," 2022, *arXiv:2208.02878*.

[15] S. Goethals, K. Sörensen, and D. Martens, "The privacy issue of counterfactual explanations: Explanation linkage attacks," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 5, pp. 1–24, Oct. 2023.

[16] S. Pentyala, S. Sharma, S. Kariyappa, F. Lecue, and D. Magazzeni, "Privacy-preserving algorithmic recourse," 2023, *arXiv:2311.14137*.

[17] U. Aïvodji, A. Bolot, and S. Gambs, "Model extraction from counterfactual explanations," 2020, *arXiv:2009.01884*.

[18] Y. Wang, H. Qian, and C. Miao, "DualCF: Efficient model extraction attack from counterfactual explanations," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2022, pp. 1318–1329.

[19] P. Dissanayake and S. Dutta, "Model reconstruction using counterfactual explanations: A perspective from polytope theory," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, Dec. 2024, pp. 83397–83429.

[20] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations," *ACM Comput. Surveys*, vol. 55, no. 5, pp. 1–29, Dec. 2022.

[21] S. Servan-Schreiber, S. Langowski, and S. Devadas, "Private approximate nearest neighbor search with sublinear communication," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 911–929.

[22] S. Vithana, M. Cardone, and F. P. Calmon, "Private approximate nearest neighbor search for vector database querying," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 3666–3671.

[23] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

[24] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[25] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[26] X. Yao, N. Liu, and W. Kang, "The capacity of private information retrieval under arbitrary collusion patterns for replicated databases," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6841–6855, Oct. 2021.

[27] K. Banawan and S. Ulukus, "Private information retrieval through wiretap channel II: Privacy meets security," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4129–4149, Jul. 2020.

[28] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[29] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[30] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[31] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3407–3416, Jun. 2020.

[32] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[33] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3590–3602, Jun. 2020.

[34] S. Meel, X. Kong, T. J. Maranzatto, I. Tamo, and S. Ulukus, "Private information retrieval on multigraph-based replicated storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2025, pp. 1–6.

[35] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[36] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7613–7627, Nov. 2019.

[37] L. Holzbaur, R. Freij-Hollanti, and C. Hollanti, "On the capacity of private information retrieval from coded, colluding, and adversarial servers," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

[38] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[39] Z. Jia, H. Sun, and S. A. Jafar, "Cross subspace alignment and the asymptotic capacity of X -secure T -private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5783–5798, Sep. 2019.

[40] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 2953–2964, Dec. 2018.

[41] M. Nomeir, S. Vithana, and S. Ulukus, "Asymmetric X-secure T-private information retrieval: More databases is not always better," in *Proc. 58th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2024, pp. 1–6.

[42] Z. Wang and S. Ulukus, "Symmetric private information retrieval at the private information retrieval rate," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 2, pp. 350–361, Jun. 2022.

[43] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6617–6634, Nov. 2020.

[44] Q. Wang and M. Skoglund, "Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5160–5175, Aug. 2019.

[45] N. Woolsey, R.-R. Chen, and M. Ji, "Uncoded placement with linear sub-messages for private information retrieval from storage constrained databases," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6039–6053, Oct. 2020.

[46] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

[47] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2635–2652, Apr. 2022.

[48] J. Cheng, N. Liu, W. Kang, and Y. Li, "The capacity of symmetric private information retrieval under arbitrary collusion and eavesdropping patterns," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3037–3050, 2022.

[49] S. Ulukus, S. Avestimehr, M. Gastpar, S. A. Jafar, R. Tandon, and C. Tian, "Private retrieval, computing, and learning: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 729–748, Mar. 2022.

[50] A. Cassandra. (2019). *2.1 for DSE, Data Replication*. Accessed: Feb. 21, 2025. [Online]. Available: https://docs.datastax.com/en/archived/cassandra/2.1/cassandra/architecture/architectureDataDistributeReplication_c.html

[51] Hadoop Distributed File System. (2025). *Architecture Guide-Data Replication*. Accessed: Feb. 21, 2025. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Data

[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.

[53] Z. Jia and S. A. Jafar, "Cross subspace alignment codes for coded distributed batch computation," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2821–2846, May 2021.

[54] Z. Jia and S. A. Jafar, "X-secure T-private federated submodel learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[55] S. Vithana and S. Ulukus, "Private read update write (PRUW) in federated submodel learning (FSL): Communication efficient schemes with and without sparsification," *IEEE Trans. Inf. Theory*, vol. 70, no. 2, pp. 1320–1348, Feb. 2024.

[56] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[57] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decis. Support Syst.*, vol. 47, no. 4, pp. 547–553, Nov. 2009.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.

[58] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. (2016). *Machine Bias*. ProPublica. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Shreya Meel (Graduate Student Member, IEEE) received the B.E. degree in electronics and telecommunication engineering from Jadavpur University, India, in 2019, and the M.Tech. degree in electrical communication engineering from Indian Institute of Science, Bengaluru, in 2021, with a specialization in communications and networks. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park, under the supervision of Prof. Sennur Ulukus. Her research interests include information-theoretic security and privacy, distributed storage, and private information retrieval.

Mohamed Nomeir (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Alexandria University, Egypt, and the M.Sc. degree from the Electronics and Communications Department, American University, Cairo, Egypt. He is currently pursuing the Ph.D. degree with the ECE Department, UMD. Currently, his research focuses on topics related to classical and quantum information theoretic privacy and security. He received the Medhat Haroun award for scholarly excellence as the best thesis of the academic year for his master's thesis.

Pasan Dissanayake (Graduate Student Member, IEEE) received the B.Sc. degree in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park. His research interests include explainability-driven approaches for efficient and sustainable AI applications. He was a recipient of George Corcoran Memorial Award for a graduate student and the Outstanding Graduate Assistant Award in 2025.

Sanghamitra Dutta (Member, IEEE) received the B.Tech. degree in electrical and computer engineering from IIT Kharagpur and the master's and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University.

She has been an Assistant Professor with the Department of Electrical and Computer Engineering, University of Maryland, College Park, since Fall 2022. She is also affiliated with the Center for Machine Learning (CML), UMIACS, the Department of Computer Science, the Values-Centered Artificial Intelligence (VCAI), the Applied Mathematics and Statistics, and Scientific Computation (AMSC), and the Artificial Intelligence Interdisciplinary Institute at Maryland (AIM). Prior to joining UMD, she was a Senior Research Associate with JPMorgan Chase AI Research New York in the Explainable AI Centre of Excellence (XAI CoE). She has also pursued research internships at IBM Research and Dataminr. Her research interests include efficient, reliable, and trustworthy machine learning. She is particularly interested in addressing the challenges concerning explainability, efficiency, robustness, and ethics by bringing in novel foundational perspectives deep-rooted in information theory, statistics, causality, and optimization. Her research has featured in New Scientist and Montreal AI Ethics Brief, and also been adopted as part of the fair-lending model review at JPMorgan. In her prior work, she has also examined problems in reliable computing, proposing novel algorithmic solutions for large-scale distributed machine learning, using tools from coding theory (an emerging area called "coded computing").

Dr. Dutta was a recipient of the 2024 NSF CAREER Award, the 2024 George Corcoran Memorial Award, the 2023 JPMorgan Faculty Award, the 2022 Simons Institute Fellowship for Causality, the 2021 AG Milnes Outstanding Thesis Award from CMU, and the 2019 K&L Gates Presidential Fellowship.

Sennur Ulukus (Fellow, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University and the Ph.D. degree in electrical and computer engineering from the Wireless Information Network Laboratory (WINLAB), Rutgers University.

She is currently a Distinguished University Professor and the Anthony Ephremides Professor in information sciences and systems with the Department of Electrical and Computer Engineering, University of Maryland, College Park, where she also holds a joint appointment with the Institute for Systems Research (ISR). She is a Distinguished Scholar-Teacher of the University of Maryland. Prior to joining UMD, she was a Senior Technical Staff Member with AT&T Labs-Research. Her research interests include information theory, wireless communications, machine learning, and signal processing and networks; with recent focus on age of information, private information retrieval, blockchain consensus, quantum information, machine learning for wireless, federated learning, and large language models.

Dr. Ulukus received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 IEEE Communications Society Best Tutorial Paper Award, the 2020 IEEE Communications Society Women in Communications Engineering (WICE) Outstanding Achievement Award, the 2020 IEEE Communications Society Technical Committee on Green Communications and Computing (TCGCC) Distinguished Technical Achievement Recognition Award, the 2024 IEEE Communications Society Communication Theory Technical Committee (CTTC) Technical Achievement Award, the 2005 NSF CAREER Award, the 2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She is the TPC Chair of 2021 IEEE Globecom and the TPC Co-Chair of 2024 IEEE Globecom, 2024 IEEE DySPAN, 2023 IEEE MILCOM, 2019 IEEE ITW, 2017 IEEE ISIT, 2016 IEEE Globecom, 2014 IEEE PIMRC, and 2011 IEEE CTW. She is a Senior Editor of IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKINGin 2020. She was an Area Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONSfrom 2019 to 2023, an Area Editor of IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKINGfrom 2016 to 2020, an Editor of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS-Series on Green Communications and Networking from 2015 to 2016, an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORYfrom 2007 to 2010, and an Editor of IEEE TRANSACTIONS ON COMMUNICATIONSfrom 2003 to 2007. She was a Guest Editor of IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORYfrom 2021, 2022, and 2023, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONSfrom 2008, 2015, 2021, and 2022; IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSINGin 2021; *Journal of Communications and Networks*in 2012; and IEEE TRANSACTIONS ON INFORMATION THEORYin 2011. She was a Distinguished Lecturer of the IEEE Information Theory Society from 2018 to 2019.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on October 26,2026 at 22:04:07 UTC from IEEE Xplore. Restrictions apply.