PRIVATE INFORMATION RETRIEVAL WITH SIDE INFORMATION

A Thesis

by

BRENDEN MARK GARCIA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Alex Sprintson
Committee Members,	I-Hong Hou
	Catherine Yan
Head of Department,	Miroslav Begovic

May 2018

Major Subject: Computer Engineering

ABSTRACT

The objective of the classical Private Information Retrieval (PIR) problem is to enable a user to download a message from a database that is replicated across a collection of non-colluding servers without revealing the identity of the demanded message to the servers. In the classical PIR problem the user has no prior information about the content of messages in the database. It is easy to verify in the special case of the PIR problem when there is only one server in the system, the user must download all messages from the database in order keep information about the message they want private.

In a real environment the user may have other sources to download and obtain messages from such as trusted peer-to-peer communication. In this way, the user has the potential to obtain some of the messages that are contained in the database to use as side information in a PIR scheme with the servers. Accordingly, we introduce the Private Information Retrieval with Side Information (PIR-SI) problem that focuses on settings in which the user has side information about some messages in the database. To capture the different levels of privacy a user may want to enforce in PIR-SI schemes two metrics of privacy, W-privacy and (W, S)-privacy, are introduced. W-privacy only protects information about the identity of the message that the user wants and is most similar to the measure of privacy in the original PIR problem. (W, S)-privacy protects the identity of the wanted message as well as the indentities of the messages they have as side information and is a stronger sense of privacy than W-privacy. When enforcing either measure of privacy the user no longer has to download all the messages in the database, even if there is only one server in the system; side information reduces the amount of data that one has to download in a PIR scheme.

The first case of the PIR-SI problem that we consider is when the user has M messages for side information and wants a different message from the database of K messages. When there is only one server in the system, we show that the optimal download rate for a W-private scheme is $\frac{K}{M+1}^{-1}$ and the optimal download rate for a (W, S)-private scheme is $\frac{1}{K-M}$. When there is more than one server in the system a W-private scheme is presented that has a larger rate than the classical PIR

scheme, but its optimality is not shown.

The second case of the PIR-SI problem that is considered is when the user has M messages as side information, the user wants $D > 1$ distinct messages from the database, and there is only one server in the system. In the case when $M = 1$ a (W, S)-private optimal scheme is presented and shown to be optimal. In the case when $M \geq D$ and $D = 2$ a W-private scheme that can increase the rate from the (W, S)-private scheme with the same parameters is presented. This scheme's optimality reminds an open problem. We highlight the difficulty of finding an optimal scheme and determining the capacity of the multi-message PIR-SI problem.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Alex Sprintson - advisor, and Professor I-Hong Hou - member, of the Department of Electrical and Computer Engineering, and Professor Catherine Yan - member, of the Department of Mathematics.

The analyses depicted in Chapter 3 were conducted in collaboration with Swanand Kadhe of the Department of Electrical Engineering and Computer Sciences of UC Berkeley, Salim El Rouayheb of the Department of Electrical and Computer Engineering of Rutgers, and Anoosheh Heidarzadeh of the Department of Electrical and Computer Engineering of Texas A&M, and were published in 2017 in an article listed in the proceedings of the Allerton Conference on Communication and Control.

All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from the Office of Graduate and Professional Studies of Texas A&M University and assistantships provided by the Department of Electrical and Computer Engineering of Texas A&M. Part of the material is based on work supported by the National Science Foundation through grant number 1718658.

TABLE OF CONTENTS

Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The classical Private Information Retrieval (PIR) problem introduced in [1] considers how a user can query a collection of non-colluding servers in order to obtain a desired bit from a database that is replicated on each of the servers without revealing the identity of the bit the user wants. In this setting, when there is only one server it is easy to verify that the user must download all the bits in the database in order to keep the identity of the bit the user wants private. When there are more servers in the problem the user does not have to download the entire database from each server resulting in a better download rate.

In [2] the authors studied the more realistic case when the messages stored in the database are made up of a large number of bits. In this case the authors of [2] considered privacy in an information-theoretic sense, defining the privacy metric of the PIR problem to be related to the mutual information between the queries that the user sends to the servers and the indices of the messages that the user wants from the database. The optimal download cost of any PIR scheme was derived using information theoretic tools and an optimal scheme for an arbitrary number of servers was described.

In many scenarios the user has some form of prior information about the messages in the database. A user could obtain this information through eavesdropping or through trusted peer-to-peer communication for instance. To address these instances the Private Information Retrieval with Side Information (PIR-SI) problem is introduced to study schemes where a user has prior side information about messages on the servers. Works that address the PIR-SI problem differentiate themselves by the variations in the model used for a user's side information [3, 4].

This work focuses on the side information model where the user has a subset of messages from the database as side information, and the servers do not know which messages are available to the user to use as side information. With the introduction of side information to the classical PIR problem model, we introduce two privacy metrics to capture different types of privacy the user may want against the server. (W, S)-privacy is introduced when the user wants to protect both the index

of their wanted message and the indices of the messages in their side information. W-privacy is introduced when the user wants only to protect information about the message they want from the server, but are allowed to leak information about their side information.

This work focuses first on the case when a database of K messages is replicated across N servers, a user wants one of the K messages, and the user has M messages as side information. When considering situations in which there is one server in the system and enforcing (W, S)-privacy it is shown that a Maximum Distance Separable (MDS) coding scheme that downloads $K - M$ messages from the server is optimal. When enforcing W-privacy we present a partition based coding scheme that downloads $\lceil \frac{K}{M+1} \rceil$ messages from the server is an optimal scheme. It is noted that when the user has no side information ($M = 0$) these results agree with the previous PIR results. When there are $N \geq 2$ servers in the system we present a scheme that combines the partition based coding scheme with the scheme presented in [2] in order to lower the rate when compared to classical PIR.

It is known that in the PIR problem when multiple messages are requested by the user they can be obtained more efficiently when retrieved jointly as opposed to downloading them through multiple PIR rounds [5]. The second situation that we consider is when the user wants more than one message from the database on a single server; the user wants some D messages from the database and has M messages as side information. We show that the MDS scheme that is used in the single message case ($D = 1$) achieves (W, S)-privacy in the multi-message case for any value of D and is optimal in the case when $M = 1$. We show that the partition based coding scheme used in the single demanded message case can be generalized for $D = 2$ and $M \geq D$ in order to achieve W-privacy and reduce download cost to $\frac{2K}{D+\frac{M}{D}}$ in the multi-message case. The generalized partition based scheme has a lower download cost then the MDS scheme for large values of K.

1.1 Related Works

The classical PIR problem was introduced by Chor *et al.* [1]. In the classical PIR problem there is a database of n bits replicated across a collection of non-colluding servers and a user that wishes to download a single bit from the database without revealing to the servers which bit

they are interested in downloading. In [1] the authors show that when there is only one server in the system, the user must download every bit in the database in order to maintain privacy. The authors then demonstrate that when there is more than one server the download cost can be reduced. Because the database in the problem stores bits, the upload cost of the user's queries was taken into consideration when computing the rate of different schemes. The user does not have access to side information of any form in the work as well.

The work of Sun and Jafar [2] studied the capacity of PIR in an information theoretic sense. In [2] the messages replicated across the servers consist of an arbitrarily large number of bits, rendering the upload cost negligible. The authors characterize the optimal rate and give an optimal scheme for the classical PIR problem where the user wants one message from the database. The results of [2] showed again that in the case of one server the best a user can do in order to preserve privacy is to download all messages from the database. In this work the user does not have any prior information about messages on the server.

The work of Banawan and Ulukus [5] studied the classical PIR problem when the user demands more than one message from the database. The work shows that in classical PIR when the user demands more than one message from the database the user can download the messages more efficiently together, rather than running single message PIR schemes repeatedly to retrieve their desired messages. The authors give bounds for the optimal rate with respect to parameters in the problem, and give nearly optimal schemes for different operating points. The user in this problem does not have access to prior information about messages on the server.

Different variations of the classical PIR problem have been studied as well. The works [6, 7, 8, 9, 10, 11] look at PIR when the servers are allowed to collude, experience failures, and introduce Byzantine errors into the answers sent back to the user. The problem of Symmetric PIR considers situations where the user wants to protect information about the index of the message they want from the server and the server wants to keep the content of the messages that the user does not want private from the user [12, 13, 14, 15]. Classical PIR when the database is coded across servers, as apposed to replicated across the servers, has been studied [16, 17, 18]. There are a number of

works that introduce side information to the user in the PIR problem as well.

The work of Tandon [3] studies the capacity of the PIR problem when a particular side information model is used. The side information that the user has in [3] is known to the server and can be any function of the messages in the database. The work of Wei et. al [4] looks at the PIR-SI problem when the side information of the user is unknown to the server, but it takes the form of a random number of bits from each message of the database. In this side information model the user gets some information about each message, including the message that is demanded by the user. In contrast the side information in this work is only considered to be a subset of messages from the database that doesn't include a user's wanted message and the particular realization of a user's side information is not known to the server.

The work of Chen et. al [19] considers the same PIR with side information model as this work. The authors only consider the case when the user is interested in protecting information about demanded messages jointly with their side information set; they do not consider the case when the user wants only to protect information about their demanded message. The authors compute the capacity of PIR-SI schemes when the user is interested in protecting their demand and side information jointly for an arbitrary set of parameters by presenting a scheme based on MDS codes and the work of [2] and showing its optimality.

2. PROBLEM SETUP

For a positive integer K let $[K] = \{1, 2, \ldots, K\}$. For a set of K elements $\{X_1, \ldots, X_K\}$ and a set $S \subseteq [K]$, let $X_S = \{X_i : i \in S\}$. Let $\mathbb{F}_q$ be a finite field of order q. For $S \subseteq [K]$ let $\mathbf{1}_S \in \{0,1\}^K$ be a vector whose i^{th} entry is a 1 if $i \in S$ and a 0 otherwise. Given a directed graph G with vertex set V, for $i \in V$ denote the out-neighbors of i as $\mathcal{N}(i) = \{v : \text{there exists an arc } (i, v)\}$. We denote the entropy function by $H(x)$.

In the Private Information Retrieval with Side Information (PIR-SI) problem there is a database of K ($K \geq 1, K \in \mathbb{Z}$) messages that are replicated across N ($N \geq 1, N \in \mathbb{Z}$) servers. The messages are each t bits in length and are denoted by $X_1, X_2, \ldots, X_K$. All X_i's are identically and independently distributed, each being distributed uniformly at random from the finite field $\mathbb{F}_{2^t}$. This relationship is captured in the equation

$$H(X_1, X_2, \ldots, X_K) = \sum_{i=1}^{K} H(X_i) = Kt. \tag{2.1}$$

There is one user in the system that is interested downloading D messages from the servers privately, where D is a positive integer and $1 \leq D \leq K$. The user then is interested in downloading the messages X_W where $W \subseteq [K], |W| = D$. The set W is referred to as the *demand index set* and X_W is referred to as the *demand* of the user.

The user also has a subset of M messages from the database as side information. Let $S \subset [K]$ be the set of indices of the M messages available to the user ($|S| = M$); the user has the messages X_S as side information. S is referred to as the *side information index set* and X_S is the *side information*.

Let $\mathbf{W}$ and $\mathbf{S}$ be random sets corresponding to the demand index set and the side information index set. In this work it is assumed that the demand index set is distributed such that all possible

demand index sets have a non-zero probability,

$$\mathbb{P}(\mathbf{W} = W) \neq 0, \forall\, W \in \{\mathcal{W} \subseteq [K] : |\mathcal{W}| = D\}. \tag{2.2}$$

The distributions on $\mathbf{W}$ and $\mathbf{S}$ that are assumed when analyzing PIR-SI schemes are given below:

$$\mathbb{P}(\mathbf{W} = W) = \begin{cases} \frac{1}{\binom{K}{D}} & \text{if } W \subseteq [K] \text{ and } |W| = D \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

$$\mathbb{P}(\mathbf{S} = S | \mathbf{W} = W) = \begin{cases} \frac{1}{\binom{K-D}{M}} & \text{if } S \subseteq [K] \setminus W \text{ and } |S| = M \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Denote by $\mathcal{S}$ the set $\mathcal{S} \triangleq \{(W, S) : W \subseteq [K], S \subseteq [K], |S| = M, |W| = D, \text{ and } S \cap W = \emptyset\}$ The joint probability of $\mathbf{W}$ and $\mathbf{S}$ is given below,

$$\mathbb{P}(\mathbf{S} = S, \mathbf{W} = W) = \begin{cases} \frac{1}{\binom{K}{D}\binom{K-D}{M}} = \frac{1}{\binom{K}{M}\binom{K-M}{D}} & \text{if } (W, S) \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

It is assumed as well that the server only knows the value of M and the a priori distributions $\mathbb{P}(\mathbf{W})$, $\mathbb{P}(\mathbf{S}|\mathbf{W})$, and $\mathbb{P}(\mathbf{W}, \mathbf{S})$ but not the realizations of W and S.

In order to download the messages in X_W given X_S, the user sends a query denoted $Q_i^{[W,S]}$ to server i. Server i responds to the query with an answer $A_i^{[W,S]}$. The set of queries, $Q^{[W,S]}$, and answers, $A^{[W,S]}$, are referred to as the *Private Information Retrieval with Side Information* (PIR-SI) scheme. A valid PIR-SI scheme should satisfy the following requirements:

1. Any query $Q_i^{[W,S]}$ should be a (potentially stochastic) function of W, S, and X_S.

2. Any answer $A_i^{[W,S]}$ is a deterministic function of the query sent to server i and the messages,

$$H(A_i^{[W,S]}|Q_i^{[W,S]}, X_1, X_2, \ldots, X_K) = 0. \tag{2.6}$$

3. The demand set of the user should be decodable with the answer and the side information set of the user,

$$H(X_W|A^{[W,S]}, X_S) = 0. \tag{2.7}$$

In addition to the above requirements a valid PIR-SI scheme must uphold one of two privacy conditions denoted as W-privacy and (W, S)-privacy.

Definition 1. *W-privacy: To achieve W-privacy the server should not learn anything about the demand index set of the user after seeing the query. To capture this the W-privacy is defined as:*

$$I(\mathbf{W}; Q^{[\mathbf{W},\mathbf{S}]}, X_1, \ldots, X_K) = 0. \tag{2.8}$$

Definition 2. *(W, S)-privacy: To achieve (W, S)-privacy the server should not learn anything about the demand index set and side information index set of the user jointly. To capture this (W, S)-privacy is defined as:*

$$I(\mathbf{W}, \mathbf{S}; Q^{[\mathbf{W},\mathbf{S}]}, X_1, \ldots, X_K) = 0. \tag{2.9}$$

In order to show W-privacy and (W, S)-privacy these definitions are not used directly. In order to show W-privacy it is sufficient to show that the queries sent by the user are independent of the demand index set and satisfy

$$\mathbb{P}(\mathbf{W}|Q^{[W,S]}) = \mathbb{P}(\mathbf{W}). \tag{2.10}$$

Similarly to show (W, S)-privacy it is sufficient to show that the query sent by the user is

independent of the side information index set an demand index set jointly and satisfies

$$\mathbb{P}(\mathbf{W}, \mathbf{S}|Q^{[W,S]}) = \mathbb{P}(\mathbf{W}, \mathbf{S}). \tag{2.11}$$

A PIR-SI scheme upholding W-privacy will be referred to as a W-PIR-SI scheme and a PIR-SI scheme upholding (W, S)-privacy will be referred to as a (W, S)-PIR-SI scheme.

The *rate* of a PIR-SI team is defined as

$$R = \frac{H(X_W)}{H(A^{[W,S]})} = \frac{t}{H(A^{[W,S]})}. \tag{2.12}$$

The *capacity* of the PIR-SI scheme is defined as the supremum of the rate over all PIR-SI schemes. The capacity of W-PIR-SI schemes will be denoted as C_W and the capacity of (W, S)-PIR-SI schemes will be denoted as $C_{(W,S)}$. The objective of this work is to find C_W and $C_{W,S}$ when considering different settings by finding upper bounds for the rate and designing schemes that can achieve the rate upper bound.

3. PIR-SI - SINGLE MESSAGE RESULTS*

The first scenario that is considered is the case that the user only demands one message from the database, $D = 1$. In this case Equation 2.3 becomes

$$\mathbb{P}(\mathbf{W} = W) = \begin{cases} \frac{1}{K} & \text{if } W \subseteq [K] \text{ and } |W| = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{3.1}$$

Letting $\mathcal{S} \triangleq \{(W, S) : |W| \subseteq [K], |W| = 1, S \subseteq [K], |S| = M, \text{ and } S \cap W = \emptyset\}$, Equation 2.5 becomes

$$\mathbb{P}(\mathbf{W} = W, \mathbf{S} = S) = \begin{cases} \frac{1}{(K-M)\binom{K}{M}} & \text{if } (W, S) \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}. \tag{3.2}$$

3.1 (W, S)-PIR-SI Single Server

First (W, S)-PIR-SI schemes are considered in the single server case. In the single server case it is shown that utilizing the side information introduced in this PIR-SI model, the user can increase the rate of retrieval, when compared to having no side information, by an additive factor. The capacity of (W, S)-PIR-SI is given below in a theorem and the rest of the section gives the upper bound of the rate and an achievable scheme that attains this rate.

Theorem 1. *For (W, S)-PIR-SI schemes when the size of the replicated database is K messages, there is one server $N = 1$, the user demands one message $D = 1$, and there are M messages in the users side information set, the capacity is*

$$C_{(W,S)} = (K - M)^{-1}. \tag{3.3}$$

*Parts of this section were reprinted with permission from: S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private Information Retrival with Side Information: The Single Server Case," *2017 55th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1099–1106, Oct 2017. ©2017 IEEE. Parts reprinted, with permission, from the authors and IEEE

3.1.1 Achievability

A scheme that achieves the rate of $(K-M)^{-1}$ is based off of a Maximum Distance Separable (MDS) code and is called the *MDS PIR Scheme*. An assumption that is needed for this scheme to be feasible is for the size of each message, t, to have the following relationship $2^t \geq K$, or $t \geq \log_2(K)$.

MDS PIR Scheme: For this scheme, each message will be viewed as an element in $\mathbb{F}_{2^t}$, $X_i \in \mathbb{F}_{2^t}$ and it is assumed all operations done in the scheme will be done over $\mathbb{F}_{2^t}$.

Step 1. The user will pick K distinct values from $\mathbb{F}_{2^t}$, call these elements $\{\alpha_i\}_{i=1}^{K}$. The user sends these values in order to the server.

Step 2. The user and server will then generate a $(K-M) \times K$ Vandermonde matrix from the elements selected of the form

$$V = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_K \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_K^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{K-M-1} & \alpha_2^{K-M-1} & \dots & \alpha_K^{K-M-1} \end{pmatrix}. \tag{3.4}$$

Step 3. The server responds by sending the rows of the matrix T, back to the user and the user decodes for their message;

$$T = V \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix}. \tag{3.5}$$

In the following Lemmas the feasibility and privacy of the scheme are shown.

Lemma 1. *In the (W, S)-PIR-SI problem when there is a single server ($N = 1$), the user has M messages as side information, and wants one message from the server ($D = 1$), the MDS PIR*

Scheme allows the user to decode for the message they want. Given the matrix T defined in the scheme the user is able to decode for any other message in $X_{[K]} \setminus X_S$. In particular the user can decode for X_W.

Proof. Without loss of generality assume the user has $X_S = X_{K-M+1}, X_{K-M+2}, \ldots, X_K$; if the user does not have these messages, reorder the messages at the server and permute the columns of V and T to reflect this, this will not change the problem. The user's side information can be expressed as the matrix expression below, where the coefficient matrix is a $M \times K$ matrix denoted by C

$$C \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix} = \begin{pmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_{K-M} \\ X_{K-M+1} \\ \vdots \\ X_K \end{pmatrix}.$$

When decoding for messages the equations that the user has to solve can be viewed as

$$T = \begin{pmatrix} V \\ C \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_{K-M} \\ X_{K-M+1} \\ \vdots \\ X_K \end{pmatrix}.$$

Looking at the coefficient matrix, with row operations the contributions of $X_{K-M}, \ldots, X_K$ can be used to give the resulting block matrix

$$\begin{pmatrix} \mathbf{V}' & \mathbf{0}_{(K-M)\times M} \\ \mathbf{0}_{M\times(K-M)} & \mathbf{I}_{M\times M} \end{pmatrix}.$$

Where $\mathbf{V}^{'}$ is a generalized Vandermonde matrix of dimension $(K-M)\times(K-M)$, $\mathbf{0}_{X\times Y}$ is the all zero matrix of dimension $X\times Y$, and $\mathbf{I}_{M\times M}$ is the identity matrix. The determinant of this matrix is

$$\det\begin{pmatrix}\mathbf{V}^{'} & \mathbf{0}_{(K-M)\times M}\\ \mathbf{0}_{M\times(K-M)} & \mathbf{I}_{M\times M}\end{pmatrix}=\det\left(\mathbf{V}^{'}\right).$$

Because $\mathbf{V}^{'}$ is a minor of the original Vandermonde matrix, its determinant is non-zero and so the matrix has full-rank. Because the matrix is full-rank, all messages can be decoded. □

Lemma 2. *In the (W,S)-PIR-SI problem when there is one server ($N=1$), the user has M messages as side information ,and wants one message from the server ($D=1$), the MDS PIR Scheme preserves (W,S)-privacy.*

Proof. The query sent to the server is independent of the user's demand and side information set. The server gains no information on (W,S). Because of this the scheme is (W,S)-private. □

Lemma 3. *The rate of the MDS PIR scheme is $(K-M)^{-1}$.*

Proof. Each message sent to the user corresponds to a row in V. Each row in V represents a linear combination of the messages $X_{[K]}$. Because the messages and coefficients are from the field $\mathbb{F}_{2^t}$ and the coefficients are from the same field, each transmission is independent and uniformly distributed over $\mathbb{F}_{2^t}$. Then we have that $H(A^{[W,S]})=(K-M)t$ and $R=(K-M)^{-1}$. □

3.1.2 Converse

To show an upper bound on the rate of any (W,S)-PIR-SI scheme a necessary condition for (W,S) privacy is described, a relation to a particular Index Coding problem is shown, a sub-problem of that Index Coding problem is considered, then a bound on that Index Coding problem is obtained.[1]

Lemma 4. *Let $\mathcal{S}\triangleq\{(W,S):|W|=1,W\subseteq[K],|S|=M,S\subseteq[K],$ and $S\cap W=\emptyset\}$ For the (W,S)-PIR-SI problem when there is one server ($N=1$), the user demands one message ($D=1$),*

[1] Background information on Index Coding is given in Appendix A

and has M messages as side information it must hold that any particular answer, $A^{[W_o,S_o]}$, sent to the user by the server must have the property that for all $(W,S) \in \mathcal{S}$, there exists a decoding function $D_{W,S}$ such that $D_{W,S}(A^{[W_o,S_o]}, X_S) = X_W$.

Proof. Let $A^{[W_o,S_o]}$ be a particular answer of a (W,S)-PIR-SI scheme. Suppose that with this answer there exists some $(W,S) \in \mathcal{S}$ such that there does not exist any $D_{W,S}$ where $D_{W,S}(A^{[W_o,S_o]}, X_S) = X_W$. Then the server knows the user could not have possibly contained that combination of W and S, i.e. $\mathbb{P}(W,S|Q^{[W,S]}) = 0$. It was assumed in the problem model that each (W,S) grouping had a non-zero a priori probability, $\mathbb{P}(W,S) \neq 0$ (Equation 2.2) and so (W,S)-privacy is not preserved. □

Lemma 5. *A (W,S)-PIR-SI answer $A^{[W,S]}$ that satisfies Lemma 4 is a solution to an Index Coding problem with the following properties:*

- *There are K messages, $X_1, ..., X_K$ located at the server*
- *There are $(K-M)\binom{K}{M}$ total clients.*
- *For each $i \in [K]$ and for each $S \in \{S : |S| = M, S \subseteq [K] \setminus \{i\}\}$, there is a client who wants X_i and has X_S as side information.*

Proof. A solution to an Index Coding problem is an answer from the server such that each user can decode for the message they demand using their answer and their side information.

In order for $A^{[W,S]}$ to be a solution to a (W,S)-PIR-SI problem, by Proposition 4 this condition holds and so it is a solution to the described Index Coding problem. □

Then the lower bound on the number of transmissions, and upper bound on the rate can be established by considering bounds on the solution to the given Index Coding problem.

Lemma 6. *For an Index Coding problem described in Lemma 5 the number of transmissions required to solve the Index Coding problem is at least $K - M$.*

Proof. Let J denote the Index Coding problem presented in Proposition 5. Choose some $S \in \{S : |S| = M, S \subseteq [K]\}$. Form another Index Coding problem $J^{'}$ such that the server has messages $X_1, X_2, \ldots, X_K$ and there are $(K - M)$ users, each with side information X_S and where client R_i demands X_i, where $i \in [K] \setminus S$.

A solution that solves J will also solve $J^{'}$ as the users in $J^{'}$ are a subset of the users in J. The number of transmissions required to solve J is then at least the number required to solve $J^{'}$.

Because each user in $J^{'}$ has the same side information, each user wants a message not in X_S, and the messages are independent, this means the solution to $J^{'}$ requires at least $K - M$ transmissions. □

Lemma 7. *For the (W, S)-PIR-SI problem the with $N = 1$, $D = 1$, and K messages, and side information size M, the capacity is at most $C_{(W,S)} \leq (K - M)^{-1}$*

Proof. Lemmas 4 and 5 together with Lemma 6 implies that the length of the an answer $A^{[W,S]}$ is at least $(K - M)t$ for any W and S. It follows then that and $C_{(W,S)} \leq (K - M)^{-1}$. □

3.2 W-Privacy Single Server

In this section it is shown that enforcing the weaker W-privacy can increase the capacity of PIR-SI by a multiplicative factor. This section gives the upper bound for the rate in the W-PIR-SI problem and also gives a partitioning based scheme that can achieve this upper bound.

Theorem 2. *For a W-PIR-SI schemes when the size of the replicated database is K messages, there is one server ($N = 1$), the user demands one message $D = 1$, and there are M messages in the user's side information set, the capacity is*

$$C_W = \left\lceil \frac{K}{M+1} \right\rceil^{-1}. \tag{3.6}$$

3.2.1 Achievability

To achieve a rate of $\lceil \frac{K}{M+1} \rceil$ a scheme referred to as the *Partition and Code Scheme* is described and the privacy and decodability of the scheme is shown.

Partition and Code Scheme: Given $N = 1$, and K and M denote $g \triangleq \lceil \frac{K}{M+1} \rceil$. The scheme then is

Step 1. The user creates a partition of $[K]$ into g sets in the following way

Case 1. $(M+1)|K$: In this case one set of the partition is $P_1 \triangleq W \cup S$, the remaining sets $\{P_i\}_{i=2}^{g}$ of size $M+1$ are randomly chosen from $[K] \setminus P_1$.

Case 2. $(M+1) \nmid K$: In this case let $P_1, P_2, ..., P_g$ be a collection of empty sets. $P_1, \ldots, P_{g-1}$ are to be of size $M+1$ and P_g is to be of size $K-(g-1)(M+1)$. The user begins by assigning probabilities to the sets according to their size; $P_1, \ldots, P_{g-1}$ are assigned a probability of $\frac{M+1}{K}$ and the set P_g is assigned a probability of $\frac{K-(g-1)(M+1)}{K}$. The user then chooses one of the sets $\{P_i\}_{i=1}^{g}$ randomly according to their assigned probabilities.

Let P denote the set chosen by the user. If $P \in \{P_1, \ldots, P_{g-1}\}$ then the user does as in the case where $(M+1)|K$ and sets $P = W \cup S$ and randomly fills the other sets with elements from $[K] \setminus P$. If the set P is chosen to be P_g then the user puts W in the set and randomly chooses elements from S to fill the set P_g. In this case, not all of the elements in S are put into P. The user then fills the remaining sets randomly from $[K] \setminus P$.

Step 2. The user then sends $\{P_1, \ldots, P_g\}$ to the server in a random order.

Step 3. The server sends back the answer $A^{[W,S]}$ which is a set of g inner products; $A^{[W,S]} \triangleq \{A_{P_1}, A_{P_2}, \ldots, A_{P_g}\}$ where $A_{P_i} = \sum_{j \in P_i} X_j$.

Step 4. The user decodes for X_W by subtracting off the contributions of their side information X_S from the appropriate $A_P \in A^{[W,S]}$.

Example $(\mathbf{M+1})|\mathbf{K}$: Suppose $K = 6$, $M = 2$, $W = \{5\}$, and $S = \{1, 2\}$. In the Partition and Code scheme the user would create the sets $P_1 = W \cup S = \{1, 2, 5\}$ and $P_2 = \{3, 4, 6\}$. The user then sends these sets to the server and receives $X_1 + X_2 + X_5$ and $X_3 + X_4 + X_6$ back from the server. The user decodes for X_5 by subtracting $X_1 + X_2$ from $X_1 + X_2 + X_5$.

From the server's perspective it knows the user's demand is in either $\{1, 2, 5\}$ or $\{3, 4, 6\}$, each with probability $\frac{1}{2}$. Given that the user's demand is in a particular set, the probability that it is one particular element in that set is $\frac{1}{3}$, giving $\mathbb{P}(\mathbf{W} = W | Q^{[W,S]}) = \frac{1}{6} = \mathbb{P}(\mathbf{W} = W)$.

Example $(\mathbf{M}+\mathbf{1}) \nmid \mathbf{K}$ **:** Suppose $K = 8$, $M = 2$, $W = \{5\}$, and $S = \{1, 2\}$. In the Partition and Code Scheme the user begins by labeling three empty sets as P_1, P_2, and P_3. The user assigns the probability $\frac{3}{8}$ to both P_1 and P_2 and the probability $\frac{2}{8}$ to P_3. The user then chooses one of these sets according to the probability distribution.

Suppose the user chooses the set P_3. The user then has to choose 1 element from S uniformly at random to place in P_3 along with 5; say the user chooses 1. Then the user sets $P_3 = \{1, 5\}$ and chooses the other two sets at random; say $P_1 = \{2, 8, 6\}$ and $P_2 = \{3, 4, 7\}$. The user sends these sets to the server and the server sends back the transmissions $X_2 + X_8 + X_6$, $X_3 + X_4 + X_7$, and $X_1 + X_5$. The user decodes by subtracting X_1 from $X_1 + X_5$ to decode for X_5.

From the server's perspective it knows that the user's demanded index is in either P_1, P_2, or P_3, and it knows that $\mathbb{P}(\mathbf{W} \in P_1 | Q^{[W,S]}) = \mathbb{P}(\mathbf{W} \in P_2 | Q^{[W,S]}) = \frac{3}{8}$ and $\mathbb{P}(\mathbf{W} \in P_3 | Q^{[W,S]}) = \frac{2}{8}$. Then the probability that $\mathbf{W}$ is a particular element in P_1 given that $\mathbf{W} \in P_1$ is $\frac{1}{3}$, and similarly for elements in P_2. The probability that $\mathbf{W}$ is a particular element in P_3 given that $\mathbf{W} \in P_3$ is $\frac{1}{2}$. Any case yields $\mathbb{P}(\mathbf{W} = W | Q^{[W,S]}) = \frac{1}{8} = \mathbb{P}(\mathbf{W} = W)$.

From the construction of the Partition and Code Scheme the user will always be able to decode for X_W by utilizing the transmission from the server that includes X_W.

To show the W-privacy of the scheme, the a posteriori probability after seeing a query is computed similarly to in the examples.

Lemma 8. *In the W-PIR-SI problem when there is one server ($N = 1$), the user has M messages as side information, and wants $D = 1$ message from the server, the Partition and Code scheme preserves W-privacy.*

Proof. To show W-privacy, the probability of $\mathbb{P}(\mathbf{W} | Q^{[W,S]})$ is computed to show that $Q^{[W,S]}$ is independent of $\mathbf{W}$. To compute this probability consider two cases.

Case 1. $\mathbf{W}$ is in one of the sets in $\{P_1, \ldots, P_{g-1}\}$: In this case for $i \in [g-1]$,

$$\mathbb{P}(\mathbf{W} \in P_i | Q^{[W,S]}) = \frac{M+1}{K}$$

and

$$\mathbb{P}(\mathbf{W} = W | \mathbf{W} \in P_i, Q^{[W,S]}) = \begin{cases} \frac{M+1}{K} & \text{if } W \in P_i \\ 0 & \text{if } W \notin P_i \end{cases}.$$

Case 2. $\mathbf{W}$ is in the set P_g. In this case,

$$\mathbb{P}(\mathbf{W} \in P_g | Q^{[W,S]}) = \frac{K - (g-1)(M+1)}{K}$$

and

$$\mathbb{P}(\mathbf{W} = W | \mathbf{W} \in P_g, Q^{[W,S]}) = \begin{cases} \frac{1}{K-(g-1)(M+1)} & \text{if } W \in P_g \\ 0 & \text{if } W \notin P_g \end{cases}.$$

Then to compute $\mathbb{P}(\mathbf{W} = W | Q^{[W,S]})$

$$\begin{aligned} \mathbb{P}(\mathbf{W} = W | Q^{[W,S]}) = & \\ & = \sum_{i=1}^{g} \mathbb{P}(\mathbf{W} = W, \mathbf{W} \in P_i | Q^{[W,S]}) \\ & = \sum_{i=1}^{g} \mathbb{P}(\mathbf{W} \in P_i | Q^{[W,S]}) \mathbb{P}(\mathbf{W} = W | \mathbf{W} \in P_i, Q^{[W,S]}) \\ & = \frac{1}{K}. \end{aligned}$$

□

Lemma 9. *The rate of the Partition and Code Scheme is* $\left\lceil \frac{K}{M+1} \right\rceil^{-1}$.

Proof. Because each transmission from the server is a linear combination of independent and uniformly distributed messages, each transmission $A_{P_i} \in A^{[W,S]}$ has entropy $H(A_{P_i}) = t$. Then $H(A^{[W,S]}) = \sum_{i=1}^{g} H(A_{P_i}) = gt$. The rate of the partition and code scheme is then

$$R = \frac{t}{gt} = \frac{1}{g} = \left\lceil \frac{K}{M+1} \right\rceil^{-1}.$$

□

3.2.2 Converse

To give an upper bound on the rate of W-PIR-SI schemes when $D = 1$ and $N = 1$ a necessary condition for W-privacy is introduced, a relation to a family of Index Coding problems is shown, and a bound of any Index Coding solution for a problem in the family is used to bound the rate of W-PIR-SI schemes.

Lemma 10. *Let $A^{[W,S]}$ be an answer from the server in a W-PIR-SI scheme, then the following is a necessary condition for $A^{[W,S]}$. For each message X_i, $i \in [K]$, there exists a subset of messages X_{S_i} with $S_i \subseteq [K] \backslash \{i\}$ and $|S_i| = M$ and a decoding function D_i satisfying $D_i(A^{[W,S]}, X_{S_i}) = X_i$.*

Proof. Suppose that there exists a W-PIR-SI solution $A^{[W,S]}$ such that there exists some X_i for which there does not exist any subset $S_i \subseteq [K] \backslash \{i\}$, $|S_i| = M$ and D_i such that $D_i(A^{[W,S]}, X_{S_i}) = X_i$. This means that there is no way the user can decode for X_i and further the server would know that $W \neq i$. Because of the assumption given in Equation 2.2 this leaks information and $A^{[W,S]}$ cannot be a W-PIR-SI solution. □

Lemma 11. *A W-PIR-SI answer $A^{[W,S]}$ that satisfies the necessary condition in Lemma 10 is a solution to an Index Coding problem with the following properties:*

- *There are K messages, $X_1, \ldots, X_K$ located at the server*
- *There are K total clients*
- *For each $i \in [K]$ there is a client that wants to decode X_i*
- *Each client i has some M messages as side information; $X_{S_i} \subseteq X_{[K]} \setminus X_i$*

Proof. Because $A^{[W,S]}$ satisfies the necessary condition in Proposition 10, each X_i has some side information X_{S_i} such that X_i can be decoded with $A^{[W,S]}$ and X_{S_i}. An Index Coding problem can then be constructed in the following way:

- There are K messages at the server, $X_1, X_2, \ldots, X_K$.

- There are K clients, $\{R_1, R_2, \ldots, R_K\}$, where R_i wants message X_i
- Each client R_i has side information X_{S_i}.

By the decodability condition of $A^{[W,S]}$ in Lemma 10 each client can decode for their wanted message ($D_i(\cdot)$ exists for each i) and is a solution to an the Index Coding problem described in this proposition. □

Lemma 11 shows that an answer to a W-PIR-SI problem must also be a solution to an Index Coding Problem instance with a corresponding side information graph G where each vertex of the graph has out degree M. The broadcast rate of an Index Coding problem can be lower bounded by bounding the Maximum Acyclic Induced Subgraph (MAIS) of G, denoted by MAIS(G). [2]

Lemma 12. *Let G be a directed graph with K verticies such that each vertex has out-degree M. MAIS(G) is lower bounded by $\lceil \frac{K}{M+1} \rceil$.*

Proof. Given a $G = (V, E)$ such that $|V| = K$ and each vertex has out-degree M, an acyclic induced subgraph of G, given by the set of verticies $Z \subset V$ of size $|Z| = \lceil \frac{K}{M+1} \rceil$ can be found with the following procedure.

Step 1. Set $Z = \emptyset$ and a candidate set of verticies to be $V' = V$

Step 2. Add an arbitrary vertex $i \in V'$ into Z, i.e. $Z = Z \cup \{i\}$

Step 3. Set $V' = V' \setminus (\mathcal{N}(i) \cup \{i\})$

Step 4. There are two cases

Case 1. If $V' \neq \emptyset$ repeat Steps 2-4.

Case 2. If $V' = \emptyset$, then terminate the procedure and return Z

To show the induced subgraph Z is acyclic first order the verticies of Z in the order they were added to Z according to the procedure. For $i, j \in Z$ and $i \neq j$ an arc (i, j) can only be an arc in Z if j was added to Z before i due to the ordering. There can not be an arc from a lower order vertex to a higher ordered vertex, so the induced subgraph Z is acyclic.

[2] Background on Index Coding can be found in Appendix A

Furthermore, $|Z| \geq \lceil \frac{K}{M+1} \rceil$. Notice that at each removal step, at most $M+1$ verticies are removed from the candidate set of verticies V'; each iteration exhausts at most $M+1$ verticies from V and there is one vertex added to Z on each iteration. □

Lemma 13. *For the W-PIR-SI problem with a single server that has K messages, and side information size M, the capacity is at most $\lceil \frac{K}{M+1} \rceil^{-1}$.*

Proof. Propositions 10 and 11 with Lemma 12 imply that the length of an answer to a W-PIR-SI problem $A^{[W,S]}$ is at least $t \lceil \frac{K}{M+1} \rceil$ for any W and S and the capacity then is upper bounded by $C_W \leq \lceil \frac{K}{M+1} \rceil^{-1}$. □

3.3 W-Private Multi-Server Scheme

In the case of a single demanded message by the user $D = 1$ there is a W-PIR-SI scheme for multiple servers that can be used which leverages the partition and code scheme and the scheme presented in [2].[3] This multi server PIR-SI scheme is denoted the *Multi-Server W-PIR-SI Scheme* and is described below.

Multi-Server W-PIR-SI Scheme: The user has one message that is demanded $D = 1$ and has M messages and that there are N servers in the PIR-SI problem. Also assume that each message is $t = N^{\frac{K}{M+1}}$ bits in length.

Step 1. Form partitions $P_1, P_2, \ldots, P_g$ as prescribed in Step 1. of the Partition and Code Scheme.

Step 2. The user sends $\{P_1, P_2, \ldots, P_g\}$ to the servers in a random order. From the sets received from the user, the servers form super messages $\hat{X}_1, \ldots, \hat{X}_g$ where $\hat{X}_i = \sum_{j \in P_i} X_j$.

Step 3. The user utilizes the Sun-Jafar scheme [2] on the super-messages to obtain the super message containing the user's demanded message.

Lemma 14. *For $D = 1$, the multi-server W-PIR-SI scheme satisfies W-privacy.*

Proof. The Sun-Jafar scheme is a valid PIR scheme [2], which means that each server does not

[3] Background on the Sun-Jafar scheme can be found in Appendix B

know which super message the user demands after seeing the query sent to it; each super message is equally likely.

Then from the super messages, the servers only learn the partitions given by Step 1 of the Partition and Code scheme. These partitions are proven to satisfy W-privacy in Lemma 8, so this scheme is W-private. □

Lemma 15. *When $D = 1$, the Multi-Server W-PIR-SI scheme has a rate of*

$$R = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{\left\lceil \frac{K}{M+1} \right\rceil - 1}}\right)^{-1}.$$

Proof. Each of the super messages are a linear combination of the independent and identically distributed messages, because of this $H(\hat{X}_i) = t$. The rate of the Sun Jafar scheme on $\left\lceil \frac{K}{M+1} \right\rceil$ messages is $\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{\left\lceil \frac{K}{M+1} \right\rceil - 1}}\right)^{-1}$. □

Example of Multi Message Scheme Consider the case when $K = 6$, $N = 2$, $M = 2$, and the messages are each $t = 4$ bits in size. Suppose that $W = \{1\}$ and $S = \{3, 6\}$. Then the user would send the sets $P_1 = \{1, 3, 6\}$ and $P_2 = \{2, 4, 5\}$ to the server. The server would form the super messages $\hat{X}_1 = X_1 + X_3 + X_6$ and $\hat{X}_2 = X_2 + X_4 + X_5$. Denote the bits of $\hat{X}_1 = \hat{X}_{1,1}, \hat{X}_{1,2}, \hat{X}_{1,3}, \hat{X}_{1,4}$ and $\hat{X}_2 = \hat{X}_{2,1}, \hat{X}_{2,2}, \hat{X}_{2,3}, \hat{X}_{2,4}$. The user would generate a permutation σ on $\{1, 2, 3, 4\}$, say $\sigma = (1423)$. Then the user would query the server for the following table of queries.

Queries to Server 1	Queries to Server 2
$\hat{X}_{1,4}$ $\hat{X}_{2,4}$	$\hat{X}_{1,3}$ $\hat{X}_{2,3}$
$\hat{X}_{1,1} + X_{2,3}$	$\hat{X}_{1,2} + X_{2,4}$

Table 3.1: Multi-Message Scheme Example Table of Queries

The user can reconstruct $\hat{X}_1$ from the answers of can decode for X_1 by $\hat{X}_1 - X_3 - X_6$.

4. PIR-SI - MULTI-MESSAGE RESULTS

The next part of the PIR-SI problem that is explored is when the user demands more than one message from a single server (i.e., $D \geq 2$,$N = 1$). It is noted that the MDS Coding scheme described in Section 3.1.1 can be used when the user demands more than one message as the user can decode for any message in the database (Lemma 1). It is shown that when $M = 1$ this scheme is actually optimal in the multi-message setting for both W-privacy and (W, S)-privacy.

The partitioning scheme described in Section 3.2.1 is generalized to the case where $D = 2$ and the user has M messages as side information. The W-privacy of the scheme is shown, and its rate is compared to the MDS PIR-SI scheme. The optimality of the scheme remains open.

4.1 $M = 1$ Converse

The capacity for the single server, multi-message case when the user has one message as side information is formally stated below.

Theorem 3. *For the W-PIR-SI problem when $N = 1$, $D \geq 2$, and $M = 1$,*

$$C_{(W,S)} = C_W = (K - 1)^{-1}. \tag{4.1}$$

The MDS PIR-SI scheme achieves a rate of $(K - 1)^{-1}$ and achieves (W, S)-privacy and as a consequence achieves W-privacy. In the following it will be shown that C_W is upper bounded by $(K - 1)^{-1}$ (see Lemma 20), thus proving Theorem 3.

In the single server case where the user only has one message as side information, $N = 1, M = 1$, the required number of transmissions is at least $K - 1$ for $D \geq 2$. To show this the case when $D = 2, M = 1$ is considered, a necessary condition is noted, an index coding problem is constructed from the necessary condition, the bound on the MAIS of a graph constructed from the index coding problems is found, and this bound is related to a subproblem of the original index coding problems to give a bound on the rate.

First a necessary condition is stated for the W-PIR-SI problem when the user wants two messages and has one message.

Lemma 16. *Let $A^{[W,S]}$ be an solution from the server for a W-PIR-SI problem where $N = 1$, $D = 2$, and $M = 1$. A necessary condition on $A^{[W,S]}$ is for each pair of messages X_i, X_j, $i \neq j$, there exists a message $X_{S_{ij}}$, $S \in [K] \setminus \{i, j\}$ and a pair of decoding functions D_{ij} and D_{ji} such that $D_{ij}(A^{[W,S]}, X_{S_{ij}}) = X_i$ and $D_{ji}(A^{[W,S]}, X_{S_{ij}}) = X_j$.*

Proof. Let $A^{[W,S]}$ be a solution to a W-PIR-SI problem where $D = 2$, $N = 1$, and $M = 1$ such that there exists some pair of messages X_i, X_j where there does not exist any combination of side information and decoding functions that can decode for both X_i and X_j. Then the server would know that the user could not have demanded the combination of messages X_i and X_j. Because it is assumed a priori that all pairs are demanded with a non-zero probability (Equation 2.2) this solution cannot be private because it leaks information. □

Lemma 17. *A solution $A^{[W,S]}$ to a W-PIR-SI problem where $D = 2$, $M = 1$, and $N = 1$ is a solution to an index coding problem with the following properties:*

- *There are K messages $X_1, X_2, \ldots, X_K$ located at the server.*
- *There are $\binom{K}{2}$ total receivers.*
- *For each pair $(i, j) \in [K] \times [K]$, $i \neq j$, there is a receiver that wants to decode both X_i and X_j. The receiver that wants the pair of messages (i, j) will be denoted R_{ij}.*
- *Each receiver R_{ij} has one message of side information $X_{S_{ij}} \in [K] \setminus \{i, j\}$.*

Proof. Let $A^{[W,S]}$ be a solution to a W-PIR-SI problem with $D = 2$, $M = 1$, and $N = 1$. Because $A^{[W,S]}$ is such a solution, it satisfies the necessary condition given in Lemma 16.

Construct an Index Coding problem in the following way:

- There are K messages at the server, $X_1, X_2, \ldots, X_K$
- There are $\binom{K}{2}$ clients $\{R_{12}, R_{13}, \ldots, R_{1K}, R_{23}, \ldots, R_{(K-1)K}\}$

- Each client R_{ij} has a message of side information $X_{S_{ij}}$.

The existence of side information sets and decoding functions for each pair of messages R_{ij} in Lemma 16 means each receiver in this Index Coding problem can decode for their pair of messages and $A^{[W,S]}$ is a solution to the constructed index coding problem. □

For a particular index coding problem satisfying the properties of Lemma 17 construct a directed graph $G = (V, E)$ in the following way:

- $V = \{1_2, 1_3, \ldots, 1_K, 2_3, 2_4, \ldots, (K-1)_K\}$. The two vertices i_j and j_i correspond to receiver R_{ij}.
- Let $E_1 \triangleq \{(i_j, x_y) : i = x, \text{ and } j \neq y\}$ forming K cliques, and $E_2 \triangleq \{(i_j, x_y) :$ receiver R_{ij} or R_{ji} has X_x as side information$\}$, to give a structure similar to regular side information graphs. Then $E = E_1 \cup E_2$.

Lemma 18. *Let $G = (V, E)$ be a graph for an index coding problem given in Lemma 17, constructed by the above procedure. Then there is an acyclic induced subgraph Z of size $K - 1$.*

Proof. Without loss of generality, suppose that client R_{1_2} and R_{2_1} have side information X_3. Then the arcs $(1_2, 3_i)$ and $(2_1, 3_i)$, where $i \in [K] \setminus \{3\}$, are in E. An acyclic induced subgraph Z can be found with the following procedure.

Step 1. Let $V(Z)$ denote the vertices in the subgraph Z, and intialize $V(Z)$ to $V(Z) = \emptyset$. Let V' be the current candidate set of vertices that can be added to $V(Z)$, and initialize $V' = V$.

Step 2. Add 1_2 and 2_1 to $V(Z)$. $V(Z) = \{1_2, 2_1\}$. Let $I_Z \triangleq \{i \in [K] : i_j \in V(Z)\}$; i.e. $I_Z = \{1, 2\}$. Set the label $k = 3$. Set the counter $r = 0$.

Step 3. Examine the vertices k_i, where $i \in [K] \setminus \{k\}$. There are two cases.

Case 1. There is some vertex k_i that has outgoing arcs (k_i, x_y) where $x \notin I_Z$. In this case add the vertex k_i to $V(Z)$; $V(Z) = V(Z) \cup \{k_i\}$, $I_Z = I_Z \cup \{k\}$. Set $k = i$. Set $r = r + 1$. Repeat Step 3.

Case 2. All vertices k_i have outgoing arcs (k_i, x_y) where $x \in I_Z$. Add the vertices j_k, where $j \in [K] \backslash (I_Z \cup \{k\})$, and set $V(Z) = V(Z) \cup \{j_k : j \in [K] \backslash (I_Z \cup \{k\})\}$. Terminate the procedure.

First notice that the procedure will terminate. There are a finite number of cliques in the graph G, at each step at least one clique is added to I_Z so the procedure will terminate at the worst case when there is only one clique not included in I_Z.

To show that the induced subgraph Z is acyclic a proof by contradiction is given. By way of contradiction suppose that Z is cyclic. Then at some point in the procedure a vertex added in either case must have produced a cycle in Z. Suppose that k_i caused the cycle and was added in Case 1 of Step 3 in the procedure. Then k_i must have been added with an outgoing arc of the form (k_i, x_y) where $x_y \in V(Z)$. This is not allowed in the procedure as this vertex would not have been added to $V(Z)$, the vertex cannot have been added in Case 1. Say then j_i caused the cycle and was added in Case 2 of Step 3 in the procedure. There is no arc in Z of the form (x_y, j_i), as if there was j_i would have been added already to Z in case 1 of step 3 of the procedure in an earlier iteration. This is a contradiction and Z could not have been created by the procedure.

The size of Z is $K - 1$. Before running any iteration of Step 3 of the procedure, $|Z| = 2$. Then the size of Z upon reaching Case 2 of Step 3 of the algorithm is $|Z| = 2 + r$ where r is the number of times Step 3 case 1 is iterated. When reaching Case 2 of Step 3, there are $K - r - 3$ verticies that are added to Z, making $|Z| = 2 + r + (K - r - 3) = K - 1$. □

Lemma 19. *For an index coding problem $\mathcal{I}$ fitting Lemma 17 and a graph G constructed with the given procedure from $\mathcal{I}$, let Z be the acyclic graph found using the procedure given in the proof of Lemma 18. Z corresponds to an acyclic induced subgraph of an index coding problem $\mathcal{J}$ that is a sub-problem of $\mathcal{I}$. The MAIS(Z) lower bounds the broadcast rate of the index coding problem $\mathcal{I}$.*

Proof. Construct an index coding problem $\mathcal{J}$ with the following procedure:

- There are K messages at the server $X_1, X_2, \ldots, X_K$.
- There are K receivers $R_1, R_2, \ldots, R_K$.
- For each vertex $i_j \in V(Z)$ that has an outgoing arc, receiver R_i has side information $\mathcal{N}(i_j)$.

- For each vertex $i_j \in V(Z)$ without an outgoing arc, receiver R_i has the same side information as R_{ij} or R_{ji} in $\mathcal{I}$.

- For $i \in [K]$ such that $i_j \notin V(Z)$ for any $j \in [K]$, let receiver R_i have one message of side information from the set $S_i \triangleq \{s \in [K] : \exists R_{ij} \text{ or } R_{ji} \in \mathcal{I} \text{ such that } R_{ij} \text{ has side information } X_s\}$.

Clearly any solution to $\mathcal{I}$ will solve the index coding problem $\mathcal{J}$. Because of this $\mathcal{J}$ is a subproblem of $\mathcal{I}$ so the length of an index coding solution for $\mathcal{I}$ must be at least as long as the solution to $\mathcal{J}$. (For the definition of the length of an Index Coding Solution refer to Appendix A.)

Notice further that Z is an acyclic induced subgraph of $\mathcal{J}$, and so $\text{MAIS}(Z) \geq K-1$. Together this implies that the broadcast rate of $\mathcal{I} \leq (K-1)^{-1}$. □

Lemma 20. *The capacity of the W-PIR-SI problem when $N = 1$ and $M = 1$, is upper bounded by $(K-1)^{-1}$.*

Proof. First notice that the number of transmissions required to solve a W-PIR-SI problem with $D > 2$ must be at least as many transmissions as a W-PIR-SI problem with $D = 2$; the rate is upper bounded by the case when $D = 2$. Then Lemmas 16–19 imply that the upper bound on the capacity of the W-PIR-SI problem when $D = 2$ is $(K-1)^{-1}$. Therefore for $D \geq 2$, $C_W \leq (K-1)^{-1}$. □

4.2 $D = 2$ W-PIR-SI Scheme

To achieve W-privacy in the case where $D = 2$ the *Generalized Partition and Code Scheme* will be introduced and is described below.

Generalized Partition and Code Scheme: The Generalized Partition and Code Scheme is described for $N = 1$ and $D = 2$ and $M \geq D$. [1] The following assumptions are made for the scheme:

1. Assume that $2|M$.

[1] There is a way to generalize the scheme for $D > 2$ but is omitted as the proof of privacy includes a large amount of case checking as D grows.

2. Assume that $(2 + \frac{M}{2})|K$.

3. Assume that $M \geq 2$.

4. Assume that $2^t \geq (2 + \frac{M}{2})$.

Given $D = 2$, K, and M satisfying the above conditions, denote $\beta \triangleq 2 + \frac{M}{2}$ and $g \triangleq \frac{K}{\beta}$. The scheme is then given by the following steps.

Step 1. The user creates a partition of $[K]$ into g sets, $P_1, P_2, \ldots, P_g$, each to be of size β each are initially empty and are filled in the following way. Choose an element (i, j) from $[g] \times [g]$ uniformly at random. Place W_1 into the set P_i and W_2 into the set P_j. There are two cases.

Case 1. The demand indices of the user are placed in the same set; $i = j$. In this case fill remaining places in the set P_i with $\beta - 2$ random elements from S.

Case 2. The demand indices of the user are placed in different sets; $i \neq j$. In this case choose $\beta - 2$ elements from S and place them into set P_i, place the remaining elements of S into P_j.

Fill the remaining sets randomly with elements from $[K] \setminus (P_i \cup P_j)$.

Step 2. The user then sends each P_i, $i \in [g]$, to the server in a random order, and with the elements in P_i sent in a random order.

Step 3. For each set the server gets, the server sends back

$$\mathbf{V}_{2\times\beta}\mathbf{X}_{P_i}.$$

Where $\mathbf{V}_{2\times\beta}$ is a Vandermonde matrix with distinct parameters $\alpha_1, \alpha_2, \ldots, \alpha_\beta \in \mathbb{F}_{2^t}$, and $\mathbf{X}_{P_i}$ is a $\beta \times 1$ column vector of the messages indexed by the elements in P_i given in the order the user sent them.

Step 4. There are two cases here for the user to decode

Case 1. The demand indices were placed in the same group P_i. To decode for the two messages W_1, W_2 is equivalent to solving a system of equations given by a 2×2 Vandermonde

matrix, as the user has $\beta-2$ messages in P_i as side information, this guarantees the user can decode for their packets as the determinant of any 2×2 minor is non-zero.

Case 2. The demand indices were placed in different groups P_i and P_j. Notice that this is similar to Case 1, in that both P_i and P_j both have $\beta - 2$ messages in both P_i and P_j as side information, this guarantees the user can decode for two messages in both P_i and P_j, namely W_1 and W_2.

Lemma 21. *The Generalized Partition and Code Scheme when $N = 1$, $D = 2$, $2|M$, and $\beta|K$ is W-private.*

Proof. To show the W-privacy of the scheme, $\mathbb{P}(\mathbf{W} = \{W_1, W_2\}|Q^{[W,S]})$ will be computed. To compute this privacy the events T and T^C are defined. T is the event that $\mathbf{W_1}$ and $\mathbf{W_2}$ are in the same group. T^C is the event that $\mathbf{W_1}$ and $\mathbf{W_2}$ are in different groups.

First consider the probability $\mathbb{P}(T|Q^{[W,S]})$. Because the distribution of $\mathbf{W}$ is uniform over all pairs, this probability is the number of pairs of indices in the same group in the query over the number of all possible pairs.

$$\mathbb{P}(T|Q^{[W,S]}) = \frac{g\binom{\beta}{2}}{\binom{K}{2}},$$

and

$$\mathbb{P}(T^C|Q^{[W,S]}) = 1 - \mathbb{P}(T|Q^{[W,S]}) = 1 - \frac{g\binom{\beta}{2}}{\binom{K}{2}} = \frac{\binom{K}{2} - g\binom{\beta}{2}}{\binom{K}{2}}.$$

Then the probability $\mathbb{P}(\mathbf{W} = \{W_1, W_2\}|T, Q^{[W,S]})$ is computed. Notice this probability is the inverse of the number of ways to pick a pair of elements from the same group.

$$\mathbb{P}(\mathbf{W} = \{W_1, W_2\}|T, Q^{[W,S]}) = \begin{cases} \frac{1}{g\binom{\beta}{2}} & \text{if } W_1 \text{ and } W_2 \text{ are in the same } P_i \text{ for some } i \in [g] \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbb{P}(\mathbf{W} = \{W_1, W_2\} | T^C, Q^{[W,S]}) = \begin{cases} \frac{1}{\binom{K}{2} - g\binom{\beta}{2}} & \text{if } W_1 \in P_i \text{ and } W_2 \in P_j \text{ for some } i, j,\ i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Then for any $W = \{W_1, W_2\}$

$$\begin{aligned} \mathbb{P}(\mathbf{W} = \{W_1, W_2\} | Q^{[W,S]}) = & \\ & = \mathbb{P}(\mathbf{W} = \{W_1, W_2\}, T | Q^{[W,S]}) + \mathbb{P}(\mathbf{W} = \{W_1, W_2\}, T^C | Q^{[W,S]}) \\ & = \frac{1}{\binom{K}{2}}. \end{aligned}$$

□

Notice the rate of the Generalized Partition and Code Scheme when $D = 2$ is $\left(\frac{4K}{4+M}\right)^{-1}$. Recall that the rate of the MDS PIR-SI scheme is $(K - M)^{-1}$ and is a feasible scheme for this scenario. The Generalized Partition and Code Scheme does not out perform the MDS PIR-SI scheme for small values of K but performs better for larger values of K, e.g., when $K \geq 4 + M$.

5. OPEN PROBLEMS AND CONCLUSIONS

In the classical PIR problem where there is one user and one server with K messages stored on it, the user has no choice but to query and download everything in the database in order to download a message privately. When the user has some side information in the form of a subset of messages contained on the server, the user no longer has to download everything from the server to be private.

With the introduction of side information to the problem two different privacy measures are introduced, the weaker W-privacy which just protects information about the user's demanded packet, and the stronger (W, S)-privacy which protects information about the user's demanded index and side information set jointly. When there is one server and the user has M of the database's K messages and enforcing the stronger (W, S)-privacy the user can achieve an additive reduction in download costs, downloading $K - M$ messages from the server. When enforcing the weaker W-privacy the user gets a multiplicative gain in terms of download cost, only downloading $\lceil \frac{K}{M+1} \rceil$ messages from the server.

When the user demands two messages from the server the gains are similar to the one message case. When enforcing (W, S)-privacy there is an additive gain in download cost when compared to downloading everything from the server. When enforcing W-privacy there is a multiplicative gain in download cost. It is shown that when the user has one message as side information in the multi-message case, the user can use the MDS-PIR-SI scheme to obtain his desired messages to achieve the rate of $(K - M)^{-1}$ and is shown to be optimal. When the user demands two or more messages from the server and the user has at least two messages as side information, the user can utilize the Generalized Partition and Code Scheme given in Section 4.2 to get a multiplicative advantage in rate and download the messages they want privately.

5.1 Open Problems

In the case where a user wants a single message from the database, there are multiple servers, and the user has a subset of messages from the database as side information, [19] characterizes the capacity of (W, S)-privacy presenting a scheme in this case and showing that the scheme is optimal. A scheme for W-privacy in this scenario was presented in Section 3.3 but the optimality remains an open problem.

In the multi-message case the capacity of a W-private scheme was shown when the user only has one message from the database as side information. A scheme is given when the user demands two messages from the server and has at least two messages from the database as side information that beats the rate of the MDS-PIR-SI scheme. A privacy proof that can scale with D remains an open problem; the optimality of the scheme also remains open.

Different types of side information that can be considered such as when the user has coded messages from the database as side information instead of a subset of uncoded messages from the database. An example of such a model would be when the user has one random linear combination of M messages from the database as side information. In the case where the user demands one message and there is one server it can be shown that the user can download much less while preserving W-privacy when compared to the case where the user has one uncoded message as side information. The capacity and achievable schemes for this problem remain open.

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 41–50, Oct 1995.

[2] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, vol. 63, pp. 4075–4088, Jul 2017.

[3] R. Tandon, "The capacity of cache aided private information retrieval," *arXiv*, vol. abs/1706.07035, Jun 2017.

[4] Y. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *arXiv*, vol. abs/1709.01056, Sep 2017.

[5] K. Banawan and S. Ulukus, "Multi-message private information retrieval," *IEEE International Symposium on Information Theory (ISIT)*, pp. 1898–1902, Jun 2017.

[6] K. Banawan and S. Ulukus, "The capacity of private information retrieval from byzantine and colluding databases," *arXiv*, vol. abs/1706.01442, Jul 2017.

[7] E. Y. Yang, J. Xu, and K. H. Bennett, "Private information retrieval in the presence of malicious failures," *Proceedings 26th Annual International Computer Software and Applications*, pp. 805–810, Aug 2002.

[8] Y. Zhang and G. Ge, "Multi-file private information retrieval from MDS coded databases with colluding servers," *arXiv*, vol. 1705.03186, Oct 2017.

[9] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by freij-hollanti et al," *arXiv*, vol. 1701.07807, Jan 2017.

[10] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *arXiv*, vol. 1611.02062, Aug 2017.

[11] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *arXiv*, vol. 1605.00635, May 2016.

[12] Q. Wang and M. Skoglund, "Secure symmetric private information retrieval from colluding databases with adversaries," *arXiv*, vol. abs/1707.02152, Aug 2017.

[13] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," *Journal of Computer and System Sciences*, vol. 60, pp. 592–629, Jun 2000.

[14] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *arXiv*, vol. 1606.08828, Jul 2017.

[15] Q. Wang and M. Skoglund, "Symmetric private information retrieval for mds coded distributed storage," *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2017.

[16] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *arXiv*, vol. 1609.08138, Sep 2016.

[17] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *arXiv*, vol. 1602.01458, Feb 2016.

[18] T. Chan, S. Ho, and H. Yamamoto, "Private information retrieval for coded storage," *arXiv*, vol. 1410.5489, Oct 2014.

[19] Z. Chen, Z. Wang, and S. Jafar, "The capacity of private information retrieval with private side information," *arXiv*, vol. abs/1709.03022, Sep 2017.

[20] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol, "Index coding with side information," pp. 197–206, Oct 2006.

[21] N. Alon, E. Lubetzky, U. Stav, A. Weinstein, and A. Hassidim, "Broadcasting with side information," *49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 823–832, Oct 2008.

APPENDIX A

INDEX CODING BACKGROUND

The index coding problem is introduced in [20] and has the following set up. There is a single server with a string of bits $x \in \{0,1\}^n$ and there are n receivers $R_1, R_2, \ldots, R_n$ with each receiver R_i interested in decoding for bit x_i. Each client has knows some of the bits of the string $S_i = \{j \in [n] : R_i \text{ knows bit } x_j \text{ of } x\}$. The goal of the index coding problem is to have the server broadcast the minimum number of transmissions, given by functions of x in order for each receiver to decode for x_i with the transmission and their side information.

In [20] a side information graph G is introduced to capture the relationship between desired indices that a receiver wants, and what indices the receiver has as side information.

Definition 3. *A **Side Information Graph** G for an index coding problem with receivers $R_1, R_2, \ldots, R_n$ is a directed graph $G = (V, E)$ where the vertex set $V = \{1, 2, \ldots, n\}$, and the edge set $E = \{(i, j) : R_i$ has index j as side information$\}$.*

With the side information graph G, then the side information of receiver R_i is denoted by $x[\mathcal{N}(i)]$ where $\mathcal{N}(i)$ are the out-neighbors of vertex i in G. Then an index code is formally defined in [20] in the following way.

Definition 4. *An **index code** for G, denoted $\mathcal{C}$, is a set of codewords in $\{0,1\}^l$ together with:*

1. *An encoding function E which maps inputs in $\{0,1\}^n$ to codewords*
2. *A set of decoding functions $D_1, D_2, \ldots, D_n$ such that $D_i(E(x), x[\mathcal{N}(i)]) = x_i$ for all i*

*The **length** of an index code is defined to be l.*

The work of [20] also consider a class of index codes called δ-error randomized index codes, which allow for encoding and decoding functions to be stochastic functions and also allows up

to a δ decoding error rate at the receivers. Theorem 6 from [20] lower bounds the length of a δ-error randomized index code given a side information G and is stated below. Note that MAIS(G) denotes the size of the maximum acyclic induced subgraph of G and $H_2(x)$ is the binary entropy function.

Theorem 4 (Theorem 6 [20])**.** *The length of any δ-error randomized index code for G is at least MAIS$(G)(1 - H_2(\delta))$.*

This work is focused on the special case of these codes where $\delta = 0$ and the receivers always can decode for their desired packet. Theorem 6 [20] says then that the length of a 0-error index code is lower bounded by MAIS(G) as $H_2(0) = 0$.

The work of [21] generalizes the index coding problem by having the server contain a word x of n blocks $x_1, x_2, \ldots, x_n$ where each block is of size t bits, $x_i \in \{0,1\}^t$. In their scenario there are m receivers $R_1, R_2, \ldots, R_m$ each wanting a single distinct block from x; R_j wants block $x_{f(j)}$. The authors denote the minimum length of a code for blocks of t bits as β_t. Then side information graph defined in [20] can be similarly defined for the problem in [21] and given a side information graph G, $\beta_t(G)$ can be found.

The authors in [21] consider the asymptotic behavior of $\beta_t(G)$, $\beta(G) \triangleq \lim_{t\to\infty} \frac{\beta_t(G)}{t}$ which they show is the same as $\inf_t \frac{\beta_t(G)}{t}$. They then state that MAIS$(G) \leq \beta(G)$ and therefore MAIS$(G) \leq \beta_t(G)$ for any t. This shows that the size of the maximum acyclic induced subgraph of a side information graph G gives a lower bound on the rate of an index code for a side information graph G.

APPENDIX B

SUN-JAFAR PIR SCHEME BACKGROUND

This appendix describes the Sun-Jafar scheme utilized in the $D = 1$ multi-server W-PIR-SI scheme described in Section 3.3. The Sun-Jafar scheme is introduced in [2] and the scheme is given in Algorithm 1 in [2]. This appendix gives examples of the scheme, for the formal algorithm refer to [2].

In [2] the classical PIR problem is considered in which there is a database of K independent messages $X_1, X_2, \ldots, X_K$ of size t bits replicated across N non-colluding servers. Each message X_i has entropy $H(X_i) = t$. There is a user who wants message $\mathbf{W} \in [K]$ and wants to retrieve $X_{\mathbf{W}}$ from the server while keeping $\mathbf{W}$ secret from each server. For a particular realization of $\mathbf{W} = W$ the user sends queries to the servers and the servers send the user back answers. The query sent to server n is denoted Q_n^W and the answer sent back to the user from server n is denoted A_n^W. The answer from server n is determined from the query sent to server n and the messages; $H(A_n^W|Q_n^W, X_1, \ldots, X_K) = 0$. The privacy metric of classical PIR in this context is then $I(\mathbf{W}; Q_n^{\mathbf{W}}, A_n^{\mathbf{W}}, X_1, \ldots, X_K) = 0$ for all n. The rate is defined to be $R \triangleq \frac{t}{D}$ where D is the number of bits downloaded in total from all the servers. The capacity C of classical PIR schemes is defined to be the supremum of rates over all PIR schemes.

The authors in [2] show that the capacity of a classical PIR scheme is $C = (1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}})^{-1}$. The scheme that achieves this capacity is given in Algorithm 1 of [2], an example of the scheme is given below.

Sun-Jafar Example: Suppose that $K = 3$ and $N = 2$ and the messages are $t = N^K = 8$ bits in length. Furthermore suppose that the user demands message X_1, i.e. $\mathbf{W} = 1$. Then the scheme is run as follows:

Step 1. Denote the message bits as $X_1 = X_{1,1}, X_{1,2}, \ldots, X_{1,8}$, $X_2 = X_{2,1}, X_{2,2}, \ldots, X_{2,8}$, and $X_3 = X_{3,1}, X_{3,2}, \ldots, X_{3,8}$. The user will choose a permutation of $\{1, 2, \ldots, 8\}$ randomly, denote

this permutation as σ.

Step 2. The user then queries the server's following the queries given in the table below. The rows of the table are sent to each server in a random order.

Queries to Server 1	Queries to Server 2
$X_{1,\sigma(1)}$ $X_{2,\sigma(1)}$ $X_{3,\sigma(1)}$	$X_{1,\sigma(2)}$ $X_{2,\sigma(2)}$ $X_{3,\sigma(2)}$
$X_{1,\sigma(3)} + X_{2,\sigma(2)}$ $X_{1,\sigma(5)} + X_{3,\sigma(2)}$ $X_{2,\sigma(3)} + X_{3,\sigma(3)}$	$X_{1,\sigma(4)} + X_{2,\sigma(1)}$ $X_{1,\sigma(6)} + X_{3,\sigma(1)}$ $X_{2,\sigma(4)} + X_{3,\sigma(4)}$
$X_{1,\sigma(7)} + X_{2,\sigma(4)} + X_{3,\sigma(4)}$	$X_{1,\sigma(8)} + X_{2,\sigma(3)} + X_{3,\sigma(3)}$

Table B.1: Sun-Jafar Scheme Transmissions for $\mathbf{W} = 1$, $K = 3$, $N = 2$