SYMMETRIC SCHEMES FOR EFFICIENT RANGE AND ERROR-TOLERANT SEARCH ON ENCRYPTED DATA

A Thesis
Presented to
The Academic Faculty

by

Nathan L. Chenette

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combinatorics, and Optimization

Department of Mathematics
Georgia Institute of Technology
August 2012

SYMMETRIC SCHEMES FOR EFFICIENT RANGE AND ERROR-TOLERANT SEARCH ON ENCRYPTED DATA

Approved by:

Professor Alexandra Boldyreva,
Advisor
College of Computing
Georgia Institute of Technology

Professor Richard Lipton
College of Computing
Georgia Institute of Technology

Professor Chris Peikert
College of Computing
Georgia Institute of Technology

Professor Mustaque Ahamad
College of Computing
Georgia Institute of Technology

Professor Robin Thomas
Department of Mathematics
Georgia Institute of Technology

Date Approved: 15 June 2012

For my parents,

Jon and Jeannie,

to whom I can attribute

all of the wisdom that I will ever attain

ACKNOWLEDGEMENTS

I must first thank my advisor, Sasha Boldyreva, for her outstanding support of me through my doctoral studies. I cannot imagine a better mentor to help me through the hard work, tediousness, and excitement of several years of theoretical cryptographic research. I am especially appreciative of her constant patience with me, her amazing ability to understand my often-scattered ideas, and her wisdom in knowing the right thing to do at any given time. I would also like to especially thank Adam O'Neill, my academic "big brother" who has always embodied that role for me, providing knowledge, support, and encouragement over the years. Thanks also to early research partner Younho Lee, and to my doctoral committee for their help and advice.

Much gratitude goes to the leadership in the Algorithms, Combinatorics, and Optimization program at Georgia Tech, particularly Robin Thomas, without whom I would not be where I am. I have also been lucky to have many friends in the ACO program who have given me support, moral and otherwise: Carl Yerger, Luke Postle, Arash Asadi, and others. Special thanks to Noah and Amanda Streib, great friends who have always leant a compassionate ear and a helping hand.

I owe a lot to friends who have been an active part of my life, whether I see them every day, every week, or less than once a year. My family, from grandparents to siblings to in-laws to parents and beyond, is a group of inspiring, amazing people, with no exceptions, and I thrive off of the joy they bring me every day. Finally, thanks to Heather—life is wonderful when you share all parts of it (well, besides chemical engineering and cryptography) with your best friend.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

Large-scale data management systems rely more and more on cloud storage, where the need for efficient search capabilities clashes with the need for data confidentiality. Encryption and efficient accessibility are naturally at odds, as for instance strong encryption necessitates that ciphertexts reveal nothing about underlying data. Searchable encryption is an active field in cryptography studying encryption schemes that provide varying levels of efficiency, functionality, and security, and efficient searchable encryption focuses on schemes enabling sub-linear (in the size of the database) search time. I present the first cryptographic study of efficient searchable symmetric encryption schemes supporting two types of search queries, range queries and error-tolerant queries.

The natural solution to accommodate efficient range queries on ciphertexts is to use order-preserving encryption (OPE). I propose a security definition for OPE schemes, construct the first OPE scheme with provable security, and further analyze security by characterizing one-wayness of the scheme. Efficient error-tolerant queries are enabled by efficient fuzzy-searchable encryption (EFSE). For EFSE, I introduce relevant primitives, an optimal security definition and a (somewhat space-inefficient, but in a sense efficient as possible) scheme achieving it, and more efficient schemes that achieve a weaker, but practical, security notion.

In all cases, I introduce new appropriate security definitions, construct novel schemes, and prove those schemes secure under standard assumptions. The goal of this line of research is to provide constructions and provable security analysis that

should help practitioners decide whether OPE or FSE provides a suitable efficiency-security-functionality tradeoff for a given application.

CHAPTER I

INTRODUCTION

In this chapter, I introduce, motivate, and contextualize the cryptographic study of general Efficient Searchable Encryption as well as my focal topics, Order-Preserving Encryption (OPE) and Efficient Fuzzy-Searchable Encryption (EFSE). In Section 1.5, I briefly introduce the new results and contributions of the thesis. To avoid excessive discussion in the Introduction, I relegate further introductory material to the Overview sections that begin Chapters 3, 4, and 5.

1.1 The Setting

Today's broadband, mobile, data-driven world has seen explosive growth in cloud storage, i.e., remote storage accessed over a network. Cloud storage frees clients from the burdens of data management while guaranteeing efficient access to data, and has many advantages including lower costs, more flexibility, decentralization of resources, and division of labor. The major downside of cloud storage, though, is lack of security. Providing efficient access to huge quantities of data rules out the possibility of using strong encryption (which, by design, hides all information about data), so virtually all cloud storage solutions currently store and access data in the clear. As a result, apprehension lingers over the widespread use of cloud storage, especially as concerns over cybersecurity skyrocket—to quote a top White House official[1], "our nation's security and economic prosperity depend on the security, stability, and integrity of communications and information infrastructure that are largely privately-owned and globally-operated." We are left with a critical challenge: can cloud storage solutions

[1]John Brennan, Assistant to the President for Homeland Security and Counterterrorism, quote from Congressional Bill S.773 – Cybersecurity Act of 2009

provide security against untrusted parties without compromising functionality and efficiency?

Let us conceptualize our view of this problem. Consider the so-called Database-as-a-Service (DBaaS) model, where a *client* (user) submits queries to access and manipulate data held on a remote *server* (outsourced database). The server and the communication channel are assumed to be untrusted, so data stored on the server, as well as data used in query communication, should be encrypted. On the other hand, queries from the client must be efficiently answered by the server, and several canonical types of queries should be accepted.

From this standpoint, one way to achieve security in cloud storage is to develop encryption schemes that enable efficient search on encrypted data, that is, on the ciphertexts themselves. The study of such schemes is called *efficient searchable encryption (ESE)* and has been the focus of much recent research in the database and security communities.

1.2 Efficient Searchable Encryption for Cloud Storage

The goals of ESE are threefold: *efficiency* (of search on encrypted data), *functionality* (flexibility of search and data management), and *security*. Clearly, these goals are at odds with one another. For example, consider encrypting data with a secure encryption scheme that hides all information and is always randomized (i.e., same messages yield completely different ciphertexts.) Then it can be shown [44] that some queries must require looking at every entry of the database, because if not then (intuitively) the encrypted data leaks information, namely, that the ignored database entries cannot satisfy conditions of the query. And of course, looking through every entry of a database can be very inefficient. Because of such interference between goals, research in ESE focuses on providing various schemes admitting different efficiency-functionality-security tradeoffs so that practitioners can choose the scheme that best

suits their application's needs.

For cloud storage, as emphasized by ESE efforts in the database community [53, 2, 36, 24, 38, 41, 49, 37, 20, 64], the importance of efficiency and functionality outweighs that of security. In particular, as cloud databases are huge, to be "efficient" a query must run in time *sub-linear* in the size of the database. (Throughout this thesis, "efficient" in this context means satisfying this sub-linear condition.) Additionally, the client should not be expected to know all the data nor index it prior to outsourcing and should be able to add data to the storage on-the-fly—forcing certain conditions on functionality. Security comes only after these considerations. Thus, the basic approach of cloud-storage ESE is to provide security to the maximum extent possible *subject to* supporting efficient queries (of various types) on a dynamic database with minimal client-side computation.

This approach has received attention in the database community, with mixed results. While many database-community solutions meet efficiency and functionality goals, they usually fall short on the security side, as they tend to provide only ad-hoc, "intuitive" security analysis. From a cryptographic standpoint, the only way to provide security guarantees for a ESE scheme is to use techniques of *provable security*: first select or define an appropriate notion of security; then, provide a formal proof reducing security of a scheme to acceptable assumptions on fundamental primitives.

Thankfully, the cryptography community has recently made much progress in applying the provable-security framework to the study of schemes allowing search on encrypted data. In general, these efforts have focused on achieving ESE in a particular scenario, in order to fine-tune the balance between efficiency, functionality, and security. Scenarios include ESE in the symmetric and asymmetric settings, as well as ESE for specific query types, which I clarify before discussing relevant past results at length in Section 1.3.

Symmetric and asymmetric settings. ESE can be applied in the symmetric

(private-key) and asymmetric (public-key) settings. The asymmetric setting allows various parties to contribute to a database, but assures that only the designated user may successfully query the encrypted data using the secret key. In the symmetric setting, the same key is used to encrypt, query, and decrypt the data. Thus, the symmetric setting fits the DBaaS model, where only one party (the client) encrypts, queries, and decrypts information. I will thus concentrate on the symmetric setting, which has been less studied.

Query types. Almost all ESE schemes focus on supporting a limited number of query types. The most basic type of query is the *exact-match* query, also known as keyword search, in which the client queries a single item (or keyword) and the output is a list of records corresponding to each instance of the item. We are particularly interested in two other types of queries, defined intuitively as follows.

- In a *range* query, the client specifies two items and the output should be records of all items "between" (e.g., lexicographically) the two query items.
- In an *error-tolerant* or *fuzzy* query, the client specifies an item and the output should be records of all items "close to" (e.g., defined by a metric and threshold) the query item.

In all cases, the *access pattern* of a search query is the set of records containing "hits," that is, the records testing positive for the item/s. The *search pattern* of a client is, essentially, the symmetric 0-1 matrix indicating which search queries were performed for the same item.

1.3 Past Results

We now consider a progression of past results in *searchable symmetric encryption* (SSE), that is, search on symmetrically-encrypted data, moving toward schemes that enable more efficient search and thus qualify as ESE. A summary table is given in

Table 1, evaluating the security-efficiency-functionality tradeoff of each approach and indicating limitations on which we would like to improve.

Approach	**Security**	**Efficiency**	**Functionality**
Oblivious RAM [33]	Nothing leaked	*Impractical*	All query types
Fully homomorphic encryption [55, 29]	Nothing leaked	*Impractical*	All query types
Exact-match SSE [61, 30, 34, 21]	Search/access patterns leaked	*Linear+*	Exact-match
Exact-match ESE via static indexes [23, 58, 46]	Search/access patterns leaked	Sub-linear	Exact-match *Limited updates*
Efficiently-searchable authenticated enc. [3]	Equality leaked	Sub-linear	Exact-match
Prefix-preserving encryption [49, 8, 65]	*Vulnerable*	Sub-linear	Range; *Specialized implementation*
Order-preserving encryption (OPE) [2]	*Undefined/ unknown*	Sub-linear	Range; Transparent implementation
Fuzzy ESE via static indexes [47]	Similarity/access patterns leaked	Sub-linear	Error-tolerant *Limited updates*
Efficient fuzzy-searchable encryption [48]	*Undefined/ unknown*	Sub-linear	Error-tolerant

Table 1: Summary of relevant past results in searchable encryption, with evaluation of security, efficiency, and functionality for each. *Italics* indicate limitations where improvement is needed according to our goals for ESE. Under **Security**, evaluations indicate the "only" information that is leaked. Notice that leaking equality of messages is strictly weaker security than leaking search and/or access patterns.

The first few results, oblivious RAM and fully homomorphic encryption, provide the basis for studying ESE. These results provide SSE with excellent functionality and security, but poor efficiency.

Oblivious RAM. A powerful result by Goldreich and Ostrovsky [33] showed that any form of searchable encryption can be achieved in its full functionality. According to [33], all the functionality of a hypothetical server-side RAM can be encoded into directions for a (server-side) "oblivious RAM" whose behavior and running time on any query is consistent, so that same-type queries look identical from the perspective

of the potentially malicious server. The cost of this transformation is only poly-logarithmic overhead in all parameters, but the hidden constants are huge and it also requires a logarithmic number of rounds of interaction for each read and write. A newer paper by Pinkas and Reinman [54] improves the performance, but still, Oblivious RAM is much too inefficient for use in large-scale cloud storage.

Fully homomorphic encryption. Another exciting direction is the long-desired technique of fully homomorphic encryption [55], which allows for any computation that can be written as a circuit to be performed in the encrypted domain, thus allowing for a wide array of possible functionality on encrypted data. Gentry [29] recently constructed the first fully homomorphic encryption scheme, based on a lattice construction. However, the construction is impractical for many applications, as ciphertext size and computation time increase sharply as one increases the security level. Some small improvements have been made, including a conceptually simpler integer-based solution [62] and schemes with improved parameter size [60, 19], but current results still fall short of something practically useful.

These theoretical results are impressive, but impractical—it is clear that efficient solutions are needed. Thus, continuing research into searchable encryption has focused on finding more efficient solutions by weakening the privacy guarantees, such as revealing the access pattern and/or search pattern but nothing else. Also, to achieve improved efficiency, most research has focused on addressing particular types of queries, starting with the most basic query type, exact-match queries.

Searchable encryption for exact-match queries. Exact-match query support in symmetric encryption has been an active topic in the cryptography community. Several works [61, 30, 34, 21] provide strong security guarantees for symmetric schemes supporting faster exact-match queries, at the expense of revealing the access and/or search pattern. However, these schemes do not achieve our standard of

efficiency, as they require that the server scan the entire database for each query. Several works [23, 58, 46] do achieve sub-linear-time search on *static* databases with good security—usually leaking access and search pattern—by building "secure indexes" for data. However, these solutions incur significant memory cost, as they require an entry in the index for each keyword; and moreover, they require the database to be fully known in advance, allowing only limited updates.

Finally, schemes supporting sub-linear-time exact-match queries were developed by [3] in the symmetric-key setting (and [9, 10, 17] in the public-key setting) at the necessary cost of slightly weakened security. In particular, the schemes improve efficiency and functionality by leaking equality of underlying messages (i.e., anyone can recognize when two ciphertexts are encryptions of the same message). This solution fits the efficiency and functionality conditions for cloud storage, and still provides quite strong security. However, exact-match queries can be somewhat restrictive, and practitioners would like to be able to support more flexible queries, particularly range and error-tolerant queries.

Searchable encryption for range queries. Range queries have been less studied in the ESE context but have inspired some recent activity. [18, 59] studied range queries on encrypted data in the public-key setting—but while their schemes provably provide strong security, they are also not efficient according to our sub-linear-time standard. The work of [49] suggested enabling efficient range queries on encrypted data by using *prefix-preserving encryption* (PPE) [8, 65]. Unfortunately, as discussed in [49, 3], PPE schemes are subject to certain attacks in this context; particular queries can completely reveal some of the underlying plaintexts in the database. Moreover, PPE demands use of specialized data structures and query formats, which practitioners would prefer to avoid.

More useful, from the standpoint of the database community, would be a scheme that supports range queries naturally, as easily as if the data were unencrypted, even if

this means much weaker security. This point is implied in the 2004 paper by Agrawal et al. [2] in which they propose supporting range queries on ESE in the symmetric setting using so-called *order-preserving encryption* (OPE)[2].

ORDER-PRESERVING ENCRYPTION. OPE is deterministic encryption in which for any key K, the encryption function $\mathcal{E}nc(K, \cdot)$ is order preserving, that is, $\mathcal{E}nc(K, m_0) < \mathcal{E}nc(K, m_1)$ if and only if $m_0 < m_1$. If a database is encrypted by an OPE scheme, and kept sorted in ciphertext order, then range queries are naturally supported: a range query simply specifies the two encrypted ends of the desired range, and the server returns all ciphertexts between these values. Note that public-key OPE schemes give virtually no security because access to encryption allows an adversary to find the preimage of any ciphertext via a binary search over encryptions of chosen plaintexts. Thus, we only talk about OPE in the private-key model.

One of the most enticing qualities of OPE is the transparency of its implementation: that is, data management and query protocol/processing (for range or exact-match search) are no different for an OPE-encrypted database than for the corresponding unencrypted database. Thus, practitioners can directly implement OPE into systems that currently use no encryption, with little to no effort, and with no special expertise. Moreover, in for example the DBaaS setting, the server need not even be aware that data is encrypted via OPE, as the server's view and actions are equivalent to the that of the unencrypted scenario.

After [2] noticed the natural ability of OPE to support efficient range queries, OPE received much interest in the database community, being suggested for use in in-network aggregation on encrypted data in sensor networks [1] and as a tool for applying signal processing techniques to multimedia content protection [25]. Obviously,

[2]While [2] initiated the modern study of OPE, interestingly, OPE in fact has a long history in the form of *one-part codes*. These are lists of plaintexts and the corresponding ciphertexts, both arranged in alphabetical or numerical order so only a single copy is required for efficient encryption and decryption. One-part codes were used, for example, during World War I [4].

OPE would also be an extremely functional and efficient method for supporting range queries on encrypted data in the cloud storage setting.

However, while the seminal work [2] does provide an OPE construction, it is rather ad-hoc and the encryption algorithm relies on knowing all plaintexts in advance; furthermore, the authors do not provide a definition of security nor any formal security analysis. Though OPE was proposed in spite of its naturally low level of security, without a notion of security to start with we have no way of formally evaluating the "security" of an OPE scheme. The only option in this case is to intuitively speculate as to the advantage of using an OPE scheme over, say, nothing at all. Intuition can be useful, but is often dangerously misleading, and leaves one vulnerable to unforeseen attacks. Crucially, to understand its security with any confidence, we must formally analyze OPE from a provable security (cryptographic) standpoint.

ESE FOR ERROR-TOLERANT QUERIES. Finally, we consider one other query type. *(Efficient) fuzzy-searchable encryption* or (E)FSE refers to encryption schemes allowing (efficient/sub-linear) error-tolerant query search on encrypted data. Obvious applications of FSE include any setting where data inherently contains errors (e.g., with biometric data) or queries are allowed to be approximate in value (e.g., misspellings or different formats of keywords allowed). EFSE, in which queries must be handled in sub-linear time, comes into play in large encrypted database settings. For example, EFSE could be used to efficiently query a large secure criminal database with biometric data (fingerprint measurements, etc.) from a crime scene.

The EFSE primitive has been relatively unstudied, with the exception of two recent works [48, 47]. Very recently, [47] developed a construction for sub-linear error-tolerant search on encrypted data using static indexes. Unfortunately, like [23, 58, 46] for exact-match queries, this technique requires that the database be fixed in advance in order to build a specialized index with entries for each keyword. Thus, their scheme does not allow efficient data updates. Finally, like [2] for OPE, [48] constructs an

EFSE scheme but does not formally analyze its security (or propose an appropriate security notion)—and as we shall see, our research shows flaws in the scheme's security and space-efficiency. Thus, no provably secure EFSE scheme supporting efficient data updates exists.

1.4 Goal

The goal of this work is to provide provably secure solutions for supporting efficient range and error-tolerant search on (updatable) encrypted data, specifically through the natural primitives of order-preserving encryption (OPE) and efficient fuzzy-searchable encryption (EFSE). At the least, it aims to provide the framework for the study of these topics, by defining new security notions capturing appropriate levels of security, and constructing the first provably secure schemes.

1.5 Contributions

I now introduce the contributions made in this thesis, which is joint work (coauthors listed below) also appearing in published [15, 16] and submitted [14] literature. For each subject, I provide an abridged introduction here; more detailed overviews begin each respective chapter.

1.5.1 First cryptographic study of order-preserving encryption

This topic reflects results published at EUROCRYPT 2009 [15] with co-authors Alexandra Boldyreva, Adam O'Neill, and Younho Lee. For the detailed introduction, see Section 3.1.

This work initiates the cryptographic study of order-preserving encryption. We demonstrate that OPE schemes cannot possibly achieve the usual security standard for symmetric encryption, IND-CPA, and in order to derive any meaningful security statement for OPE, a new security definition is needed. We opt for an

"indistinguishability-based" security definition, in which an OPE scheme is considered secure if its behavior is computationally indistinguishable from that of an "ideal object," namely a random order-preserving function (OPF) on the same domain and range. We then construct a scheme that achieves this level of security, uncovering in the process a fundamental relationship between the set of OPFs on a given domain and range, and the hypergeometric distribution.

A major point left unanswered here is that the underlying "ideal object" (a random OPF) itself has not been studied cryptographically. From a practical standpoint, the proof that our construction is indistinguishable from a random OPF only guarantees that our scheme's security is on a level with a random OPF. Indeed, we showed that conventional notions of security (e.g. IND-CPA security) are necessarily broken by OPE; thus, there is an inherent "information leakage" associated with our OPE scheme that must be characterized.

1.5.2 One-wayness security analysis of the OPE ideal object

This topic reflects results published at CRYPTO 2011 [16] with co-authors Alexandra Boldyreva and Adam O'Neill. For the detailed introduction, see Section 4.1.

In this chapter, we investigate the security properties of the ideal object in the above security notion for OPE [15]. In particular, we analyze the "information leakage" of the ideal object, a random OPF, through various one-wayness definitions, which capture the ability of an adversary to invert encryptions of random plaintexts, or to find the distance between random plaintexts given their encryptions.

With some qualifications on plaintext and ciphertext space sizes, we find that a random OPF is secure under these notions when the adversary is allowed a constant-sized "guessing window," while it is insecure when the adversary is allowed a guessing window of size proportional to the square root of the message space size. These results together show that a random OPF can be good at hiding the **precise** location of a

plaintext (or distance between two plaintexts) but poor at hiding the **approximate** location of a plaintext (or distance between two plaintexts).

In addition, we show that our original OPE scheme can be generalized slightly to achieve improved security in a limited sense. Furthermore, we show how our one-wayness analysis can be applied when there are known plaintext-ciphertext pairs, namely by splitting the space into subspaces and applying the analysis to each.

1.5.3 First cryptographic study of efficient error-tolerant encryption

This topic represents results co-authored with Alexandra Boldyreva and currently under review [14]. For the detailed introduction, see Section 5.1.

This work initiates the formal cryptographic study of EFSE. We develop a theory of FSE, defining primitives *closeness domain* (a domain along with a concept of "closeness" on it), FSE on a closeness domain, and EFSE. We propose an appropriate security definition for EFSE, show that the [48] scheme is insecure under the definition, and unveil the first provably secure EFSE scheme. This scheme, like that of [48], is space-inefficient, but we show information-theoretically that the space-(in)efficiency is nevertheless optimal for FSE schemes on arbitrary closeness domains. We then seek more space-efficient schemes for specific domains. Unfortunately, the optimal security seems unattainable and thus we propose a practical level of security, called macrostructure security, that ensures (at an intuitive level) only "local" information is leaked about plaintexts. Finally, we discuss schemes achieving this security for several useful closeness domains.

1.5.4 Impact

Our results represent a significant step forward in the lineage of efficiently searchable encryption as discussed in Section 1.3. From a theoretical standpoint, our work represents the first cryptographic study of OPE and EFSE, which are the most efficient and straightforward symmetric schemes allowing (respectively) range and error-tolerant

search on encrypted data. Also, the novel primitives, security notions, and techniques we establish provide a basis on which these (and related) topics can proceed in the future. In the real world, practitioners who want to support range or error-tolerant search for, say, cloud storage on untrusted servers, can study our security analysis and determine whether the security vs. efficiency vs. functionality balance of OPE or EFSE constructions are acceptable for their applications.

In fact, our groundbreaking OPE results as published in [15, 16] have received significant attention from both the cryptography community and the database community as well as companies such as JP Morgan, Symantec, and SalesForce. Also, our OPE scheme was successfully implemented and test-driven by the large CryptDB project at MIT (`http://css.csail.mit.edu/cryptdb/`), and is freely available as part of the CryptDB package. We anticipate similar interest in our EFSE results once they are published, as error-tolerant search seems even more practicable (e.g., for searching biometric data) than range query search.

CHAPTER II

PRELIMINARIES

Topic-specific preliminaries will appear in each chapter, but here I cover primitives, definitions, and conventions common to all topics. For the most part, this chapter consists of ideas standard in mathematical or cryptographic literature.

NOTATION AND CONVENTIONS. For sets X and Y, if $f\colon X \to Y$ is a function, then we call X the domain, Y the range, and the set $\{f(x) \mid x \in X\}$ the image of the function. We refer to members of $\{0,1\}^*$ as strings. If x is a string then $|x|$ denotes its length in bits and if x, y are strings then $x\|y$ denotes an encoding from which x, y are uniquely recoverable. For $\ell \in \mathbb{N}$ we denote by 1^ℓ the string of ℓ "1" bits.

If M is a positive integer, then $[M]$ denotes the set $\{1, \ldots, M\}$. For simplicity, in many cases we will assume a domain/plaintext space $[M]$ and range/ciphertext space $[N]$, for $N \geq M \in \mathbb{N}$. In general, results for arbitrary spaces $\mathcal{D}$, $\mathcal{R}$ can be derived from those of $[|\mathcal{D}|]$, $[|\mathcal{R}|]$—though (particularly for constructions) there may be some technical challenges in this translation that are beyond the scope of my research.

For set S and $n \leq |S|$, let Cmb_n^S denote the set of n-element subsets of S. If S is a finite set then $x \xleftarrow{\$} S$ denotes that x is selected uniformly at random from S. For convenience, for any $k \in \mathbb{N}$ we write $x_1, x_2, \ldots, x_k \xleftarrow{\$} S$ as shorthand for the series of assignments $x_1 \xleftarrow{\$} S, x_2 \xleftarrow{\$} S, \ldots, x_k \xleftarrow{\$} S$. If A is a randomized algorithm and Coins is the set from where it draws its coins, then we write $a \xleftarrow{\$} A(x, y, \ldots)$ as shorthand for $R \xleftarrow{\$} \mathsf{Coins}\,;\ a \leftarrow A(x, y, \ldots; R)$, where the latter denotes that variable a obtains the result of running A on inputs $x, y, \ldots$ and coins R.

In some of the algorithm descriptions, for ease and clarity of analysis, we use abstract set notation. In a practical implementation, the sets can be implemented

by some specialized data structure, or by vectors/lists with a common predetermined order (e.g., numerical order.) If $\mathcal{E}nc$ is an encryption function with key K, $\mathbf{x} = (x_1, \ldots, x_\ell)$ is a vector, and $X = \{x_1, \ldots, x_\ell\}$ is a set, then $\mathcal{E}nc(K, \mathbf{x}) = (\mathcal{E}nc(K, x_1), \ldots, \mathcal{E}nc(K, x_\ell))$ and $\mathcal{E}nc(K, X) = \{\mathcal{E}nc(K, x_1), \ldots, \mathcal{E}nc(K, x_\ell)\}$. The same holds for decryption $\mathcal{D}ec$.

We denote the probability of event A by $\Pr[A]$. If A depends on a random variable X, we write $\Pr_{X \xleftarrow{\$} D}[A(X)]$ for the probability of A when X sampled randomly from distribution D. If B is another event, $\Pr[A \mid B]$ denotes the conditional probability of A given B, and $\Pr_{X \xleftarrow{\$} D}[A(X) \mid B]$ denotes the conditional probability of $A(X)$ given B, for random variable X sampled from distribution D. Often, the distribution being used is clear and we omit it, as in $\Pr_X[A(X)]$ (where $X \xleftarrow{\$} D$ is implied).

Let $\mathbb{E}[X]$ denote the expected value of X. Again, we use the notation $\mathbb{E}_{Y \xleftarrow{\$} D}[X(Y)]$ or $\mathbb{E}_Y[X(Y)]$ to indicate that the expected value is taken over the randomness in selecting related random variable Y from distribution D.

An adversary is an algorithm. By convention, the running time of an adversary includes that of its overlying experiment. All algorithms are assumed to be efficient, and all functions are assumed to be efficiently computable.

For security notions, we often require that any efficient adversary's advantage will be "small." This condition is intentionally left informal as in symmetric key cryptography, we usually use blockciphers, which have fixed parameters; thus, we cannot bound advantage in terms of a security parameter. Judging what constitutes "small" advantage depends on the application and security needs, and is left to the reader.

Symmetric Encryption. A *symmetric encryption scheme* $\mathcal{SE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ with associated *plaintext space* $\mathcal{D}$ and *ciphertext space* $\mathcal{R}$ consists of three algorithms.

- The randomized *key generation algorithm* $\mathcal{K}$ returns a secret key K.

- The (possibly randomized) *encryption algorithm* $\mathcal{E}nc$ takes a secret key K and a plaintext m to return a ciphertext c.
- The deterministic *decryption algorithm* $\mathcal{D}ec$ takes a secret key K and a ciphertext c to return a plaintext m or a special symbol $\perp$ indicating that the ciphertext was invalid.

We require the usual correctness condition, $\mathcal{D}ec(K, (\mathcal{E}nc(K, m)) = m$ for all K output by $\mathcal{K}$ and all $m \in \mathcal{D}$. Finally, we say that $\mathcal{SE}$ is *deterministic* if $\mathcal{E}nc$ is deterministic.

INDISTINGUISHABILITY UNDER CHOSEN-PLAINTEXT ATTACK. Let $\mathcal{LR}(\cdot, \cdot, b)$ denote the *left-or-right selector* function that on inputs m_0, m_1 returns m_b. For a symmetric encryption scheme $\mathcal{SE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$, adversary A, and $b \in \{0, 1\}$, consider the IND-CPA experiment in Figure 1, where it is required that each query (m_0, m_1) that A makes to its oracle satisfies $|m_0| = |m_1|$.

Experiment $\mathbf{Exp}^{\text{ind-cpa-}b}_{\mathcal{SE}}(A)$
$K \xleftarrow{\$} \mathcal{K}$
$b' \xleftarrow{\$} A^{\mathcal{E}nc(K, \mathcal{LR}(\cdot,\cdot,b))}$
Return b' .

Figure 1: The IND-CPA experiment.

For an adversary A, define its *IND-CPA advantage* against $\mathcal{SE}$ as

$$\mathbf{Adv}^{\text{ind-cpa}}_{\mathcal{SE}}(A) \;=\; \Pr\left[\mathbf{Exp}^{\text{ind-cpa-1}}_{\mathcal{SE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-cpa-0}}_{\mathcal{SE}}(A) = 1\right] .$$

We say that $\mathcal{SE}$ is *indistinguishable under chosen-plaintext attack* (IND-CPA-secure) if the IND-CPA advantage of any adversary against $\mathcal{SE}$ is small.

PSEUDORANDOM FUNCTIONS (PRFS). We say that $\mathcal{F} = (\mathcal{K}, F)$ is a *function family* on domain $\mathcal{D}$ and range $\mathcal{R}$ if $\mathcal{K}$ outputs random keys and for each key $K \xleftarrow{\$} \mathcal{K}$ the map $F(K, \cdot)$ is a function from $\mathcal{D}$ to $\mathcal{R}$. We refer to $F(K, \cdot)$ as an *instance* of $\mathcal{F}$. Let $\mathsf{Func}_{\mathcal{D},\mathcal{R}}$ denote the set of all functions from $\mathcal{D}$ to $\mathcal{R}$. For any adversary A, the

prf-advantage against function family $\mathcal{F} = (\mathcal{K}, F)$ is defined as

$$\mathbf{Adv}^{\mathrm{prf}}_{\mathcal{F}}(A) \;=\; \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[\, A^{F(K,\cdot)} = 1 \,\right] - \Pr_{f \xleftarrow{\$} \mathsf{Func}_{\mathcal{D},\mathcal{R}}}\left[\, A^{f(\cdot)} = 1 \,\right]$$

We say that $\mathcal{F}$ is a *pseudorandom function* (PRF) if for any efficient adversary A, $\mathbf{Adv}^{\mathrm{prf}}_{\mathcal{F}}(A)$ is small.

BLOCKCIPHERS. A *blockcipher* is a function family $\mathcal{E}nc : \{0,1\}^k \times \{0,1\}^n \to \{0,1\}^n$ where *key length* k and *input/output length* n are parameters, and it is required that for every $K \in \{0,1\}^k$, $\mathcal{E}nc(K,\cdot)$ is a permutation (bijection) from $\{0,1\}^n$ to itself. In this document, when we refer to a "blockcipher" we mean one that is (assumed to be) psuedorandom permutation secure under chosen-ciphertext attack (PRP-CCA), such as Advanced Encryption Standard (AES).

METRIC SPACES. $(\mathcal{D}, d)$ is a *metric space* if $\mathcal{D}$ is a set and d (the *metric*) is a real-valued function on $\mathcal{D} \times \mathcal{D}$ such that for all $x, y, z \in \mathcal{D}$,

$$d(x,y) \geq 0 \qquad\qquad d(x,y) = 0 \text{ iff } x = y$$

$$d(x,y) = d(y,x) \qquad\qquad d(x,z) \leq d(x,y) + d(y,z).$$

CHAPTER III

ORDER-PRESERVING ENCRYPTION AND PSEUDORANDOM ORDER-PRESERVING FUNCTIONS

As introduced in Section 1.3, order-preserving symmetric encryption (OPE) is a deterministic encryption scheme (a.k.a. cipher) whose encryption function preserves numerical ordering of the plaintexts. Modern study of OPE was initiated by [2], who suggested it for use in supporting efficient (sub-linear-time) handling of range queries on encrypted data.

One might wonder whether order-preserving encryption is *necessary* to allow efficient range search on encrypted data. It is not—for example, the work of [49] shows that prefix-preserving encryption (where messages with equal prefixes are encrypted to ciphertexts with equal prefixes) also allows efficient range search, and other solutions may be possible. However, assuming a server accurately supports range queries, note that range query responses inherently leak relative ordering of underlying plaintexts. And if we must leak this information at least query-by-query, why not leak it directly—that is, allow order to be efficiently computable from ciphertexts? Then, the database server can index data according to this leaked information, enabling much faster search. This is precisely what order-preserving encryption gives us.

Indeed, OPE not only allows efficient range queries, but allows indexing and query processing to be done exactly and as efficiently as for unencrypted data, since a query consists of just the encryptions of a and b and the server can locate the desired ciphertexts in logarithmic-time via standard tree-based data structures. Such transparency is appealing to practitioners, who can effortlessly implement OPE in applications that previously ran in the cleartext—changing nothing on the database server side, which

may be oblivious to whether OPE is used or not.

The seminal work [2] suggests the OPE primitive and provides a construction. However, the construction is rather ad-hoc and has certain limitations, particularly that its encryption algorithm must take as input all the plaintexts in the database. It is not always practical to assume that users know all these plaintexts in advance, so a stateless scheme whose encryption algorithm can process single plaintexts on the fly is preferable. Moreover, [2] does not define security nor provide any formal security analysis. Thus, our following research [15] represents the first cryptographic study of OPE in the provable-security tradition. We first give an overview of the results.

3.1 Overview

Defining security of OPE. Our first goal is to devise a rigorous definition of security that OPE schemes should satisfy. Of course, such schemes cannot satisfy standard notions of security, such as indistinguishability against chosen-plaintext attack (IND-CPA), as they are not only deterministic, but also leak the order-relations among the plaintexts. (In particular, an adversary against an OPE scheme that queries two pairs with opposite order can trivially break IND-CPA, as the ciphertexts have the same order as their plaintexts.) So, although we cannot target a notion on the level of IND-CPA, we want to define the best possible security subject to this order-preserving constraint. (Such an approach was taken previously in the case of deterministic public-key encryption [9, 17, 10], on-line ciphers [8], and deterministic authenticated encryption [56].)

Weakening IND-CPA. One approach is to try to weaken the IND-CPA definition appropriately. Indeed, in the case of deterministic symmetric encryption this was done by [11], which formalizes a notion called *indistinguishability under distinct chosen-plaintext attack* or IND-DCPA. (The notion was subsequently applied to message authentication codes in [7].) Since deterministic encryption leaks equality of

plaintexts, IND-DCPA restricts the adversary in the IND-CPA experiment to make queries to its oracle that avoid the obvious attack exploiting equality-preservation. We generalize this to a notion we call *indistinguishability under ordered chosen-plaintext attack* or IND-OCPA, asking these sequences instead to satisfy the same *order relations*. (See Section 3.2.2.) Surprisingly, we go on to show that this plausible-looking definition is not useful for us, because it cannot be achieved by an OPE scheme unless the size of its ciphertext space is prohibitively large.

An alternative approach. Instead of trying to further restrict the adversary in the IND-OCPA definition, we turn to an approach along the lines of pseudorandom functions (PRFs) or permutations (PRPs), requiring that no adversary can distinguish between oracle access to the encryption algorithm of the scheme, and a corresponding "ideal" object. In our case the latter is a (uniformly) random order-preserving function on the same domain and range. Since order-preserving functions are injective, it also makes sense to aim for a stronger security notion that additionally gives the adversary oracle access to the decryption algorithm or the inverse function, respectively. We call the resulting notion POPF-CCA for *pseudorandom order-preserving function under chosen-ciphertext attack*.

Towards a construction. After having settled on the POPF-CCA notion, we would naturally like to construct an OPE scheme meeting it. Essentially, the encryption algorithm of such a scheme should behave similarly to an algorithm that samples a random order-preserving function from a specified domain and range on-the-fly (dynamically as new queries are made). (Here we note a connection to implementing huge random objects [32] and lazy-sampling [13].) But it is not immediately clear how this can be done; blockciphers, our usual tool in the symmetric-key setting, do not seem helpful in preserving plaintext order. Our construction takes a different route, borrowing some tools from probability theory. We first uncover a relation between a random order-preserving function and the hypergeometric (HG) and negative

hypergeometric (NHG) probability distributions.

The connection to NHG. To gain some intuition, first observe that any order-preserving function f from $\{1, \ldots, M\}$ to $\{1, \ldots, N\}$ can be uniquely represented by a combination of M out of N ordered items (see Proposition 3.3.1). Now let us recall a probability distribution that deals with selections of such combinations. Imagine we have N balls in a bin, out of which M are black and $N - M$ are white. At each step, we draw a ball at random without replacement. Consider the random variable Y describing the total number of balls removed after we collect the x-th black ball. This random variable follows the so-called negative hypergeometric (NHG) distribution. Using our representation of an order-preserving function, it is not hard to show that $f(x)$ for a given point $x \in \{1, \ldots, M\}$ has a NHG distribution over a random choice of f. Assuming an efficient sampling algorithm for the NHG distribution, this suggests a rough idea for a scheme, but there are still many subtleties to take care of.

Handling multiple points. First, assigning multiple plaintexts to ciphertexts independently according to the NHG distribution cannot work, because the resulting encryption function is unlikely to even be order-preserving. One could try to fix this by keeping track of all previously encrypted plaintexts and their ciphertexts (in both the encryption and decryption algorithms) and adjusting the parameters of the NHG sampling algorithm appropriately for each new plaintext. But we want a stateless scheme, so it cannot keep track of such previous assignments.

Eliminating the state. As a first step towards eliminating the state, we show that by assigning ciphertexts to plaintexts in a more organized fashion, the state can actually consist of a static but exponentially long random tape. The idea is that, to encrypt plaintext x, the encryption algorithm performs a binary search down to x. That is, it first assigns $\mathcal{E}nc(K, M/2)$, then $\mathcal{E}nc(K, M/4)$ if $x < M/2$ and $\mathcal{E}nc(K, 3M/4)$ otherwise, and so on, until $\mathcal{E}nc(K, x)$ is assigned. Crucially, each

ciphertext assignment is made according to the output of the NHG sampling algorithm run on appropriate parameters and *coins from an associated portion of the random tape indexed by those parameters*. (The decryption algorithm can be defined similarly.) Now, it may not be clear that the resulting scheme induces a *random* order-preserving function from the plaintext to ciphertext space (does its distribution get skewed by the binary search?), but we prove (by induction on the size of the plaintext space) that this is indeed the case.

Of course, instead of making the long random tape the secret key K for our scheme, we can make it the key for a PRF and generate portions of the tape dynamically as needed. However, coming up with a practical PRF construction to use here requires some care. For efficiency it should be blockcipher-based. Since the size of parameters to the NHG sampling algorithm as well as the number of random coins it needs varies during the binary search, and also because such a construction seems useful in general, it should be both variable input-length (VIL) and variable output-length. Such a construction we call a *length-flexible* (LF)-PRF. We propose a generic construction of an LF-PRF from a VIL-PRF and a (keyless) VOL-PRG (pseudorandom generator). Efficient blockcipher-based VIL-PRFs are known, and we suggest a highly efficient blockcipher-based VOL-PRG that is apparently folklore. POPF-CCA-security of the resulting OPE scheme can then be easily proved assuming only standard security (pseudorandomness) of the underlying blockcipher.

Switching from NHG to HG. Finally, our scheme needs an efficient sampling algorithm for the NHG distribution. Unfortunately, the existence of such an algorithm seems open. It is known that NHG can be approximated by the negative binomial distribution [51], which in turn can be sampled efficiently [28, 26], and that the approximation improves as M and N grow. However, quantifying the quality of approximation for fixed parameters seems difficult.

Instead, we turn to a related probability distribution, namely the hypergeometric (HG) distribution, for which a very efficient exact (not approximated) sampling algorithm is known [42, 43]. In our balls-and-bin model with M black and $N - M$ white balls, the random variable X specifying the number of black balls in our sample as soon as y balls are picked follows the HG distribution. The scheme based on this distribution, which is the one described in the body of the chapter, is rather more involved, but nearly as efficient: instead of $O(\log M) \cdot T_{\text{NHGD}}$ running time it is $O(\log N) \cdot T_{\text{HGD}}$ (where $T_{\text{NHGD}}, T_{\text{HGD}}$ are the running times of the sampling algorithms for the respective distributions), but we show that it is $O(\log M) \cdot T_{\text{HGD}}$ on average.

We note that the hypergeometric distribution was also used in [35] for sampling pseudorandom permutations and constructing blockciphers for short inputs. The authors of [35] were unaware of the efficient sampling algorithms for HG [42, 43] and provided their own realizations based on general sampling methods.

Further security. It is important to realize that the "ideal" object in our POPF-CCA definition (a random order-preserving function), and consequently our OPE construction meeting it, inherently leak some information about the underlying plaintexts. Characterizing this leakage is an important next step in the study of OPE and is covered in Chapter 4.

For now, the POPF-CCA definition captures in some sense a "best-possible" security notion for OPE. Note that it is usually the case that a security notion for a cryptographic object is met by a "random" one (which is sometimes built directly into the definition, as in the case of PRFs and PRPs). So it is natural to demand that an OPE scheme satisfy POPF-CCA.

Nevertheless, we caution that the further security analysis in Chapter 4, and possibly more analysis not yet performed, may be important for practitioners to evaluate how our POPF-CCA-secure OPE fits the security needs of an application.

On a more general primitive. To allow efficient range queries on encrypted data,

it is sufficient to have an order-preserving hash function family H (not necessarily invertible). The overall OPE scheme would then have secret key $(K_{\mathcal{E}nc}, K_H)$ where $K_{\mathcal{E}nc}$ is a key for a normal (randomized) encryption scheme and K_H is a key for H, and the encryption of x would be $\mathcal{E}nc(K_{\mathcal{E}nc}, x)\|H(K_H, x)$ (cf. efficiently searchable encryption (ESE) in [9]). Our security notion (in the CPA case) can also be applied to such H. In fact, there has been some work on hash functions that are order-preserving or have some related properties [50, 27, 39]. But none of these works are concerned with security in any sense. Since our OPE scheme is efficient and already invertible, we have not tried to build any secure order-preserving hash separately.

On the public-key setting. Finally, it is interesting to note that in a public-key setting one cannot expect OPE to provide any privacy at all. Indeed, given a ciphertext c computed under public key pk, anyone can decrypt c via a simple binary-search. In the symmetric-key setting a real-life adversary cannot encrypt messages itself, so such an attack is unlikely to be feasible.

3.2 Order-Preserving Encryption and Its Security

We begin by defining a primitive for deterministic encryption schemes that preserve order on their plaintext space.

3.2.1 Order-Preserving Encryption

For $A, B \subseteq \mathbb{N}$ with $|A| \leq |B|$, a function $f\colon A \to B$ is *order-preserving* (a.k.a. monotonically increasing) if for all $i, j \in A$, $f(i) > f(j)$ iff $i > j$. We say that deterministic encryption scheme $\mathcal{SE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ with plaintext and ciphertext spaces $\mathcal{D}, \mathcal{R}$ is an *order-preserving encryption* (OPE) scheme if $\mathcal{E}nc(K, \cdot)$ is an order-preserving function from $\mathcal{D}$ to $\mathcal{R}$ for all K output by $\mathcal{K}$ (with elements of $\mathcal{D}, \mathcal{R}$ interpreted as numbers, encoded as strings).

We now address the goal of establishing an appropriate notion of security for objects of this primitive.

3.2.2 Seeking an OPE security notion by weakening IND-CPA

OPE obviously cannot be IND-CPA-secure (see Chapter 2), as it leaks order of plaintexts. To see this, consider an adversary that submits two "crossing" left/right-queries $(m_0, m_1), (m'_0, m'_1)$, where (say) $m_0 < m'_0$ and $m_1 > m'_1$. Then corresponding OPE-encrypted ciphertexts c, c' will satisfy $c < c'$ in ind-cpa-0 and $c > c'$ in ind-cpa-1, so the adversary can easily achieve IND-CPA advantage 1. Thus, a natural question arises: can we weaken the IND-CPA-notion just enough so that it is achievable by an OPE scheme, but is still as strong as possible?

Past, related efforts have succeeded in this approach. Security of *deterministic symmetric encryption* was introduced in [11], as a notion they call *security under distinct chosen-plaintext attack (IND-DCPA)*. (It will not be important to consider chosen-ciphertext attacks now.) The idea is that because deterministic encryption leaks plaintext equality, the adversary A in the IND-CPA experiment is restricted to make only *distinct* queries on either side of its oracle (as otherwise there is a trivial attack). That is, supposing A makes queries $(m_0^1, m_1^1), \ldots, (m_0^q, m_1^q)$, they require that $m_b^1, \ldots m_b^q$ are all distinct for $b \in \{0, 1\}$. One could equivalently require that left queries and right queries have the same *equality pattern*, i.e., $m_0^i = m_0^j$ if and only if $m_1^i = m_1^j$ for all indices i, j.

Noting that any OPE scheme analogously leaks order relations of plaintexts, consider extending the above approach to take this into account. In particular, let us require the above queries made by A to have the same "order pattern."

IND-OCPA. Formally, let $\mathcal{LR}(\cdot, \cdot, b)$ denote the function that on inputs m_0, m_1 returns m_b. For a symmetric order-preserving encryption scheme $\mathcal{OPE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$, adversary A, and $b \in \{0, 1\}$, consider the following experiment:

Experiment $\mathbf{Exp}^{\text{ind-ocpa-}b}_{\mathcal{OPE}}(A)$

$K \xleftarrow{\$} \mathcal{K}$

$d \xleftarrow{\$} A^{\mathcal{E}nc(K,\mathcal{LR}(\cdot,\cdot,b))}$

Return d

We require that each query (m_0, m_1) that A makes to its oracle satisfies $|m_0| = |m_1|$, and also that the left/right-queries have the same *order pattern*, i.e. $m_0^i < m_0^j$ iff $m_1^i < m_1^j$ for all $1 \le i, j \le q$. For an adversary A, define its *indistinguishability under ordered chosen-plaintext attack (IND-OCPA) advantage* against $\mathcal{OPE}$ as

$$\mathbf{Adv}^{\text{ind-ocpa}}_{\mathcal{OPE}}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{\text{ind-ocpa-1}}_{\mathcal{OPE}}(A) = 1\,\right] - \Pr\left[\,\mathbf{Exp}^{\text{ind-ocpa-0}}_{\mathcal{OPE}}(A) = 1\,\right].$$

IND-OCPA IS NOT USEFUL. IND-OCPA security seems like a promising way to analyze security for OPE. Surprisingly, it turns out to be not useful for us. Below, we show that IND-OCPA is unachievable by a practical order-preserving encryption scheme, in that an OPE scheme cannot be IND-OCPA assuming its ciphertext space size can be bounded by an exponential function in the message space size. (This extends a result from our published paper [15].)

Theorem 3.2.1. *Let $\mathcal{OPE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ be an order-preserving encryption scheme on plaintext-space $[M]$ and ciphertext-space $[N]$, where $N < t^{\lfloor M/4 \rfloor}$ for some integer $t > 1$. There exists an IND-OCPA adversary A against $\mathcal{OPE}$ such that*

$$\mathbf{Adv}^{\text{ind-ocpa}}_{\mathcal{OPE}}(A) > \frac{1}{2t}.$$

Furthermore, A runs in time $O(\log N)$ and makes at most 3 oracle queries.

The proof of this result is in Appendix A.1. Notice then that if N is exponential in M, there exists an adversary with constant, nonzero IND-OCPA advantage. To have N super-exponential in M would be inconceivable, so for all intents and purposes, Theorem 3.2.1 shows IND-OCPA is unachievable for all practical OPE schemes.

In addition, the attack (like the weaker "big-jump attack" of our published paper [15]) implies that an OPE scheme inherently leaks more information about the plaintexts than just their order, namely some information about their relative distances. We return to this point in Chapter 4.

3.2.3 OPE security through pseudorandom order-preserving functions

Since OPE inherently leaks distance information about plaintexts, further weakening of IND-CPA does not seem very fruitful, as long as attacks can still sample far-apart versus close-together plaintexts. We instead assume a new starting point in the search for an OPE security notion: namely, security of pseudorandom permutations (PRPs) [31] or on-line PRPs [8], in which oracle access to the function in question should be indistinguishable from access to the corresponding "ideal" random object, e.g., a random permutation or a random on-line permutation. As order-preserving functions are injective, we consider the "strong" version of such a definition where an inverse oracle is also given.

Fix an order-preserving encryption scheme $\mathcal{SE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ with plaintext space $\mathcal{D}$ and ciphertext space $\mathcal{R}$, $|\mathcal{D}| \leq |\mathcal{R}|$. For an adversary A against $\mathcal{SE}$, define its *pseudorandom order-preserving function advantage under chosen-ciphertext attack (POPF-CCA-advantage)* against $\mathcal{SE}$ as

$$\mathbf{Adv}^{\text{popf-cca}}_{\mathcal{SE}}(A) = \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[\, A^{\mathcal{E}nc(K,\cdot),\mathcal{D}ec(K,\cdot)} = 1 \,\right] - \Pr_{g \xleftarrow{\$} \mathsf{OPF}_{\mathcal{D},\mathcal{R}}}\left[\, A^{g(\cdot),g^{-1}(\cdot)} = 1 \,\right],$$

where $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$ denotes the set of all order-preserving functions from $\mathcal{D}$ to $\mathcal{R}$. We say a scheme is *POPF-CCA-secure* if the POPF-CCA-advantage of any efficient adversary against the scheme is small.

Lazy sampling. To build a scheme that achieves POPF-CCA-security, we cannot simply select a random order-preserving function and sample it, because it is inefficient to even describe such an object. Rather, we seek a way to "lazy-sample" (a

term from [13]) a random order-preserving function and its inverse—that is, generate points of the function on-the-fly, as needed.[1]

As shown in [13], lazy-sampling of "exotic" functions with many constraints can be tricky. In the case of a random order-preserving function, it turns out that straightforward procedures—which assign a random point in the range to a queried domain point, subject to the obvious remaining constraints—do not work (that is, the resulting function is not uniformly distributed over the set of all such functions). So how can we lazy-sample such a function, if it is possible at all? We address this issue next.

3.3 Lazy-Sampling a Random Order-Preserving Function

In this section, we show how to lazy-sample a random order-preserving function and its inverse. This result may also be of independent interest, since the more general question of what functions can be lazy-sampled is interesting in its own right, and it may find other applications as well, e.g. to [57]. We first uncover a connection between a random order-preserving function and the hypergeometric (HG) probability distribution.

3.3.1 The hypergeometric connection

To gain some intuition we start with the following claim.

Proposition 3.3.1. *There is bijection between the set* $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$ *containing all order-preserving functions from a domain* $\mathcal{D}$ *of size* M *to a range* $\mathcal{R}$ *of size* $N \geq M$ *and the set of all possible combinations of* M *out of* N *ordered items.*

Proof. Without loss of generality, it is enough to prove the result for domain $[M]$ and range $[N]$. Imagine a graph with its x-axis marked with integers from 1 to

[1] For example, in the case of a random function from the set of *all* functions one can simply assign a random point from the range to each new point queried from the domain. In the case of a random permutation, the former can be chosen from the set of all previously unassigned points in the range, and lazy-sampling of its inverse can be done similarly. A lazy-sampling procedure for a random on-line PRP and its inverse via a tree-based characterization was given in [8].

M and its $y = f(x)$-axis marked with integers from 1 to N. Given a set S of M distinct integers from $[N]$, construct an order-preserving function from $[M]$ to $[N]$ by mapping each $i \in [M]$ to the ith smallest element in S. So, an M-out-of-N combination corresponds to a unique order-preserving function. On the other hand, consider an order-preserving function f from $[M]$ to $[N]$. The image of f defines a set of M distinct objects in $[N]$, so an order-preserving function corresponds to a unique M-out-of-N combination. □

Using the above combination-based characterization it is straightforward to justify the following equality, defined for $M, N \in \mathbb{N}$ and any $x \in [M-1], y \in [N]$:

$$\Pr_{f \xleftarrow{\$} \mathsf{OPF}_{[M],[N]}} [\, f(x) \leq y < f(x+1) \,] = \frac{\binom{y}{x} \cdot \binom{N-y}{M-x}}{\binom{N}{M}} \,. \tag{1}$$

Now let us recall a particular distribution dealing with an experiment of selecting from combinations of items.

Hypergeometric distribution. Consider the following balls-and-bins model. Assume we have N balls in a bin out of which M balls are black and $N - M$ balls are white. At each step we draw a ball at random, without replacement. Consider a random variable X that describes the number of black balls chosen after a *sample size* of y balls are picked. This random variable has a hypergeometric distribution, and the probability that $X = x$ for the parameters N, M, y is

$$P_{HGD}(x; N, M, y) = \frac{\binom{y}{x} \cdot \binom{N-y}{M-x}}{\binom{N}{M}} \,. \tag{2}$$

Intuitively, Equations 1 and 2 imply that we can construct a random order-preserving function f from $[M]$ to $[N]$ as an experiment involving N balls, M of which are black. Choosing balls randomly without replacement, if the y-th ball we pick is black then the least unmapped point in the domain is mapped to y under f. Of course, this experiment is too inefficient to be performed directly when any of the parameters are large. But we will use the hypergeometric distribution to design procedures that

efficiently and recursively lazy-sample a random order-preserving function and its inverse.

3.3.2 The LazySample algorithms

Here we give our algorithms **LazySample**, **LazySampleInv** that lazy-sample a random order-preserving function from domain $\mathcal{D}$ to range $\mathcal{R}$, $|\mathcal{D}| \leq |\mathcal{R}|$, and its inverse, respectively. The algorithms share and maintain joint state. We assume that both $\mathcal{D}$ and $\mathcal{R}$ are sets of consecutive integers.

Two subroutines. Our algorithms make use of two subroutines. The first, denoted HGD, takes inputs M, N, and $y \in \{0, 1, \ldots, N\}$ to return $x \in \{0, 1, \ldots, M\}$ such that for each $x^* \in \{0, 1, \ldots, M\}$ we have $x = x^*$ with probability $P_{HGD}(x; N, M, y)$ over the coins of HGD. (Efficient algorithms for this exist, and we discuss them in Section 3.3.5.) The second, denoted GetCoins, takes inputs 1^ℓ, $\mathcal{D}$, $\mathcal{R}$, and $b\|z$, where $b \in \{0, 1\}$ and $z \in \mathcal{R}$ if $b = 0$ and $z \in \mathcal{D}$ otherwise, to return $cc \in \{0, 1\}^\ell$. The purpose of GetCoins is to provide a consistent random coin for a particular call to a sampling algorithm such as HGD, where the parameters input to GetCoins match the parameters to be sent to the sampling algorithm.

The algorithms. To define our algorithms, let us denote by $w \xleftarrow{cc} S$ that w is assigned a value sampled uniformly at random from set S using coins cc of length ℓ_S, where ℓ_S denotes the number of coins needed to do so. Let $\ell_1 = \ell(M, N, y - r)$ denote the number of coins needed by HGD on inputs $M, N, y - r$. Our algorithms are given in Figure 2. Note that the arrays F and I, initially empty, are global and shared between the algorithms; also, for now, think of GetCoins as returning fresh random coins. We later implement it by using a PRF on the same parameters to eliminate the joint state.

Overview. To determine the image of input m, **LazySample** employs a strategy of mapping "range gaps" to "domain gaps" in a recursive, binary search manner. By

LazySample$(\mathcal{D}, \mathcal{R}, m)$

01 $M \leftarrow |\mathcal{D}|$; $N \leftarrow |\mathcal{R}|$
02 $d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$
03 $y \leftarrow r + \lceil N/2 \rceil$
04 If $|\mathcal{D}| = 1$ then
05 If $F[\mathcal{D}, \mathcal{R}, m]$ is undefined then
06 $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_{\mathcal{R}}}, \mathcal{D}, \mathcal{R}, 1 \| m)$
07 $F[\mathcal{D}, \mathcal{R}, m] \xleftarrow{cc} \mathcal{R}$
08 Return $F[\mathcal{D}, \mathcal{R}, m]$
09 If $I[\mathcal{D}, \mathcal{R}, y]$ is undefined then
10 $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_1}, \mathcal{D}, \mathcal{R}, 0 \| y)$
11 $I[\mathcal{D}, \mathcal{R}, y] \xleftarrow{\$} \text{HGD}(M, N, y - r; cc)$
12 $x \leftarrow d + I[\mathcal{D}, \mathcal{R}, y]$
13 If $m \leq x$ then
14 $\mathcal{D} \leftarrow \{d + 1, \ldots, x\}$
15 $\mathcal{R} \leftarrow \{r + 1, \ldots, y\}$
16 Else
17 $\mathcal{D} \leftarrow \{x + 1, \ldots, d + M\}$
18 $\mathcal{R} \leftarrow \{y + 1, \ldots, r + N\}$
19 Return **LazySample**$(\mathcal{D}, \mathcal{R}, m)$

LazySampleInv$(\mathcal{D}, \mathcal{R}, c)$

20 $M \leftarrow |\mathcal{D}|$; $N \leftarrow |\mathcal{R}|$
21 $d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$
22 $y \leftarrow r + \lceil N/2 \rceil$
23 If $|\mathcal{D}| = 1$ then $m \leftarrow \min(\mathcal{D})$
24 If $F[\mathcal{D}, \mathcal{R}, m]$ is undefined then
25 $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_{\mathcal{R}}}, \mathcal{D}, \mathcal{R}, 1 \| m)$
26 $F[\mathcal{D}, \mathcal{R}, m] \xleftarrow{cc} \mathcal{R}$
27 If $F[\mathcal{D}, \mathcal{R}, m] = c$ then return m
28 Else return $\perp$
29 If $I[\mathcal{D}, \mathcal{R}, y]$ is undefined then
30 $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_1}, \mathcal{D}, \mathcal{R}, 0 \| y)$
31 $I[\mathcal{D}, \mathcal{R}, y] \xleftarrow{\$} \text{HGD}(M, N, y - r; cc)$
32 $x \leftarrow d + I[\mathcal{D}, \mathcal{R}, y]$
33 If $c \leq y$ then
34 $\mathcal{D} \leftarrow \{d + 1, \ldots, x\}$
35 $\mathcal{R} \leftarrow \{r + 1, \ldots, y\}$
36 Else
37 $\mathcal{D} \leftarrow \{x + 1, \ldots, d + M\}$
38 $\mathcal{R} \leftarrow \{y + 1, \ldots, r + N\}$
39 Return **LazySampleInv**$(\mathcal{D}, \mathcal{R}, c)$

Figure 2: Algorithms **LazySample** and **LazySampleInv** for lazy-sampling a pseudorandom order-preserving function and its inverse by sampling the hypergeometric distribution.

"range gap" or "domain gap," we mean an imaginary barrier between two consecutive points in the range or domain, respectively. When run, the algorithm first maps the middle range gap y (the gap between the middle two range points) to a domain gap. To determine the mapping, on line 11 it sets, according to the hypergeometric distribution, how many points in $\mathcal{D}$ are mapped up to range point y and stores this value in array I. (In the future the array is referenced instead of choosing this value anew.) Thus we have that $f(x) \leq y < f(x + 1)$ (cf. (1)), where $x = d + I[\mathcal{D}, \mathcal{R}, y]$ as computed on line 12. So, we can view the range gap between y and $y + 1$ as having been mapped to the domain gap between x and $x + 1$.

If the input domain point m is below (resp. above) the domain gap, the algorithm recurses on line 19 on the lower (resp. upper) half of the range and the lower

(resp. upper) part of the domain, mapping further "middle" range gaps to domain gaps. This process continues until the gaps on either side of m have been mapped to by some range gaps. Finally, on line 07, the algorithm samples a range point uniformly at random from the "window" defined by the range gaps corresponding to m's neighboring domain gaps. This result is assigned to array F as the image of m under the lazy-sampled function.

3.3.3 Correctness

When GetCoins returns truly random coins, it should be clear that **LazySample** and **LazySampleInv** are consistent and sample an order-preserving function and its inverse respectively. But we need a stronger claim; namely, that our algorithms sample a (uniformly) *random* order-preserving function and its inverse. We show this by arguing that any (even computationally unbounded) adversary has no advantage in distinguishing oracle access to a random order-preserving function and its inverse from that to the algorithms **LazySample**, **LazySampleInv**. The following theorem states this claim.

Theorem 3.3.2. *Suppose* GetCoins *returns truly random coins on each new input. Then for any (even computationally unbounded) algorithm A we have*

$$\Pr_{g \xleftarrow{\$} \mathsf{OPF}_{\mathcal{D},\mathcal{R}}} \left[A^{g(\cdot),g^{-1}(\cdot)} = 1 \right] \;=\; \Pr \left[A^{\mathbf{LazySample}(\mathcal{D},\mathcal{R},\cdot),\mathbf{LazySampleInv}(\mathcal{D},\mathcal{R},\cdot)} = 1 \right] ,$$

where g^{-1} denotes the inverse of OPF g.

We clarify that in the theorem, A's oracles for **LazySample**, **LazySampleInv** in the right-hand-side experiment share and update joint state. It is straightforward to check, via simple probability calculations, that the theorem holds for an adversary A that makes one query. The case of multiple queries is harder. The reason is that the distribution of the responses given to subsequent queries depends on which queries A has already made, and this distribution is difficult to compute directly. Instead

our proof uses strong induction in a way that parallels the recursive nature of our algorithms. The proof is located in Appendix A.2.

3.3.4 Efficiency

We characterize efficiency of our algorithms in terms of the number of recursive calls made by **LazySample** or **LazySampleInv** before termination. (The proposition below is just stated in terms of **LazySample** for simplicity; the analogous result holds for **LazySampleInv**.)

Proposition 3.3.3. *The number of recursive calls made by* **LazySample** *is at most* $\log N + 1$ *in the worst-case and at most* $5 \log M + 12$ *on average.*

The proof is located in Appendix A.3 and relies on some bounds by Chvátal on the tail of the hypergeometric distribution. However, we must note that one of the results of Proposition 3.3.3 was recently improved by Yum and Lee [66] through more in-depth analysis. They showed that our scheme in fact recurs less than $\log M + 3$ times on average. We will thus use this bound.

Note that the algorithms make one call to HGD on each recursion, so an upper-bound on their running times is then at most $(\log N + 1) \cdot T_{\text{HGD}}$ in the worst-case and at most $(\log M + 3) \cdot T_{\text{HGD}}$ on average, where T_{HGD} denotes the running time of HGD on inputs of size at most $\log N$. However, this does not take into account the fact that the size of these inputs decrease on each recursion. Thus, better bounds may be obtained by analyzing the running time of a specific realization of HGD.

3.3.5 Realizing HGD

Kachitvichyanukul and Schmeiser [42] designed an efficient implementation of a sampling algorithm HGD for the hypergeometric distribution. Their algorithm is exact; it is not an approximation by a related distribution. It is implemented in Wolfram Mathematica and other libraries, and is fast even for large parameters. However, on

small parameters the algorithms of [63] perform better. Since the parameter size to HGD in our **LazySample** algorithms shrinks across the recursive calls from large to small, it could be advantageous to switch algorithms at some threshold. We refer the reader to [63, 42, 43, 26] for more details.

We comment that the algorithms of [42] are technically only "exact" when the underlying floating-point operations can be performed to infinite precision. In practice, one has to be careful of truncation error. For simplicity, Theorem 3.3.2 does not take this into account, as in theory the error can be made arbitrarily small by increasing the precision of floating-point operations (independently of M, N). But we make this point explicit in Theorem 3.4.3 where we analyze security of our actual scheme.

3.4 Our OPE Scheme and its Analysis

Algorithms **LazySample**, **LazySampleInv** cannot be directly converted into encryption and decryption procedures because they share and update a joint state, namely arrays F and I, which store the outputs of the randomized algorithm HGD. For our actual scheme, we can eliminate this shared state by implementing the subroutine GetCoins (which produces coins for HGD) as a PRF, and re-constructing entries of F and I on-the-fly as needed. However, coming up with a practical yet provably secure construction requires some care. Below we give the details of our PRF implementation, which we call `TapeGen`.

3.4.1 The `TapeGen` PRF

Length-Flexible PRFs. In practice, it is desirable that `TapeGen` be both variable input-length (VIL)- and variable output-length (VOL)-PRF,[2] a primitive we call a *length-flexible* (LF)-PRF. (In particular, the number of coins used by HGD can be beyond one block of an underlying blockcipher in length, ruling out the use of most

[2]That is, a VIL-PRF takes inputs of varying lengths. A VOL-PRF produces outputs of varying lengths specified by an additional input parameter.

practical pseudorandom VIL-MACs.) That is, LF-PRF `TapeGen` with key-space *Keys* takes as input a key $K \in \mathit{Keys}$, an output length 1^ℓ, and $x \in \{0,1\}^*$ to return $y \in \{0,1\}^\ell$. Define the following oracle R taking inputs 1^ℓ and $x \in \{0,1\}^*$ to return $y \in \{0,1\}^\ell$, which maintains as state an array D (initially empty; i.e., $D[x]$ is the empty string for all x):

$$\begin{array}{l} \textbf{Oracle } R(1^\ell, x) \\ \quad \text{If } |D[x]| < \ell \text{ then} \\ \quad\quad r \xleftarrow{\$} \{0,1\}^{\ell - |D[x]|} \\ \quad\quad D[x] \leftarrow D[x] \| r \\ \quad \text{Return } D[x]_1 \ldots D[x]_\ell \end{array}$$

Above and in what follows, m_i denotes the i-th bit of a string m, and we require everywhere that $\ell < \ell_{\max}$ for an associated maximum output length $\ell_{\max}$. For an adversary A, define its *length-flexible pseudorandom function (LF-PRF) advantage* against `TapeGen` as

$$\mathbf{Adv}^{\mathrm{prf}}_{\mathtt{TapeGen}}(A) \;=\; \Pr\left[\, A^{\mathtt{TapeGen}(K,\cdot,\cdot)} = 1 \,\right] - \Pr\left[\, A^{R(\cdot,\cdot)} = 1 \,\right],$$

where the left probability is over the random choice of $K \in \mathit{Keys}$. Most practical VIL-MACs (message authentication codes) are PRFs and are therefore VIL-PRFs, but the VOL-PRF requirement does not seem to have been addressed previously. To achieve it we suggest using a VOL-PRG (pseudorandom generator) as well. Let us define the latter.

Variable-output-length PRGs. Let G be an algorithm that on input a seed $s \in \{0,1\}^k$ and an output length 1^ℓ returns $y \in \{0,1\}^\ell$. Let $\mathcal{O}_G$ be the oracle that on input 1^ℓ chooses a random seed $s \in \{0,1\}^k$ and returns $G(s,\ell)$, and let S be the oracle that on input 1^ℓ returns a random string $r \in \{0,1\}^\ell$. For an adversary A, define its *variable-output-length pseudorandom function (VOL-PRG) advantage* against G as

$$\mathbf{Adv}^{\text{vol-prg}}_{G}(A) \;=\; \Pr\left[\, A^{\mathcal{O}_G(\cdot)} = 1 \,\right] - \Pr\left[\, A^{S(\cdot)} = 1 \,\right].$$

As mentioned above, we require above that $\ell < \ell_{\max}$ for an associated maximum output length $\ell_{\max}$. Call G *consistent* if $\Pr[\, G(s,\ell') = G(s,\ell)_1 \ldots G(s,\ell)_{\ell'} \,] = 1$ for all $\ell' < \ell$, with the probability over the choice of a random seed $s \in \{0,1\}^k$. Many PRGs are consistent due to their "iterated" structure.

Our LF-PRF construction. We propose a general construction of an LF-PRF that composes a VIL-PRF with a consistent VOL-PRG by using the output of the former as the seed for the latter. Formally, let F be a VIL-PRF and G be a consistent VOL-PRG, and define the associated pseudorandom tape generation function `TapeGen` which on inputs $K, 1^\ell, x$ returns $G(1^\ell, F(K,x))$.

The following, proved in Appendix A.4, says that `TapeGen` is indeed an LF-PRF if F is a VIL-PRF and G is a VOL-PRG.

Proposition 3.4.1. *Let A be an adversary against* `TapeGen` *that makes at most q queries to its oracle of total input length ℓ_{in} and total output length ℓ_{out}. Then there exists an adversary B_1 against F and an adversary B_2 against G such that*

$$\mathbf{Adv}^{\text{prf}}_{\texttt{TapeGen}}(A) \le \mathbf{Adv}^{\text{prf}}_{F}(B_1) + \mathbf{Adv}^{\text{vol-prg}}_{G}(B_2)\ .$$

Adversaries B_1, B_2 make at most q queries of total input length ℓ_{in} and total output length ℓ_{out} to their respective oracles and run in the time of A.

Concretely, we suggest the following blockcipher-based consistent VOL-PRG for G. Let $E\colon \{0,1\}^k \times \{0,1\}^n \to \{0,1\}^n$ be a blockcipher. Define the associated VOL-PRG $G[E]$ with seed-length k and maximum output length $n \cdot 2^n$, where G[E] on input $s \in \{0,1\}^k$ and 1^ℓ outputs the first ℓ bits of $E(s,\langle 1\rangle) \| E(s,\langle 2\rangle) \| \cdots$ (Here $\langle i\rangle$ denotes the n-bit binary encoding of $i \in \mathbb{N}$.) The following, proved in Appendix A.5, says that $G[E]$ is a consistent VOL-PRG if E is a PRF.

Proposition 3.4.2. *Let $E\colon \{0,1\}^k \times \{0,1\}^n \to \{0,1\}^n$ be a blockcipher, and let A be an adversary against $G[E]$ making at most q oracle queries whose responses total*

at most $p \cdot n$ bits. Then there is an adversary B against E such that

$$\mathbf{Adv}^{\text{vol-prg}}_{G[E]}(A) \leq q \cdot \mathbf{Adv}^{\text{prf}}_{E}(B) \ .$$

Adversary B makes at most p queries to its oracle and runs in the time of A. Furthermore, $G[E]$ is consistent.

Now, to instantiate the VIL-PRF F in the `TapeGen` construction, we suggest OMAC (a.k.a. CMAC) [40], which is also blockcipher-based and introduces no additional assumption. Then the secret key for `TapeGen` consists only of that for OMAC, which in turn consists of just one key for the underlying blockcipher (e.g. AES).

3.4.2 OPE scheme and analysis

The scheme. Let `TapeGen` be as above, with key-space *Keys*. Our associated order-preserving encryption scheme $\mathcal{OPE}^{\text{HGD}}[\texttt{TapeGen}] = (\mathcal{K}^{\text{HGD}}, \mathcal{E}nc^{\text{HGD}}, \mathcal{D}ec^{\text{HGD}})$ is defined as follows. The plaintext and ciphertext spaces are sets of consecutive integers $\mathcal{D}, \mathcal{R}$, respectively. Algorithm $\mathcal{K}^{\text{HGD}}$ returns a random $K \in \mathit{Keys}$. Algorithms $\mathcal{E}nc^{\text{HGD}}, \mathcal{D}ec^{\text{HGD}}$ are the same as **LazySample**, **LazySampleInv**, respectively, except that HGD is implemented by the algorithm of [42] and GetCoins by `TapeGen` (so there is no need to store the elements of F and I). See Figure 3 for the formal descriptions of $\mathcal{E}nc^{\text{HGD}}$ and $\mathcal{D}ec^{\text{HGD}}$, where as before $\ell_1 = \ell(M, N, y - r)$ is the number of coins needed by HGD on inputs $M, N, y - r$, and $\ell_{\mathcal{R}}$ is the number of coins needed to select an element of $\mathcal{R}$ uniformly at random. (The length parameters to `TapeGen` are just for convenience; one can always generate more output bits on-the-fly by invoking `TapeGen` again on a longer such parameter. In fact, our implementation of `TapeGen` can simply pick up where it left off instead of starting over.)

Security. The following theorem, proved in Appendix A.6, characterizes security of our OPE scheme, saying that it is POPF-CCA-secure if `TapeGen` is a LF-PRF. Applying Proposition 3.4.2, this is reduced to pseudorandomness of an underlying

$\underline{\mathcal{E}nc_K^{\mathrm{HGD}}(\mathcal{D},\mathcal{R},m)}$	$\underline{\mathcal{D}ec_K^{\mathrm{HGD}}(\mathcal{D},\mathcal{R},c)}$
01 $M \leftarrow \lvert\mathcal{D}\rvert$; $N \leftarrow \lvert\mathcal{R}\rvert$	17 $M \leftarrow \lvert\mathcal{D}\rvert$; $N \leftarrow \lvert\mathcal{R}\rvert$
02 $d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$	18 $d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$
03 $y \leftarrow r + \lceil N/2 \rceil$	19 $y \leftarrow r + \lceil N/2 \rceil$
04 If $\lvert\mathcal{D}\rvert = 1$ then	20 If $\lvert\mathcal{D}\rvert = 1$ then $m \leftarrow \min(\mathcal{D})$
05 $\quad cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_{\mathcal{R}}}, (\mathcal{D},\mathcal{R},1\|m))$	21 $\quad cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_{\mathcal{R}}}, (\mathcal{D},\mathcal{R},1\|m))$
06 $\quad c \xleftarrow{cc} \mathcal{R}$	22 $\quad w \xleftarrow{cc} \mathcal{R}$
07 $\quad$ Return c	23 $\quad$ If $w = c$ then return m
	24 $\quad$ Else return $\perp$
08 $cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_1}, (\mathcal{D},\mathcal{R},0\|y))$	25 $cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_1}, (\mathcal{D},\mathcal{R},0\|y))$
09 $x \xleftarrow{\$} d + \mathrm{HGD}(M, N, y - r; cc)$	26 $x \xleftarrow{\$} d + \mathrm{HGD}(M, N, y - r; cc)$
10 If $m \le x$ then	27 If $c \le y$ then
11 $\quad \mathcal{D} \leftarrow \{d+1, \ldots, x\}$	28 $\quad \mathcal{D} \leftarrow \{d+1, \ldots, x\}$
12 $\quad \mathcal{R} \leftarrow \{r+1, \ldots, y\}$	29 $\quad \mathcal{R} \leftarrow \{r+1, \ldots, y\}$
13 Else	30 Else
14 $\quad \mathcal{D} \leftarrow \{x+1, \ldots, d+M\}$	31 $\quad \mathcal{D} \leftarrow \{x+1, \ldots, d+M\}$
15 $\quad \mathcal{R} \leftarrow \{y+1, \ldots, r+N\}$	32 $\quad \mathcal{R} \leftarrow \{y+1, \ldots, r+N\}$
16 Return $\mathcal{E}nc_K^{\mathrm{HGD}}(\mathcal{D},\mathcal{R},m)$	33 Return $\mathcal{D}ec_K^{\mathrm{HGD}}(\mathcal{D},\mathcal{R},c)$

Figure 3: Encryption $\mathcal{E}nc^{\mathrm{HGD}}$ and decryption $\mathcal{D}ec^{\mathrm{HGD}}$ algorithms for our hypergeometric distribution-based OPE scheme, $\mathcal{OPE}^{\mathrm{HGD}}[\mathtt{TapeGen}]$.

blockcipher.

Theorem 3.4.3. *Let $\mathcal{OPE}^{\mathrm{HGD}}[\mathtt{TapeGen}]$ be the above OPE scheme with plaintext space size M, ciphertext space size N. For adversary A against $\mathcal{OPE}^{\mathrm{HGD}}[\mathtt{TapeGen}]$ making at most q queries to its oracles combined, there is an adversary B against* TapeGen *such that*

$$\mathbf{Adv}^{\text{popf-cca}}_{\mathcal{OPE}^{\mathrm{HGD}}[\mathtt{TapeGen}]}(A) \leq \mathbf{Adv}^{\mathrm{prf}}_{\mathtt{TapeGen}}(B) \ + \ \lambda \ .$$

Adversary B makes at most $q_1 = q \cdot (\log N + 1)$ queries of size at most $5 \log N + 1$ to its oracle, whose responses total $q_1 \cdot \lambda'$ bits on average, and its running time is that of A. Above, λ, λ' are constants depending only on HGD *and the precision of the underlying floating-point computations (not on M, N).*

EFFICIENCY. The efficiency of our scheme follows from our previous analyses. Using the suggested implementation of `TapeGen` in Subsection 3.4.1, encryption and decryption require the time for at most $\log N + 1$ invocations of HGD on inputs of size at most $\log N$ plus at most $(\log M + 3) \cdot (5 \log N + \lambda' + 1)/128$ invocations of AES on average for λ' in the theorem.

3.4.3 On choosing N

Practitioners interested in implementing our scheme might naturally wonder how large we recommend making the ciphertext space size N. In fact, different choices of N have no bearing on our scheme's POPF-CCA-security. Rather, different choices of N will affect how the ideal object, a random OPF, behaves. Thus, in order to say something meaningful about the choice of N, we first need a security definition and analysis for the ideal object, which is the subject of Chapter 4.

3.5 On Using the Negative Hypergeometric Distribution

In the balls-and-bins model described in Section 3.3.1 with M black and $N-M$ white balls in the bin, consider the random variable Y describing the total number of balls in our sample after we pick the x-th black ball. This random variable follows the *negative hypergeometric* (NHG) distribution. Formally,

$$P_{NHGD}(y; N, M, x) = \frac{\binom{y-1}{x-1} \cdot \binom{N-y}{M-x}}{\binom{N}{M}}.$$

As we discussed in Section 3.1, use of the NHG distribution instead of the HG permits slightly simpler and more efficient lazy-sampling algorithms and corresponding OPE scheme. The problem is that we require an efficient NHG sampling algorithm, and the existence of such an algorithm is apparently open. What is known is that the NHG distribution can be approximated by the negative binomial distribution [51], the latter can be sampled efficiently [28, 26], and the approximation improves as M and N grow. However, quantifying the quality of the approximation for fixed parameters

seems difficult. If future work either develops an efficient exact sampling algorithm for the NHG distribution or shows that the approximation by the negative binomial distribution is sufficiently close, then our NHG-based OPE scheme could be a good alternative to the HG-based one. Here are the details.

3.5.1 Construction of the NHGD-based OPE scheme

Assume there exists an efficient algorithm NHGD that efficiently samples according to the NHG distribution, possibly using an approximation to a related distribution as we discussed. NHGD takes inputs M, N, and $x \in \{0, 1, \ldots, M\}$ and returns $y \in \{0, 1, \ldots, N\}$ such that for each $y^* \in \{0, 1, \ldots, N\}$ we have $y = y^*$ with probability $P_{NHGD}(y^*; N, M, x)$ over the coins of NHGD. Let $\ell_1 = \ell(M, N, y - r)$ denote the number of coins needed by NHGD on inputs $M, N, y - r$.

Definte $\mathcal{OPE}^{\text{NHGD}}[\texttt{TapeGen}] = (\mathcal{K}, \mathcal{E}nc^{\text{NHGD}}, \mathcal{D}ec^{\text{NHGD}})$, our NHGD-based order-preserving encryption scheme, as follows. Let `TapeGen` be the PRF described in Section 3.4, with key-space *Keys*. The plaintext and ciphertext spaces are sets of consecutive integers $\mathcal{D}, \mathcal{R}$, respectively. Algorithm $\mathcal{K}$ returns a random $K \in Keys$. Algorithms $\mathcal{E}nc^{\text{NHGD}}, \mathcal{D}ec^{\text{NHGD}}$ are described in Figure 4.

3.5.2 Correctness

We prove correctness of the NHGD scheme in the same manner as the HGD scheme. First, see in Figure 5 the revised versions **LazySample***, **LazySampleInv*** of the stateful algorithms from before. The algorithms re-use the subroutine GetCoins, which takes inputs 1^ℓ,$\mathcal{D}$,$\mathcal{R}$, and $b\|z$, where $b \in \{0, 1\}$ and $z \in \mathcal{R}$ if $b = 0$ and $z \in \mathcal{D}$ otherwise, to return $cc \in \{0, 1\}^\ell$. Also, recall that the array I, initially empty, is global and shared between the algorithms.

With these revised versions of **LazySample***, **LazySampleInv***, we supply a revised version of Theorem 3.3.2 for the NHGD case. It is proved in Appendix A.7.

$\underline{\mathcal{E}nc_K^{\mathrm{NHGD}}(\mathcal{D},\mathcal{R},m)}$	$\underline{\mathcal{D}ec_K^{\mathrm{NHGD}}(\mathcal{D},\mathcal{R},c)}$
	16 If $\lvert\mathcal{D}\rvert = 0$ then return $\perp$
01 $M \leftarrow \lvert\mathcal{D}\rvert$; $N \leftarrow \lvert\mathcal{R}\rvert$	17 $M \leftarrow \lvert\mathcal{D}\rvert$; $N \leftarrow \lvert\mathcal{R}\rvert$
02 $d \leftarrow \min(\mathcal{D}) - 1$	18 $d \leftarrow \min(\mathcal{D}) - 1$
03 $r \leftarrow \min(\mathcal{R}) - 1$	19 $r \leftarrow \min(\mathcal{R}) - 1$
04 $x \leftarrow d + \lceil M/2 \rceil$	20 $x \leftarrow d + \lceil M/2 \rceil$
05 $cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_1}, (\mathcal{D},\mathcal{R},x))$	21 $cc \xleftarrow{\$} \mathtt{TapeGen}(K, 1^{\ell_1}, (\mathcal{D},\mathcal{R},x))$
06 $y \leftarrow r + \mathrm{NHGD}(N, M, x-d; cc)$	22 $y \leftarrow r + \mathrm{NHGD}(N, M, x-d; cc)$
07 If $m = x$ then	23 If $c = y$ then
08 Return y	24 Return x
09 If $m < x$ then	25 If $c < y$ then
10 $\mathcal{D} \leftarrow \{d+1, \ldots, x-1\}$	26 $\mathcal{D} \leftarrow \{d+1, \ldots, x-1\}$
11 $\mathcal{R} \leftarrow \{r+1, \ldots, y-1\}$	27 $\mathcal{R} \leftarrow \{r+1, \ldots, y-1\}$
12 Else	28 Else
13 $\mathcal{D} \leftarrow \{x+1, \ldots, d+M\}$	29 $\mathcal{D} \leftarrow \{x+1, \ldots, d+M\}$
14 $\mathcal{R} \leftarrow \{y+1, \ldots, r+N\}$	20 $\mathcal{R} \leftarrow \{y+1, \ldots, r+N\}$
15 Return $\mathcal{E}nc_K^{\mathrm{NHGD}}(\mathcal{D},\mathcal{R},m)$	31 Return $\mathcal{D}ec_K^{\mathrm{NHGD}}(\mathcal{D},\mathcal{R},c)$

Figure 4: Encryption $\mathcal{E}nc^{\mathrm{NHGD}}$ and decryption $\mathcal{D}ec^{\mathrm{NHGD}}$ algorithms for our negative hypergeometric distribution-based OPE scheme, $\mathcal{OPE}^{\mathrm{NHGD}}[\mathtt{TapeGen}]$.

Theorem 3.5.1. *Suppose* GetCoins *returns truly random coins on each new input. Then for any (even computationally unbounded) algorithm A we have*

$$\Pr\left[A^{g(\cdot),g^{-1}(\cdot)} = 1\right] \;=\; \Pr\left[A^{\mathbf{LazySample}^{\star}(\mathcal{D},\mathcal{R},\cdot),\mathbf{LazySampleInv}^{\star}(\mathcal{D},\mathcal{R},\cdot)} = 1\right],$$

where g, g^{-1} denote an order-preserving function picked at random from $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$ and its inverse, respectively.

Now, it is straightforward to prove the formal statement of correctness as before.

Theorem 3.5.2. *Let $\mathcal{OPE}^{\mathrm{NHGD}}[\mathtt{TapeGen}]$ be the OPE scheme defined above with plaintext-space of size M and ciphertext space of size N. Then for any adversary A against $\mathcal{OPE}^{\mathrm{NHGD}}[\mathtt{TapeGen}]$ making at most q queries to its oracles combined, there is an adversary B against* TapeGen *such that*

$$\mathbf{Adv}^{\text{popf-cca}}_{\mathcal{OPE}^{\mathrm{NHGD}}[\mathtt{TapeGen}]}(A) \leqq \mathbf{Adv}^{\mathrm{prf}}_{\mathtt{TapeGen}}(B) \;+\; \lambda\,.$$

Adversary B makes at most $q_1 = q \cdot (\log N + 1)$ queries of size at most $5 \log N + 1$ to its oracle, whose responses total $q_1 \cdot \lambda'$ bits on average, and its running time is that

LazySample$^\star(\mathcal{D}, \mathcal{R}, m)$

$M \leftarrow |\mathcal{D}|$; $N \leftarrow |\mathcal{R}|$
$d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$
$x \leftarrow d + \lceil M/2 \rceil$
If $I[\mathcal{D}, \mathcal{R}, x]$ is undefined then
  $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_1}, \mathcal{D}, \mathcal{R}, 1 \| x)$
  $I[\mathcal{D}, \mathcal{R}, x] \xleftarrow{\$}$ NHGD$(M, N, x - d; cc)$
$y \leftarrow r + I[\mathcal{D}, \mathcal{R}, x]$
If $m = x$ then
  Return y
If $m < x$ then
  $\mathcal{D} \leftarrow \{d + 1, \ldots, x - 1\}$
  $\mathcal{R} \leftarrow \{r + 1, \ldots, y - 1\}$
Else
  $\mathcal{D} \leftarrow \{x + 1, \ldots, d + M\}$
  $\mathcal{R} \leftarrow \{y + 1, \ldots, r + N\}$
Return **LazySample$^\star(\mathcal{D}, \mathcal{R}, m)$**

LazySampleInv$^\star(\mathcal{D}, \mathcal{R}, c)$

If $|\mathcal{D}| = 0$ then return $\perp$
$M \leftarrow |\mathcal{D}|$; $N \leftarrow |\mathcal{R}|$
$d \leftarrow \min(\mathcal{D}) - 1$; $r \leftarrow \min(\mathcal{R}) - 1$
$x \leftarrow d + \lceil M/2 \rceil$
If $I[\mathcal{D}, \mathcal{R}, x]$ is undefined then
  $cc \xleftarrow{\$} \text{GetCoins}(1^{\ell_1}, \mathcal{D}, \mathcal{R}, 1 \| x)$
  $I[\mathcal{D}, \mathcal{R}, x] \xleftarrow{\$}$ NHGD$(M, N, x - d; cc)$
$y \leftarrow r + I[\mathcal{D}, \mathcal{R}, x]$
If $c = y$ then
  Return x
If $c < y$ then
  $\mathcal{D} \leftarrow \{d + 1, \ldots, x - 1\}$
  $\mathcal{R} \leftarrow \{r + 1, \ldots, y - 1\}$
Else
  $\mathcal{D} \leftarrow \{x + 1, \ldots, d + M\}$
  $\mathcal{R} \leftarrow \{y + 1, \ldots, r + N\}$
Return **LazySampleInv$^\star(\mathcal{D}, \mathcal{R}, c)$**

Figure 5: Algorithms **LazySample*** and **LazySampleInv*** for lazy-sampling a pseudorandom order-preserving function and its inverse by sampling the negative hypergeometric distribution.

of A. Above, λ, λ' are constants depending only on NHGD *and the precision of the underlying floating-point computations (not on M, N).*

Proof. The proof of this theorem is identical to that of Theorem 3.4.3, except that it uses Theorem 3.5.1 as a lemma rather than Theorem 3.3.2. □

3.5.3 Efficiency of the NHGD scheme

Efficiency-wise, it is not hard to see that to encrypt a single plaintext, each algorithm performs $\log M + 1$ recursions in the worst-case (as opposed to $\log N + 1$ for the HG-based algorithms), as the algorithm finds the desired plaintext via a binary search over the plaintext space, at each recursion calling NHGD to determine the encryption of the midpoint (defined as the last plaintext in the first half of the current plaintext

domain). The expected number of recursions is easily deduced as

$$\frac{1}{M} \cdot \left[(\log M + 1) + \sum_{k=1}^{\log M} 2^{k-1} k \right] .$$

A simple inductive proof shows that this value is between $\log M - 1$ and $\log M$. This falls in line with what we expect from a binary-search strategy, where the expected number of iterations is typically only about 1 fewer than the worst-case number of iterations.

The algorithms of the corresponding OPE scheme can be obtained following the same idea of eliminating state by using a length-flexible PRF as described in Section 3.4.2. The security statement is the same as that of Theorem 3.4.3, where the last term now corresponds to the error probability of the NHGD algorithm.

CHAPTER IV

ONE-WAYNESS OF PSEUDORANDOM ORDER-PRESERVING FUNCTIONS

Chapter 3 laid out the first formal cryptographic treatment of OPE, as published in [15]. In it, we proposed a security requirement for OPE and demonstrated an efficient blockcipher-based scheme provably meeting the security definition. However, we also noted that the POPF-CCA security notion demands further security analysis. To see why, we revisit the definition.

The POPF-CCA notion (hereafter shortened to POPF) calls an OPE scheme secure if oracle access to its encryption algorithm is indistinguishable from oracle access to a random order-preserving function (ROPF) on the same domain and range. An ROPF is thus the "ideal object" in the POPF definition, analogous to the way that a random function is the ideal object in the classical security notion of pseudorandom function (PRF). However, the ideal objects here, an ROPF and a random function, have fundamentally different security properties. A random function's behavior is well understood: on a new input the output is a uniformly random point in the range, independent of other outputs. Hence, an adversary seeing a function value learns absolutely no information about the pre-image, unless the former happens to coincide with one it has previously seen. But the situation with a random OPF is much harder to describe. It is clear that a random OPF cannot provide such strong security, but what exactly is leaked about the data and what is protected? The distribution of ciphertexts is known and it is not immediately clear if encryption is even one-way.

In this chapter, which contains material published in [16], we make a step towards

addressing these questions. We revisit the security of the ideal object, an ROPF, and provide results that help characterize what it leaks and what it protects about the underlying data. In addition, we observe that it may be possible to achieve stronger security notions than POPF using schemes that fall outside the OPE class but nevertheless allow efficient range queries on encrypted data, and propose two such schemes. We now discuss our contributions in more detail.

4.1 Overview

New Definitions for Studying ROPF Security. As just explained, a random order-preserving function—the ideal object in the POPF definition from Chapter 3—itself (perhaps surprisingly) requires a cryptographic treatment.

In order to better understand the strengths and limitations of encryption with an ROPF we first propose several security notions. One captures a basic one-wayness security and measures the probability that an adversary, given a set of ciphertexts of random messages, decrypts one of them. (The fact that messages are chosen uniformly at random we call the "uniformity assumption," and it will be discussed later.) We give the adversary multiple challenge ciphertexts because this corresponds to practical settings and because the ciphertexts are not independent from each other: learning more points of the OPE function may give the adversary additional information. We actually consider a more general security notion that asks the adversary, given a set of ciphertexts of random messages, to guess an interval (window) within which the underlying challenge plaintext lies. This definition helps us get a better sense of how accurately the adversary can identify the location of a data point. The size of the window and the number of challenge ciphertexts are parameters of the definition. When the window size is one, the notion collapses to the case of simple one-wayness.

Our subsequent definitions address leakage of information not about the *location* of the data points but rather the *distances* between them, which seems crucial in other

applications (e.g., a database of salaries). Indeed, Theorem 3.2.1 of Chapter 3 showed that on a practically-sized ciphertext space, an ROPF, like any OPE, must leak some information relating to distances between plaintexts. We attempt to clarify this intuition. We consider a definition measuring the adversary's success in (precisely) guessing the distance[1] between the plaintexts corresponding to any two out of the set of ciphertexts of random messages given to the adversary. Again, we also consider a more general definition where the adversary is allowed to specify a window in which the distance falls.

We analyze security of an ROPF under these definitions as we believe this helps to understand secure pseudorandom OPE schemes' security guarantees and limitations, and also to evaluate the risk of their usage in various applications. (Indeed, we believe they capture the information about data, namely location and relative distances, that real-world practitioners are most likely to care about.) However, especially in light of the uniformity assumption (which is unlikely to be satisfied in practice), we view our results as providing important steps in the direction of this understanding (as even under this assumption our results are challenging to prove). Still, we advise against practical use of OPE unless a practitioner is fully aware, and accepting, of what has and has not been provably shown about its security.

Analysis of an ROPF. We first give an upper bound on the one-wayness advantage of any adversary attacking an ROPF. The proof is quite involved (and is explained in detail in the Appendix), but the result is a very concise, understandable bound that, under reasonable assumptions, does not even depend on the size of the ciphertext space. (Intuitively, an ROPF's one-wayness comes from the function's probability of deviating from points on the linear OPF $m \mapsto (N/M)m$. Increasing the ciphertext space size beyond a certain amount has little to no effect on these deviations.) We

[1]Technically, for purposes that will become clear in the paper, "distance" actually refers to "directed modular distance," i.e. the distance from one point "up" to the other point, possibly wrapping around the space. As such, distance in this context is non-commutative.

evaluate the bound for several parameters to get an idea of its quality. Our evaluation demonstrates that on practical parameters ROPF and POPF-secure OPEs significantly resist one-wayness attacks, i.e. the maximum one-wayness advantage of any adversary is quite low.

On the other hand, our ROPF analysis under the window one-wayness definition shows that a very efficient adversary can successfully break window one-wayness if the size of the window is not very small. In particular, for message space size M and arbitrary constant b, if the window size is approximately $b\sqrt{M}$, there exists an adversary A whose window one-wayness advantage is at least $1 - 2e^{-b^2/2}$. Thus, for b large enough (say, $b \geq 8$), there exists an adversary with window one-wayness advantage very close to one.

We then extend our analysis of an ROPF to the distance one-wayness and window distance one-wayness definitions. Using similar techniques we show entirely analogous results, namely that the former is very small but the latter becomes large when the adversary is allowed to specify a window of size approximately $b\sqrt{M}$.

We conclude our ROPF analysis with several important supplemental remarks regarding the effect of known-plaintext attacks in the schemes, choosing an appropriate ciphertext space size, and the need to satisfy the uniformity assumption in practical implementations.

Achieving Stronger Security. We next consider the question of whether different types of schemes that support efficient range queries can achieve stronger security than POPF. To capture such schemes we introduce a general notion of *efficiently orderable encryption* (EOE), that covers all schemes supporting standard range queries by requiring a publicly computable function that determines order of the underlying plaintexts given any two ciphertexts. Since EOE leaks order of ciphertexts, the indistinguishability under ordered chosen-plaintext attack (IND-OCPA) definition, which Chapter 3 introduced and showed OPE cannot achieve, is an ideal level of security

for EOE schemes.

An Optimally Secure Committed EOE Scheme. We focus on a scenario where we can show something like IND-OCPA security is possible. We define "committed" versions of EOE and IND-OCPA, called CEOE and IND-CCPA, corresponding to a setting where the database is static and completely known to the user in advance of encryption. Such a scenario is apparently important as it was considered in the first paper to propose an order-preserving scheme [2], and was also studied in several works including [23] for the case of exact-match queries. We observe that the more restrictive functionality in this setting allows one to achieve IND-CCPA. We propose a new scheme that uses a monotone minimal perfect hash function (MMPHF) directly as an "order preserving tagging algorithm" for the given message set, together with a secure encryption. The construction allows for easy implementation of range queries while also achieving the strongest security. Moreover, while MMPHFs are known to require long keys [5], recent constructions [5] are close to being space-optimal. Thus, this application of MMPHFs for tagging seems to be a novel, nearly efficient-as-possible way to support range queries, leaking nothing but the order of ciphertexts, when the database is fixed in advance.

A New Modular OPE Scheme and its Analysis. Finally, we propose a technique that improves on the security of any OPE scheme without sacrificing efficiency. Recall that our ROPF analysis reveals that OPE leaks information about the *locations* of the data points in addition to the distances between them. We suggest a modification to (that can be viewed as a generalization of) an OPE scheme that overcomes this. The resulting scheme is not order-preserving per se, but still permits range queries—in this case, modular range queries. (When the left end of the queried range is greater than the right end, a modular range query returns the "wrap-around range," i.e. everything greater than the left end or less than the right end.) The modification to the scheme is simple and generic: the encryption algorithm just adds

(modulo the size of the message space) a secret offset to the message before encryption. (The secret offset is the same for all messages.) We call a scheme obtained this way a modular OPE scheme, and generalize the security notion: the ideal object is now a random modular OPF (RMOPF), i.e. a random OPF applied to messages with a randomly picked offset. It is easy to see that any secure OPE scheme yields a secure modular OPE scheme using the above transformation.

We show that a random modular OPF, unlike a random OPF, completely hides the locations of the data points (but has the same leakage with respect to distance and window-distance one-wayness). On the other hand, if the adversary is able to recover a single known plaintext-ciphertext pair, security falls back to that of a random OPF.

We also note that the technique with a secret offset can be applied to the CEOE scheme to enhance its security even beyond IND-CCPA when support for modular range queries is sufficient.

4.2 Primitives and Definitions

Types of Range Queries. For fixed plaintext and ciphertext spaces $[M]$ and $[N]$, a range query *target* is a pair of plaintexts (m_L, m_R) that comes in two varieties: *standard* if $m_L \leq m_R$, or *wrap-around* if $m_L > m_R$. If (m_L, m_R) is a target, its associated *range* is $[m_L, m_R]$ in the standard case and $[m_L, M] \cup [1, m_R]$ in the wrap-around case.

To model the intended application, suppose a server has a database encrypted under a scheme $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ with key $K \xleftarrow{\$} \mathcal{K}$. In a *standard range query*, the user submits two unordered ciphertexts $\{c_1, c_2\}$ to the server. Let $(m_1, m_2) = \mathcal{D}ec(K, (c_1, c_2))$. Then the target is $(\min\{m_1, m_2\}, \max\{m_1, m_2\})$, and the server must return the set of ciphertexts in the database whose decryptions fall into the associated range. Notice that these targets are always standard.

In a *modular range query*, the user submits two ordered ciphertexts (c_L, c_R). Let

$(m_L, m_R) = \mathcal{D}ec(K, (c_L, c_R))$. Then the range query target is (m_L, m_R), and the server must return the set of ciphertexts in the database whose decryptions fall into the associated range. Notice that these targets can be standard or wrap-around.

WEAK POPF NOTION. We now define the weak (chosen-plaintext-only) version of the POPF security definition from Section 3.2. (For simplicity, we do not discuss chosen-ciphertext attacks in detail. Note that symmetric schemes such as these can be made resistant to chosen-ciphertext attacks by using Encrypt-then-MAC [6] generic constructions that prevent adversaries from constructing valid ciphertexts.)

Fix an order-preserving encryption scheme $\mathcal{SE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ with plaintext space $\mathcal{D}$ and ciphertext space $\mathcal{R}$, $|\mathcal{D}| \leq |\mathcal{R}|$. For an adversary A against $\mathcal{SE}$, recall its *pseudorandom order-preserving function (POPF) advantage* against $\mathcal{SE}$:

$$\mathbf{Adv}^{\text{popf}}_{\mathcal{SE}}(A) = \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[A^{\mathcal{E}nc(K,\cdot)} = 1\right] - \Pr_{g \xleftarrow{\$} \mathsf{OPF}_{\mathcal{D},\mathcal{R}}}\left[A^{g(\cdot)} = 1\right].$$

Informally, an OPE scheme is POPF-secure if this quantity is small, i.e., if oracle access to its encryption function is indistinguishable from oracle access to the "ideal object," a random order-preserving function (ROPF) on the same domain and range. Accordingly, we focus in this chapter on analyzing the ideal object, an ROPF. The analysis will then apply also to POPF-secure OPE schemes such as the blockcipher-based scheme $\mathcal{OPE}^{\text{HGD}}$ from Section 3.4.

AN "IDEAL" SCHEME ROPF. We define the "ideal" ROPF scheme as follows. Let $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$ denote the set of all order-preserving functions from $\mathcal{D}$ to $\mathcal{R}$. Define $\mathsf{ROPF}_{\mathcal{D},\mathcal{R}} = (\mathcal{K}_{\text{r}}, \mathcal{E}nc_{\text{r}}, \mathcal{D}ec_{\text{r}})$ as the following deterministic order-preserving encryption scheme:

- $\mathcal{K}_{\text{r}}$ returns a random element g of $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$.
- $\mathcal{E}nc_{\text{r}}$ takes the key and a plaintext m to return $g(m)$.
- $\mathcal{D}ec_{\text{r}}$ takes the key and a ciphertext c to return $g^{-1}(c)$.

Of course the above scheme is not computationally efficient, but our goal is its security analysis for the purpose of clarifying security of all POPF-secure constructions.

MOST LIKELY PLAINTEXT. Let $\mathcal{SE}_{\mathcal{D},\mathcal{R}} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ be a symmetric encryption scheme on domain $\mathcal{D}$, range $\mathcal{R}$. For given $c \in \mathcal{R}$, if $m_c \in \mathcal{D}$ is a message such that $\Pr_{K \xleftarrow{\$} \mathcal{K}}[\mathcal{E}nc(K, m) = c]$ achieves a maximum at $m = m_c$, then we call m_c a (if unique, "the") *most likely plaintext* for c.

MOST LIKELY PLAINTEXT DISTANCE. Let $\mathcal{SE}_{[M],[N]} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ be a symmetric encryption scheme on domain $[M]$, range $[N]$. For given $c_1, c_2 \in \mathcal{R}$, if $d_{c_1,c_2} \in \{0, \ldots, M-1\}$ such that $\Pr_{K \xleftarrow{\$} \mathcal{K}}[\, m_2 - m_1 \bmod M = d \mid (m_1, m_2) = \mathcal{D}ec(K, (c_1, c_2))\,]$ achieves a maximum at $d = d_{c_1,c_2}$, then we call d_{c_1,c_2} a (if unique, "the") *most likely plaintext distance* from c_1 to c_2.

4.3 One-wayness Security Definitions

As explained in the introduction, the "ideal" ROPF scheme defined in Section 4.2 itself requires a cryptographic treatment. Toward this end, we propose several generalized security definitions that help us understand its security.

Let $\mathcal{SE}_{[M],[N]} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ be a deterministic symmetric encryption scheme.

WINDOW ONE-WAYNESS. The most basic question left unanswered by [15] is whether a POPF-secure scheme is even one-way. Towards this end we start with the one-wayness definition. Our definition is a stronger and more general version of the standard notion of one-wayness. For $1 \leq r \leq M$ and $z \geq 1$, the adversary is given a set of z ciphertexts of (uniformly) random messages and is asked to come up with an interval of size r within which one of the underlying plaintexts lies. We call our notion r, z-window one-wayness (or r, z-WOW). Note that when $r = 1$, the definition collapses to the standard one-wayness definition (for multiple ciphertexts), and we will call it one-wayness for simplicity.

The r, z-*window one-wayness* (r, z-*WOW*) *advantage* of an adversary A against $\mathcal{SE}_{[M],[N]}$ is

$$\mathbf{Adv}^{r,z\text{-wow}}_{[M],[N]}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{r,z\text{-wow}}_{\mathcal{SE}_{[M],[N]}}(A) = 1\,\right],$$

where the experiment $\mathbf{Exp}^{r,z\text{-wow}}_{\mathcal{SE}_{[M],[N]}}(A)$ above is defined in Figure 6.

Experiment $\mathbf{Exp}^{r,z\text{-wow}}_{\mathcal{SE}_{[M],[N]}}(A)$

$K \xleftarrow{\$} \mathcal{K}$; $\mathbf{M} \xleftarrow{\$} \mathrm{Cmb}^{[M]}_z$; $\mathbf{C} \leftarrow \mathcal{E}nc(K, \mathbf{M})$
$(m_L, m_R) \xleftarrow{\$} A(\mathbf{C})$
Return 1 if $(m_R - m_L) \bmod M + 1 \le r$ and there exists $m \in \mathbf{M}$ so that
  either $m \in [m_L, m_R]$ or $(m_L > m_R$ and $m \in [m_L, M] \cup [1, m_R])$
Return 0 otherwise

Figure 6: The window one-wayness experiment.

Notice that the latter success condition in the experiment allows the adversary to specify a window that "wraps around" the message space. Granting this extra power to the adversary will be useful in analyzing the MOPE scheme of Section 4.5.2.

Window Distance One-Wayness. To identify the extent to which an OPE scheme leaks distance between plaintexts, we also provide a definition in which the adversary attempts to guess the interval of size r in which the distance between any two out of z random plaintexts lies, for $1 \le r \le M$ and $z \ge 2$. We call the notion r, z-window distance one-wayness (r, z-WDOW). When $r = 1$, the adversary has to guess the exact distance between any two of z ciphertexts.

The r, z-*window distance one-way* (r, z-*WDOW*) *advantage* of adversary A against scheme $\mathcal{SE}_{[M],[N]}$ is

$$\mathbf{Adv}^{r,z\text{-wdow}}_{[M],[N]}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{r,z\text{-wdow}}_{\mathcal{SE}_{[M],[N]}}(A) = 1\,\right],$$

where the experiment $\mathbf{Exp}^{r,z\text{-wdow}}_{\mathcal{SE}_{[M],[N]}}(A)$ above is defined in Figure 7.

4.4 One-Wayness of a Random OPF

This section is devoted to analyzing the "ideal" scheme $\mathsf{ROPF}_{[M],[N]}$ under the security definitions given in the previous section. The first result shows an upper bound

Experiment $\mathbf{Exp}^{r,z\text{-wdow}}_{\mathcal{SE}_{[M],[N]}}(A)$

$K \xleftarrow{\$} \mathcal{K}$; $\mathbf{M} \xleftarrow{\$} \mathrm{Cmb}_z^{[M]}$; $\mathbf{C} \leftarrow \mathcal{E}nc(K, \mathbf{M})$

$(d_1, d_2) \xleftarrow{\$} A(\mathbf{C})$

Return 1 if $d_2 - d_1 + 1 \leq r$ and there exist distinct $m_i, m_j \in \mathbf{M}$ with $m_j - m_i \bmod M \in [d_1, d_2]$

Return 0 otherwise

Figure 7: The distance window one-wayness experiment.

on $1, z$-WOW advantage against the scheme. This demonstrates that on practical parameters, ROPF and POPF-secure OPEs significantly resist (size-1-window) one-wayness attacks. In contrast, the second result shows the ideal ROPF scheme is susceptible to an efficient large-window (a constant times $\sqrt{M}$) one-wayness attack, by constructing an adversary and lower-bounding its r, z-WOW advantage.

The analysis then proceeds similarly for window distance one-wayness definitions: we will show analogous contrasting results for small- versus large-window experiments. We now turn to the details of the analysis.

4.4.1 Upper and lower bounds on window one-wayness

An Upper Bound on the $1, z$-WOW Advantage. The following theorem states an upper bound on the $1, z$-WOW advantage of any adversary against $\mathsf{ROPF}_{[M],[N]}$.

Theorem 4.4.1. *For any challenge set of size z and adversary A, if $N \geq 2M$ and $M \geq 15 + z$ then*

$$\mathbf{Adv}^{1,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) < \frac{4z}{\sqrt{M - z + 1}} \,.$$

The formal proof is quite involved and is in Appendix B.1. The idea is to first bound $1, z$-WOW security in terms of $1, 1$-WOW security; because ciphertexts are correlated, a simple hybrid argument does *not* work and our reduction instead uses a combinatorial approach, demonstrating a bijection between objects in relevant definitions. Then, to bound $1, 1$-WOW security, we again take a combinatorial strategy, as follows. We consider a ciphertext's most likely plaintext (m.l.p.) and recall the

negative hypergeometric distribution (NHGD). We first relate the middle ciphertext's m.l.p.'s NHGD probability for a given plaintext/ciphertext space to that of a space twice the size; iterating this result produces a formula for the middle ciphertext's m.l.p.'s NHGD probability in a large space given the analogous value in a small space. We then relate *any* ciphertext's m.l.p.'s NHGD probability to that of the middle ciphertext in the space. Finally, we approximate the sum of m.l.p. NHGD probabilities over the ciphertext space in terms of that of the middle ciphertext, and hence to that of the middle ciphertext in a smaller space. Plugging in a value for the m.l.p. NHGD probability on the small space and simplifying yields the bound.

Evaluating the Bound. The bound of Theorem 4.4.1 is quite succinct—it does not even rely on N (as long as $N \geq 2M$). The result in essence shows that as long as the challenge set size z is small compared to M, the bound is a small constant times $z/\sqrt{M}$. This in turn is small as long as z is small compared to $\sqrt{M}$.

Table 2 shows some sample evaluations of the bound for several message space and challenge set sizes.

M	z	Clean bound	M	z	Clean bound
2^{24}	1	2^{-10}	2^{80}	1	2^{-38}
2^{40}	1	2^{-18}	2^{80}	2^{20}	2^{-18}
2^{80}	1	2^{-38}	2^{80}	2^{38}	1
2^{120}	1	2^{-58}			

Table 2: Sample evaluation of Theorem 4.4.1's clean bound for various plaintext space sizes M and challenge set sizes z. All require ciphertext space size $N \geq 2M$.

We see that $\mathsf{ROPF}_{[M],[N]}$ has very good one-wayness security for reasonably-sized parameters. Given the results of [15] our bound for ROPF can be easily adjusted for their POPF construction, by taking into account pseudorandomness of an underlying blockcipher. But as we discussed in the introduction, standard one-wayness may not be sufficient in all applications and we have to also analyze the schemes under other security notions. Thus, we turn to the next result.

A Lower Bound on Large Window One-Wayness. Here we show that there exists a very efficient adversary attacking the window one-wayness of an ROPF for a sufficiently large window size. A more intuitive explanation of the result follows the theorem.

Theorem 4.4.2. *For any window size r and challenge set size z, there exists an efficient adversary A such that*

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \;\geq\; \mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \;\geq\; 1 - 2e^{-\frac{(r-1)^2}{2}\frac{(M-1)}{M^2}}.$$

The proof is in Appendix B.3. There, we construct a straightforward adversary and demonstrate that it has the above probability of success, using some bounds by Chvátal on the tail probabilities of the hypergeometric distribution.

Intuitively, Theorem 4.4.2 implies that for $r \approx b\sqrt{M}$, where b is a large enough constant (say $b \geq 8$), there exists an adversary A whose r-window one-wayness is very close to 1. More precisely, let $r = b\frac{M}{\sqrt{M-1}} + 1$, and the theorem implies there exists an A such that

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \geq 1 - 2e^{-b^2/2}\ .$$

4.4.2 Upper and lower bounds on distance window one-wayness

An Upper Bound on the $1, z$-WDOW Advantage. The following theorem, with the proof in Appendix B.4, states an upper bound on the $1, z$-distance one-wayness of a random OPF that is very similar to the bound in Theorem 4.4.1.

Theorem 4.4.3. *For any challenge set size z and adversary A, if $N \geq 2M$ and $M \geq 16 + z$ then*

$$\mathbf{Adv}^{1,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \leq \frac{4z(z-1)}{\sqrt{M-z+1}}\ .$$

Naturally, as this result looks very much like that of Theorem 4.4.1, the proof follows the same strategy and achieves similar results. The only differences are that

the initial reduction relates r, z-WDOW security to $r, 2$-WDOW security, incurring a factor $z(z-1)$ advantage increase as opposed to just z, and the initial (tight) bound formula replaces parameters N, M with $N-1$, $M-1$. See Appendix B.4 for proof details.

Thus, the $1, z$-window distance one-wayness of a random OPF is upper-bounded in a similar fashion as the $1, z$-window one-wayness, and we conclude that random OPFs have good $1, z$-WDOW security. Again, though, that is not the whole story, as we see next.

A Lower Bound on Window Distance One-Wayness of ROPF. Here, we derive a result similar to that of Theorem 4.4.2, but for the window distance one-wayness of a random OPF.

Theorem 4.4.4. *For any window size r and challenge set size z, there exists an efficient adversary A such that*

$$\mathbf{Adv}^{r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \;\geq\; \mathbf{Adv}^{r,1\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \;\geq\; 1 - 2e^{-\frac{(r-1)^2}{2}\frac{(M-2)}{(M-1)^2}}.$$

The proof appears in Appendix B.5. Intuitively, Theorem 4.4.4 implies that for $r \approx b\sqrt{M}$, where b is a large enough constant (say, $b \geq 8$), there exists an efficient adversary A whose r-window distance one-wayness advantage is very close to 1. More precisely, let $r = b\frac{M-1}{\sqrt{M-2}} + 1$, and the theorem implies there exists an A such that

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \geq 1 - 2e^{-b^2/2}.$$

4.4.3 Further security considerations for ROPFs

In this section, we explore several important questions regarding our ROPF security analysis.

Effect of Known-Plaintext Attacks. It is a natural question to ask what happens to the security of an ROPF scheme when the adversary knows a certain

number of plaintext-ciphertext pairs. In general, we can answer this question for each definition of one-wayness using a simple extension of the arguments above.

In the scheme $\mathsf{ROPF}_{\mathcal{D},\mathcal{R}}$, known plaintext-ciphertext pairs split the plaintext and ciphertext spaces into subspaces. On each subspace, the analysis under each one-wayness definition reduces to that of an ROPF on the domain and range of the subspace. For instance, if (m_1, c_1) and (m_2, c_2) are known for $m_1 < m_2$, and no other known plaintext-ciphertext pairs occur between these two, then for $\mathcal{D}' = \{m \in \mathcal{D} \mid m_1 < m < m_2\}$ and $\mathcal{R}' = \{c \in \mathcal{R} \mid c_1 < c < c_2\}$, we analyze the behavior of the function on this subspace by considering the one-wayness bounds on $\mathsf{ROPF}_{\mathcal{D}',\mathcal{R}'}$.

This brings up an important issue. For much of our analysis to apply to a scheme, it must be the case that the ciphertext space is at least twice the size of the message space. Therefore, in order to make sure that our analysis will still apply to most subspaces once several plaintext-ciphertext pairs are discovered by the adversary, we would like to choose the initial parameters in such a way that subspaces are unlikely to violate this condition.

Choosing the Ciphertext Space Size. This brings us to the question posed in Section 3.4.2: given a plaintext space of size M, what should be the size N of the ciphertext space? Now that we have ways of characterizing the security of an ROPF using our one-wayness definitions, we can more justifiably discuss the question of what to choose for N.

For $g \in \mathsf{OPF}_{[M],[N]}$, if $m_1 < m_2 \in [M]$ exist such that $g(m_2) - g(m_1) < 2(m_2 - m_1)$, then we say that g is *shallow* on the ciphertext interval $[g(m_1), g(m_2)]$. The bounds found in the previous sections assume that $N \geq 2M$. Thus, any non-shallow interval can be analyzed through our theorems about one-wayness, and as a result we would like to choose N to avoid shallow intervals, both in the original space and in potential subspaces.

In particular, consider the following result, which bounds the probability that an

interval between encryptions of two random plaintexts is shallow.

Proposition 4.4.5. *Let $t = (N-1)/(M-1)$, and assume $t \geq 7$. Let $m_1 \xleftarrow{\$} [M]$, $m_2 \xleftarrow{\$} [M] \setminus \{m_1\}$, $K \xleftarrow{\$} \mathcal{K}_{\mathrm{r}}$, $\mathcal{E}nc_{\mathrm{r}}(K,(m_1,m_2)) = (c_1,c_2)$, $w = c_2 - c_1 \bmod M$, and $d = m_2 - m_1 \bmod M$. Then*

$$\Pr_{K,m_1,m_2}[\,2d > w\,] < \frac{3}{t}\frac{1}{\sqrt{(M-1)/\ln M}}\,.$$

The proof can be found in Appendix B.6. Besides using Lemma B.3.1, the proof is mostly algebraic fiddling.

This bound gives us an idea of good values for $t \approx N/M$. In particular, it seems that choosing a constant for $t \geq 7$, that is, taking N to be a constant multiple of M, is sufficient in order to make the above probability negligible. Whether the constant should be large or small depends on one's tolerance for random intervals to be shallow.

On Implementing a Scheme to Support Range Queries using POPF. We stress that most of our analysis relies on the uniformity assumption, namely that challenge messages come from a uniform distribution. Thus, practitioners relying on our one-wayness analysis should take steps to satisfy the uniformity assumption. In particular, underlying messages that are encrypted in a database, as well as queries, should "look" uniform in terms of their location in the message space. These uniformity restrictions could possibly be met by a scheme that performs "dummy" queries, in addition to legitimate queries, in order to make queries look uniformly random.

It is an open problem to extend our analysis to other input distributions other than uniform. However, it seems unlikely that anything positive can be said about OPE schemes' one-wayness for arbitrary distributions or for models where the adversary can choose challenge messages or distributions.

4.5 Achieving Stronger Security

We study new ways to achieve better security than the OPE scheme of [15] while still allowing for efficient range queries on encrypted data. But first, we define a general

primitive, Efficiently Orderable Encryption (EOE), that includes all schemes that support efficient standard range queries, including OPE. We show that IND-OCPA, defined and shown to be unachievable by OPE in [15], is the ideal security definition for such schemes.

We define "committed" analogues of EOE and IND-OCPA, namely CEOE and IND-CCPA, that apply to the practical scenario where the database to encrypt is predetermined and static. Such a setting has been studied in several works on searchable encryption, including the first paper to propose an order-preserving scheme [2, 23]. We then propose a new CEOE scheme that is CCPA-secure.

Finally, we develop a generic modification of an OPE that supports modular range queries (but not standard range queries) and overcomes some of the security weaknesses of any OPE that we studied in Section 4.4. The scheme is not EOE because it does not leak order; rather, it leaks only "modular" order.

Efficiently Orderable Encryption. We say that $\mathcal{EOE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, W)$ is an *efficiently-orderable encryption* (EOE) scheme if $\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec$ are the algorithms of a symmetric encryption scheme, W is an efficient algorithm that takes two ciphertexts as input, and defining $C_K = \{\mathcal{E}nc(K, m) \mid m \in \mathcal{M}\}$ as the set of valid ciphertexts for key K,

$$W(c_0, c_1) = \begin{cases} 1 & \text{if } \mathcal{D}ec(K, c_0) < \mathcal{D}ec(K, c_1) \\ 0 & \text{if } \mathcal{D}ec(K, c_0) = \mathcal{D}ec(K, c_1) \\ -1 & \text{if } \mathcal{D}ec(K, c_0) > \mathcal{D}ec(K, c_1) \end{cases}$$

for any key K and all $c_0, c_1 \in C_K$. It is easy to see that such a scheme permits efficient standard range queries, as the server can keep the encrypted database sorted using W.

It is also clear that any OPE scheme $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ corresponds to an EOE scheme with the same key generation, encryption, and decryption algorithms, and $W(c_0, c_1)$ outputting 1, 0, or -1 if the relation between c_0 and c_1 is $<$, $=$, or $>$, respectively.

But in general an EOE scheme does not have to be deterministic.

4.5.1 Committed Efficiently-Orderable Encryption

Range Queries on a Predetermined Static Database. Now we consider schemes for the settings when it is possible for the user to preprocess the whole data before encrypting and sending it to the server. For that we allow the key generation of an EOE scheme to take the message set as input, which we rename a *committed* EOE scheme.

Committed efficiently-orderable encryption. A *committed efficiently-orderable encryption* (CEOE) scheme on domain $\mathcal{D}$ is a tuple $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, W)$ satisfying the following.

- The randomized key generation algorithm $\mathcal{K}$ takes a message space $\mathcal{M} \subset \mathcal{D}$ (called the *committed* message space) as input and outputs a secret key K.
- For any committed message space $\mathcal{M} \subset \mathcal{D}$, $(\mathcal{K}(\mathcal{M}), \mathcal{E}nc, \mathcal{D}ec, W)$ is an EOE scheme on $\mathcal{M}$.

We will show that a CEOE scheme can achieve very strong security. In particular, it can achieve the "committed" adaptation of the IND-OCPA notion from [15], where the adversary outputs two vectors of plaintexts with the same order and equality pattern and is asked to guess whether it is given encryptions of the first or second vector. We define *indistinguishability under committed chosen plaintext attack (IND-CCPA)*. The definition mimics IND-OCPA except that the adversary chooses the challenge vectors (now viewed as message spaces) before key generation, and the scheme's key generation algorithm takes the appropriate message space as input.

IND-CCPA. Let $\mathcal{CEOE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, W)$ be a CEOE scheme on message space $\mathcal{M}$. For an adversary $A = (A_1, A_2)$, define its *indistinguishability under committed*

chosen plaintext attack (IND-CCPA) advantage against $\mathcal{SE}$ as

$$\mathbf{Adv}^{\text{ind-com-cpa}}_{\mathcal{CEOE}}(A) \;=\; \Pr\left[\mathbf{Exp}^{\text{ind-com-cpa-1}}_{\mathcal{CEOE}}(A)=1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-com-cpa-0}}_{\mathcal{CEOE}}(A)=1\right],$$

where for $b \in \{0,1\}$ the experiments $\mathbf{Exp}^{\text{ind-ccpa-b}}_{\mathcal{CEOE}}(A)$ are defined in Figure 8. Note

Experiment $\mathbf{Exp}^{\text{ind-com-cpa}b}_{\mathcal{CEOE}}(A)$
$(\mathcal{M}_0, \mathcal{M}_1, \sigma) \xleftarrow{\$} A_1$
If $|\mathcal{M}_0| \neq |\mathcal{M}_1|$ then output $\perp$
Let $l = |\mathcal{M}_0| = |\mathcal{M}_1|$
Let $m_1^j < m_2^j < \ldots < m_l^j$ be the elements of $\mathcal{M}_j$, for $j = 0, 1$
If there exist $1 \leq i \leq l$ so that $|m_i^0| \neq |m_i^1|$ then output $\perp$
$K \xleftarrow{\$} \mathcal{K}(\mathcal{M}_b)$
$c_j \leftarrow \mathcal{E}nc(K, m_j^b)$ for $j = 1, \ldots, l$
$d \xleftarrow{\$} A_2(\sigma, c_1, c_1, \ldots, c_l)$
Return d

Figure 8: The IND-CommittedCPA experiment.

that σ denotes a state the adversary can preserve. We say that $\mathcal{CEOE}$ is *IND-CCPA-secure* if the IND-CCPA advantage of any adversary against $\mathcal{CEOE}$ is small.

Our CEOE construction and its security. We now propose a CEOE scheme that will achieve IND-CCPA security. A ciphertext in our scheme consists of a semantically-secure ciphertext of the message concatenated with the tag, which indicates the order of the message in the ordered message list. As a building block for our scheme we use monotone minimal perfect hash functions, defined as follows.

Let $\mathcal{M}$ be a set with a total (lexicographical) order. h is a *monotone minimal perfect hash function* [5] (MMPHF) on $\mathcal{M}$ if h sends the ith largest element of $\mathcal{M}$ to i, for $i = 0, 1, \ldots, |\mathcal{M}| - 1$. Notice that the MMPHF on any given domain $\mathcal{M}$ is unique. So that we can use MMPHFs in the upcoming construction, let an *index tagging scheme* $(\mathcal{K}, \tau)$ be a pair of algorithms such that $\mathcal{K}$ takes a domain $\mathcal{M}$ and outputs a secret key $K_{\mathcal{M}}$ so that $\tau(K_{\mathcal{M}}, \cdot)$ is the (unique) MMPHF for $\mathcal{M}$, while $\tau(K, m) = \perp$ for any $m \notin \mathcal{M}$.

Our CEOE construction is based on two building blocks: MMPHF tagging and any symmetric encryption scheme.

Let $(\mathcal{K}_t, \tau)$ be an index tagging scheme. Let $\mathcal{SE} = (\mathcal{K}', \mathcal{E}nc', \mathcal{D}ec')$ be any symmetric encryption scheme on a fixed universe $\mathcal{D}$. We construct a CEOE scheme $\mathcal{CEOE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, W)$ as follows.

- $\mathcal{K}$ takes $\mathcal{M} \subset \mathcal{D}$ as input, runs $K_t \leftarrow \mathcal{K}_t(\mathcal{M})$ and $K_e \leftarrow \mathcal{K}'$, and returns $K = K_t \| K_e$.
- $\mathcal{E}nc$ takes key $K = K_t \| K_e$ and message m as input, and computes $i = \tau(K_t, m)$. If $i = \perp$ then $\mathcal{E}nc$ returns $\perp$, otherwise it returns $i \| \mathcal{E}nc'(K_e, m)$.
- $\mathcal{D}ec$ takes key $K = K_t \| K_e$ and ciphertext $c = i \| c'$ as input, and returns $\mathcal{D}ec'(K_e, c')$.
- W takes ciphertexts $c_0 = i_0 \| c_0'$ and $c_1 = i_1 \| c_1'$ as input, and returns 1 if $i_0 < i_1$, 0 if $i_0 = i_1$, and -1 if $i_0 > i_1$.

We note that unlike the scheme with pre-processing for exact-match queries [23], when using the above scheme the server does indexing and query processing as for unencrypted data, which is a practical advantage. Also, as the following result shows, the scheme is secure under IND-CCPA. The proof is in Appendix B.7.

Theorem 4.5.1. *The CEOE scheme $\mathcal{CEOE}$ is IND-CCPA-secure provided the underlying symmetric encryption scheme is IND-CPA secure.*

Note that our secure CEOE construction relies on an efficient MMHPF implementation. Luckily, MMHPFs were studied recently by [5]. They showed that for a universe of size 2^w and for $n \geq \log w$, the shortest possible description of an MMPHF function (and thus, best possible key length for a tagging scheme) on n elements is unfortunately quite large at $\Omega(n)$ bits. This is somewhat disheartening, as a naive solution, in which the MMPHF key consists of an n-entry array whose ith entry is

the ith largest element in the domain, has a key length of $O(nw)$. Nevertheless, the authors of [5] were able to generate MMPHF descriptions that are closer to the optimal bound: one construction uses $O(n \log\log w)$ bits and has query time $O(\log w)$, and the other uses $O(n \log w)$ bits and has constant query time. This is still large, but may be practical depending on the parameters involved.

4.5.2 Modular OPE and analysis of an ideal MOPE scheme

Modular OPE. We propose a modification to (that can be viewed as a generalization of) an OPE scheme that improves the security performance of any OPE. The resulting scheme is no longer strictly order-preserving, but it still permits range queries. However, now the queries must be *modular* range queries. Standard range queries are not supported, as only "modular order" rather than order is leaked. The modification from OPE is simple, generic, and basically free computation-wise.

Let $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ be an order-preserving encryption scheme. Define a *modular order-preserving encryption scheme* (MOPE) $\mathcal{SE}_{[M],[N]} = (\mathcal{K}_{\mathrm{m}}, \mathcal{E}nc_{\mathrm{m}}, \mathcal{D}ec_{\mathrm{m}})$ as follows.

- $\mathcal{K}_{\mathrm{m}}$ runs $\mathcal{K}$ to get K, picks $j \xleftarrow{\$} [M]$ and returns (K, j).
- $\mathcal{E}nc_{\mathrm{m}}$ on input (K, j) and m returns $\mathcal{E}nc(K, m - j \bmod M)$.
- $\mathcal{D}ec_{\mathrm{m}}$ on inputs (K, j) and c returns $\mathcal{D}ec(K, c) + j \bmod M$.

Notice that a MOPE is suitable for modular range query support as follows. To request the ciphertexts of the messages in the range $[m_1, m_2]$ (if $m_1 \leq m_2$), or $[m_1, M] \cup [1, m_2]$ (if $m_1 > m_2$), the user computes $c_1 \leftarrow \mathcal{E}nc_{\mathrm{m}}(K, m_1), c_2 \leftarrow \mathcal{E}nc_{\mathrm{m}}(K, m_2)$ and submits ciphertexts (c_1, c_2) as the query. The server returns the ciphertexts in the interval $[c_1, c_2]$ (if $c_1 \leq c_2$) or $[c_1, N] \cup [1, c_2]$ (if $c_1 > c_2$).

MOPE Security and Random MOPF. In order to define the security of an MOPE scheme, we introduce a generalization of OPFs. For $j \in [M]$, let $\phi_j : [M] \rightarrow$

$[M]$ be the cyclic transformation $\phi_j(x) = (x - j - 1) \bmod M + 1$. We define the set of *modular order preserving functions* from $[M]$ to $[N]$ as

$$\mathsf{MOPF}_{[M],[N]} = \{f \circ \phi_j \mid f \in \mathsf{OPF}_{[M],[N]}, j \in [M]\} .$$

Note that all OPFs are MOPFs; on the other hand, most MOPFs are not OPFs. However, a MOPF g is "modular order-preserving" in that the function $g - g(0) \bmod N$ is order-preserving.

Now, define $\mathsf{RMOPF}_{[M],[N]} = (\mathcal{K}_{\mathrm{rm}}, \mathcal{E}nc_{\mathrm{rm}}, \mathcal{D}ec_{\mathrm{rm}})$, the *random modular order-preserving function* scheme, as the following (inefficient) encryption scheme:

- $\mathcal{K}_{\mathrm{rm}}$ returns a random instance g of $\mathsf{MOPF}_{[M][N]}$.
- $\mathcal{E}nc_{\mathrm{rm}}$ takes the key g and a plaintext m to return $g(m)$.
- $\mathcal{D}ec_{\mathrm{rm}}$ takes the key g and a ciphertext c to return $g^{-1}(c)$.

Note that an MOPF could alternatively be defined with a random ciphertext shift following the OPF rather than a random plaintext shift preceding it. The advantage of the above definition is that the map from (OPF, ciphertext offset) pairs to MOPFs is bijective whereas in the alternative it is not one-to-one.

We now are ready to define MOPE security. Fix an MOPE scheme $\mathcal{SE}_{[M],[N]} = (\mathcal{K}_{\mathrm{m}}, \mathcal{E}nc_{\mathrm{m}}, \mathcal{D}ec_{\mathrm{m}})$. Let $\mathsf{RMOPF}_{[M],[N]} = (\mathcal{K}_{\mathrm{rm}}, \mathcal{E}nc_{\mathrm{rm}}, \mathcal{D}ec_{\mathrm{rm}})$ be as defined above. For an adversary A, define its *pseudorandom modular order-preserving function (PMOPF) advantage* against $\mathcal{SE}$ as

$$\mathbf{Adv}^{\mathrm{pmopf}}_{\mathcal{SE}}(A) = \Pr_{K \xleftarrow{\$} \mathcal{K}_{\mathrm{m}}}\left[\, A^{\mathcal{E}nc_{\mathrm{m}}(K,\cdot)} = 1 \,\right] - \Pr_{g \xleftarrow{\$} \mathsf{RMOPF}_{[M],[N]}}\left[\, A^{g(\cdot)} = 1 \,\right] .$$

It is straightforward to show that the MOPE scheme obtained from any POPF-secure OPE scheme via the transformation defined in the beginning of Section 4.5.2 is PMOPF-secure, under the same assumption as the base scheme. We omit the details.

We now analyze the ideal object, RMOPF, under the one-wayness definitions.

WINDOW ONE-WAYNESS OF RMOPF. The following proposition establishes that RMOPF is optimally r, z-window one-way (and hence optimally one-way, taking $r = 1$) in the sense that an adversary cannot do better than an adversary that outputs a random window independent of the challenge set. (Reminder: "window" includes windows that wrap around the edge of the space.)

Proposition 4.5.2. *Fix any window size r and challenge set size z. Let $A_{\mathrm{rand}}(r)$ be an r, z-WOW adversary that, on any input, outputs a random r-window from $[M]$. Then for any adversary A,*

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{RMOPF}_{[M],[N]}}(A) \leq \mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{RMOPF}_{[M],[N]}}(A_{\mathrm{rand}}(r)) \leq rz/M \ .$$

The proof is in Appendix B.8.

As one might surmise, the above "optimal" characterization of the one-wayness of a random MOPF fails to show a complete picture of the information a random MOPF leaks. To investigate further, we turn to distance one-wayness.

WDOW ADVANTAGE BOUNDS FOR RMOPF. We claim that the distance one-wayness analysis for RMOPF is exactly the same as for ROPF. To see this, consider the following proposition.

Proposition 4.5.3. *Let $c_1, c_2 \in [N]$. Then for any $d \in \{0, \ldots, M-1\}$,*

$$\Pr_{K \xleftarrow{\$} \mathcal{K}_{\mathrm{r}}}[\,\mathcal{D}ec_{\mathrm{r}}(K, c_2) - \mathcal{D}ec_{\mathrm{r}}(K, c_1) = d\,] = \Pr_{K \xleftarrow{\$} \mathcal{K}_{\mathrm{rm}}}[\,\mathcal{D}ec_{\mathrm{rm}}(K, c_2) - \mathcal{D}ec_{\mathrm{rm}}(K, c_1) = d\,].$$

Proof. Let $w = c_2 - c_1 \bmod N$. Note that among the $\binom{N-2}{M-2}$ OPFs f with $c_1, c_2 \in f([M])$, there are $\binom{w-1}{d-1}\binom{N-w-1}{M-d-1}$ such that $f^{-1}(c_2) - f^{-1}(c_1) \bmod M = d$. On the other hand, among the $\binom{N-2}{M-2} \cdot M$ MOPFs g with $c_1, c_2 \in g([M])$, there are $\binom{w-1}{d-1}\binom{N-w-1}{M-d-1} \cdot M$ such that $g^{-1}(c_2) - g^{-1}(c_1) \bmod M = d$. The result follows. □

Therefore, the $1, z$-WDOW advantage upper bound of Theorem 4.4.3 and the r, z-WDOW advantage lower bound of Theorem 4.4.4 against ROPF schemes also apply to RMOPF schemes on the same parameters.

So, while an RMOPF has similar security to that of an ROPF for distance and window distance one-wayness, it is better in terms of one-wayness and window one-wayness. The analysis easily transfers to any secure MOPE scheme. We now discuss a few supplemental security considerations for RMOPF schemes.

Effect of a Known-Plaintext Attack on RMOPF. In the $\mathsf{RMOPF}_{[M],[N]}$ scheme, if the adversary learns a single plaintext-ciphertext pair, then the one-wayness analysis reduces to that of $\mathsf{ROPF}_{[M-1],[N-1]}$. To see this, note that if g is a random function in $\mathsf{MOPF}_{[M],[N]}$, and it is revealed that $g(m_0) = c_0$, then $f(m) = g(m + m_0 \bmod M) - c_0 \bmod N$ is a random function in $\mathsf{OPF}_{[M-1],[N-1]}$.

On Implementing a Scheme to Support Range Queries using PMOPF. We note that when a pseudorandom MOPF scheme is used to implement a range-query-supporting database, even wrap-around target range queries must be made, for otherwise an adversary may infer the secret offset of the MOPF scheme after observing many non-wrap-around target queries.

Remark. We finally note that the tagging scheme $\mathcal{CEOE}$ defined in Section 4.5.1 could be similarly modified so that its tag receives a secret offset. The resulting scheme would support modular range queries in the predetermined static database scenario, and satisfy a stronger version of IND-CCPA, leaking only "modular" order.

CHAPTER V

EFFICIENT FUZZY-SEARCHABLE ENCRYPTION

We now consider the problem of efficient (sub-linear) search on updatable encrypted data that supports error-tolerant search queries, that is, efficient fuzzy-searchable encryption (EFSE). As explained in Section 1.5.3, this is a highly practical but relatively unexplored topic in ESE that we are the first to study in a cryptographic context, using provable security. We first give an overview of the results.

5.1 Overview

Defining closeness. To even define our problem, we first need to establish what "close" means for messages; and specifically, define the "closeness" that we would like ciphertexts to reveal. At its core, closeness is a function assigning a value (say, "close" or "far") to any pair of messages from a space. Thus, we introduce the concept of a *closeness domain* which consists of a domain along with a closeness function.[1]

Efficiently fuzzy-searchable encryption and its security. Next we define the central primitive, *efficiently fuzzy-searchable encryption (EFSE)*, defined on a closeness domain. In addition to the standard functions of a symmetric encryption scheme, an EFSE scheme should provide a public function $\mathrm{Cl}_{\mathcal{R}}$ on pairs of ciphertexts that reveals whether those ciphertexts correspond to equal, close or far messages. For EFSE, there should also exist an associated data structure supporting sub-linear search. We then discuss the details of how a user and the server perform search using

[1] One might compare the closeness domain primitive to that of a metric space. Neither is a generalization of the other: unlike a metric, a closeness function may only take on a few values; while a metric must satisfy the triangle inequality, which is not necessary for closeness functions. However, the two primitives are related, and as we explain later, a closeness domain can be defined in terms of a metric space along with numerical thresholds.

an EFSE. We note that an EFSE scheme leaks equality and "closeness" of messages in order to provide efficient exact-match and fuzzy search.

Next we define a security notion for FSE that we call *indistinguishability under same-closeness-pattern chosen-plaintext attack* or *IND-CLS-CPA*. The definition is a natural relaxation of the standard IND-CPA security definition that prohibits queries leading to trivial attacks.

Template EFSE construction. We present all EFSE constructions via a general template consisting of several components, the most important of which we call a *closeness-preserving bucketing function* (CPBF). A CPBF maps domain elements to "buckets" (arbitrary objects) so that close messages map to overlapping buckets, and far messages do not. Besides this, the construction makes use of an *efficient searchable encryption* (ESE) [3] scheme, which is essentially an encryption scheme leaking equality; and a collision-free *batch-tagging family*, each instance of which is a deterministic function from the domain to the range. We define a notion of security for a batch-tagging family which together with IND-DCPA-security [11] of the ESE scheme is sufficient to prove security of our template construction in the case that the CPBF is *consistent*, a quality we define. We also show how to create a secure collision-free batch-tagging family out of a blockcipher, and reference [3] for blockcipher-based ESE constructions that are IND-DCPA-secure. So the missing component we need for a secure scheme is a consistent CPBF.

New optimally-secure construction. We propose a new general EFSE scheme. It relies on the notion of a *closeness graph*, whose vertices are the unique elements of the message space, and edges indicate closeness between elements. Our construction defines a bucketing function that essentially sends a message to its incident edges in the closeness graph. This is a consistent CPBF and thus the associated EFSE scheme is, optimally, IND-CLS-CPA-secure.

One might worry that our construction is rather inefficient in terms of the ciphertext length, which is linear in the maximum degree of the closeness graph. However we show that in fact closeness on a *rigid* (i.e., every message pair is close or far) domain may be defined so that any EFSE scheme requires ciphertext length linear in the maximum degree of the closeness graph. The argument is information theoretic and relies on the functionality, rather than security, of the primitive. Thus, in achieving EFSE on arbitrarily-defined closeness domains the new IND-CLS-CPA-secure construction is (asymptotically) space-optimal, and moreover optimally secure.

Analysis of scheme from [48]. We also analyze security of the scheme from [48]. The scheme can be roughly translated into a scheme fitting our template construction, where the bucketing function sends a message to itself and all of its neighbors in the closeness graph. Unfortunately, this CPBF is not consistent, and we show that this is enough to guarantee IND-CLS-CPA-insecurity of the scheme. Intuitively, the attack exploits a simple observation that some close messages may have more common neighbors than others, and this is revealed in ciphertexts. Leaking such information is not required for the functionality of EFSE and hence is a security breach according to our definition. We also note that the scheme from [48] is not only insecure but is as space-inefficient as our optimally-secure edge-tagging scheme, so our construction is clearly an improvement.

Constructions with improved efficiency. In many (even most?) practical applications, vertices of the closeness graph have massive degrees. (Degrees can even be infinite, e.g. on continuous spaces.) This can happen particularly for multi-dimensional spaces, as the number of "close neighbors" increases exponentially with dimension for closeness defined on a metric. In such situations our optimally-secure scheme, as well as the insecure scheme from [48], are unacceptably inefficient.

The aforementioned lower bound result shows that we cannot expect to do better for arbitrary rigid domains. We seek the right balance between the desired efficiency

and security of EFSE, and look at non-rigid domains. We argue that IND-CLS-CPA-security is too strong to be useful in characterizing EFSEs on non-rigid closeness domains (where near messages could be encrypted to either close or far ciphertexts), and so to do this we introduce a new security definition. The new definition requires schemes to hide all information about plaintexts except nearness and a certain aspect of "local structure"—which can be intuitively understood as the least significant bits of messages corresponding to nearness clusters of known ciphertexts. Importantly, this implies that no major relative information is leaked about a pair of "disconnected messages," that is, messages that cannot be connected through a chain of near known corresponding ciphertext pairs.

Our definition and constructions focus on a practical choice of domains with associated metric along with "close" and "far" distance thresholds, that we call *metric closeness domains*. In particular we focus on the Euclidean metric on arbitrary-dimension real domains. For that, we fix a regular multi-dimensional lattice, whose short basis is assumed to be public. Before getting to specific schemes, we introduce the concept of an "anchor radius" for a metric closeness domain and a lattice, and use it to construct a bucketing function to build a EFSE via our general bucketing template. We show that a valid anchor radius then implies an EFSE construction that is secure according to our weaker notion.

Next, we pose a general problem of trying to improve space-efficiency and flexibility of such EFSE schemes by choosing lattices and anchor radii wisely. We then introduce several specific schemes on particular closeness domains. Finally, we explain how to build an EFSE scheme on a joint closeness domain (a product of small-dimension closeness domains, with closeness defined conjunctively) as might be useful in biometric-data-matching applications.

We leave it as an open problem to extend our results to the public-key setting or show that this is not possible.

5.2 Closeness Primitives

In order to study schemes preserving closeness, our first task is to develop a primitive establishing the concept of "closeness" on a message space. To this end, we define a "closeness function" Cl on 2-element subsets of the domain (we will say "pairs," meaning unordered pairs). Conceptually, the output of Cl, describing the closeness of a pair, could be defined in two disparate ways, as follows.

- Qualitative closeness: $\mathrm{Cl}(\{\cdot,\cdot\})$ takes only a few values, e.g. "close" vs. "far".
- Quantitative closeness: $\mathrm{Cl}(\{\cdot,\cdot\})$ takes on a numerical value in some (discrete or continuous) range.

Notice that quantitative closeness may be somewhat ill-suited for use with EFSE—if a scheme must leak quantitative information about closeness, that leaves very little interesting information left for encryption to protect! Thus, we focus on qualitative closeness, hereafter called just "closeness." This leads to the following primitives.

Closeness domain. We refer to the pair $\Lambda = (\mathcal{D}, \mathrm{Cl})$ as a *closeness domain* if

1. $\mathcal{D}$ is a (finite or infinite) set, called the *domain* or *message space*;
2. Cl is the *closeness function* that takes 2-element subsets of $\mathcal{D}$ and outputs a member of $\{\texttt{close}, \texttt{near}, \texttt{far}\}$.

We abuse notation: for $m \neq m' \in \mathcal{D}$, we write $\mathrm{Cl}(m, m')$ or $\mathrm{Cl}(m', m)$ as shorthand for $\mathrm{Cl}(\{m, m'\})$. For $m, m' \in \mathcal{D}$, if $\mathrm{Cl}(m, m') = \texttt{close} \mid \texttt{near} \mid \texttt{far}$ then we say m and m' are $close \mid near \mid far$, respectively. For convenience, we say Λ is *rigid* if $\mathrm{Cl}(m, m') \in \{\texttt{close}, \texttt{far}\}$ for all $m \neq m' \in \mathcal{D}$, and *flexible* if $\mathrm{Cl}(m, m') \in \{\texttt{close}, \texttt{near}\}$ for all $m \neq m' \in \mathcal{D}$—we will see the importance of this distinction in the next section.

Let d be a metric on domain $\mathcal{D}$,and let $\delta^{\mathtt{F}} \geq \delta^{\mathtt{C}} > 0$. The *metric* closeness domain $\left(\mathcal{D}, \mathcal{M}_d^{\delta^{\mathtt{C}}, \delta^{\mathtt{F}}}\right)$ on domain $\mathcal{D}$ with respect to metric d, far threshold $\delta^{\mathtt{F}}$, and

close threshold $\delta^{\mathtt{C}}$ is defined as

$$\mathcal{M}_d^{\delta^{\mathtt{C}},\delta^{\mathtt{F}}} = \begin{cases} \mathtt{close} & \text{if } d(m_1,m_2) \in (0,\delta^{\mathtt{C}}] \\ \mathtt{near} & \text{if } d(m_1,m_2) \in (\delta^{\mathtt{C}},\delta^{\mathtt{F}}] \\ \mathtt{far} & \text{if } d(m_1,m_2) > \delta^{\mathtt{F}} \end{cases}$$

For instance, $\left(\{0,1\}^{80}, \mathcal{M}_{\mathrm{Ham}}^{1,2}\right)$, where Ham is hamming distance, is a closeness domain of all length-80 strings where strings differing in 1 bit are close, differing in 2 bits are near, and differing in more than 2 bits are far.

Closeness and nearness graph. Let $\Lambda = (\mathcal{D}, \mathrm{Cl})$ be a closeness domain. Let $\mathcal{V}_\Lambda = \mathcal{D}$ and

$$\mathcal{E}_\Lambda^{\mathtt{C}} = \{\{u,v\} \mid u \neq v \in \mathcal{V}_\Lambda \text{ and } \mathrm{Cl}(u,v) = \mathtt{close}\};$$

$$\mathcal{E}_\Lambda^{\mathtt{N}} = \{\{u,v\} \mid u \neq v \in \mathcal{V}_\Lambda \text{ and } \mathrm{Cl}(u,v) \in \{\mathtt{close},\mathtt{near}\}\}.$$

Then $\mathcal{G}_\Lambda^{\mathtt{C}} = (\mathcal{D}, \mathcal{E}_\Lambda^{\mathtt{C}})$ is the *closeness graph* and $\mathcal{G}_\Lambda^{\mathtt{N}} = (\mathcal{D}, \mathcal{E}_\Lambda^{\mathtt{N}})$ the *nearness graph* of Λ.

Induced subgraph. In general, for graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and $H \subseteq \mathcal{V}$ let $\mathcal{G}(H) = (H, \mathcal{E}(H))$ be the *subgraph induced by* H where $\mathcal{E}(H) = \{\{u,v\} \in \mathcal{E} \mid u,v \in H\}$.

5.3 Efficiently Fuzzy-Searchable Symmetric Encryption

We now define our main primitive and show how can it be used for efficient search. Following that, we formulate the ideal level of security for EFSE schemes.

5.3.1 Defining Efficiently Fuzzy-Searchable Encryption

$\mathtt{FSE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}}, \mathtt{makeDS}, \mathtt{fuzzyQ})$ is a *structured fuzzy-searchable symmetric encryption* (StructFSE) scheme on closeness domain $\Lambda_{\mathcal{D}} = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$ if

- $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ is a symmetric encryption scheme on $\mathcal{D}$;
- $\mathrm{Cl}_{\mathcal{R}}$ is a function that takes two inputs and returns `close`, `far`, or `eq`. We

require that for any key K generated by $\mathcal{K}$ and $m_1, m_2 \in \mathcal{D}$,

$$\mathrm{Cl}_{\mathcal{R}}(\mathcal{E}nc(K, m_1), \mathcal{E}nc(K, m_2)) = \begin{cases} \texttt{close} & \text{if } \mathrm{Cl}_D(m_1, m_2) = \texttt{close} \\ \texttt{far} & \text{if } \mathrm{Cl}_D(m_1, m_2) = \texttt{far} \\ \texttt{eq} & \text{if } m_1 = m_2 \end{cases}$$

- `makeDS` takes a set of ciphertexts $\mathbf{C}$ (the *database*) and outputs a data structure $\texttt{DS}_{\mathbf{C}}$;
- `fuzzyQ`, given database $\mathbf{C}$ and data structure $\texttt{DS}_{\mathbf{C}}$, takes query ciphertext c, and outputs the *target* of c in $\mathbf{C}$, $\texttt{Tgt}_{\mathbf{C}}(c) = \{c' \in \mathbf{C} \mid \mathrm{Cl}_{\mathcal{R}}(c, c') \in \{\texttt{close}, \texttt{eq}\}\}$.

Notice that $\mathrm{Cl}_{\mathcal{R}}$ reveals whether ciphertexts correspond to close messages or far messages, but it may have either behavior on encryptions of near messages. Near message pairs can thus be thought of as "false positive candidates" in a fuzzy search query—as an encryption of a near message can be (but does not have to be) in the target of a fuzzy query. In this sense, FSE on a rigid closeness domain cannot have any false positives, and FSE on a flexible closeness domain can have any number of false positives.

Now, we say StructFSE scheme $\texttt{FSE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}}, \texttt{makeDS}, \texttt{fuzzyQ})$ is an *efficient fuzzy searchable symmetric encryption* (EFSE) scheme if for any (sufficiently large) database $\mathbf{C}$, data structure $\texttt{DS}_{\mathbf{C}}$, key K generated by $\mathcal{K}$, and query ciphertext c with $|\texttt{Tgt}_{\mathbf{C}}(c)|$ sub-linear in the size of $\mathbf{C}$, the running time of $\texttt{fuzzyQ}_{\mathbf{C},\texttt{DS}_{\mathbf{C}}}(c)$ is sub-linear in the size of $\mathbf{C}$.

Using an EFSE scheme. Let $\texttt{FSE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}}, \texttt{makeDS}, \texttt{fuzzyQ})$ be an EFSE scheme and K a valid key. In the DBaaS application (see Section 1.1), let $\mathbf{C}$ be the set of ciphertexts currently in the encrypted database, encrypted under K. The server runs $\texttt{makeDS}(\mathbf{C})$ to create a data structure $\texttt{DS}_{\mathbf{C}}$, and upon a new query $c = \mathcal{E}nc_K(m)$, runs $\texttt{fuzzyQ}(\mathbf{C}, \texttt{DS}_{\mathbf{C}}, c)$ and returns the result, $\texttt{Tgt}_{\mathbf{C}}(c)$, to the user.

By correctness of the scheme, this response will contain all ciphertexts in $\mathbf{C}$ whose messages are close to m and no ciphertexts whose messages are far from m. Since the scheme is efficient, such a query will take time sub-linear in the size of the database $\mathbf{C}$ (assuming the appropriate response itself is also sub-linear in the size of $\mathbf{C}$.) Also note that leaked equality permits efficient exact-match search.

As a side note, in a practical implementation, additional functions (add, remove, edit, etc.) would be useful to efficiently update the data structure as the database changes. In our analysis, we are less focused on efficiency of the data structure maintenance, so for simplicity we just let the (possibly inefficient) function `makeDS` construct the data structure from the entire database.

Finally, observe that the "difficult" part of building an EFSE scheme is ensuring that `fuzzyQ` is efficient. Thus, constructions of $\mathcal{E}nc$ and $\mathrm{Cl}_{\mathcal{R}}$ might as well be designed with the efficiency of `fuzzyQ` in mind. In our constructions, as detailed in Section 5.4, ciphertexts outputted by $\mathcal{E}nc$ will contain "tags" such that ciphertexts of close messages share a common tag. Thus, indexing ciphertexts by tags in an efficiently searchable data structure, like a binary search tree, leads to an efficient construction of `fuzzyQ`.

5.3.2 Ideal security for EFSE schemes

We construct the following indistinguishability-based security definition, called IND-CLS-CPA, for analyzing the security of EFSE schemes. Intuitively, this notion is identical to IND-CPA with the additional condition that left-right queries have the same *closeness pattern* (in the second requirement below.) Notice that we do not study chosen-ciphertext security here as it can be achieved using the encrypt-then-MAC method [6].

Let `FSE` be an EFSE scheme on closeness domain $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$. For $b \in \{0, 1\}$ and adversary A, let $\mathbf{Exp}^{\text{ind-cls-cpa-}b}_{\texttt{FSE}}(A)$ be the IND-CPA experiment $\mathbf{Exp}^{\text{ind-cpa-}b}_{\texttt{FSE}}(A)$ in

Figure 1 but with the following restrictions: if $(m_0^1, m_1^1), \ldots, (m_0^q, m_1^q)$ are the queries A makes to its LR encryption oracle $\mathcal{E}nc(K, \mathcal{LR}(\cdot,\cdot,b))$,

1. $|m_0^i| = |m_1^i|$ for all $i \in [q]$;
2. for all $i, j \in [q]$, either $\begin{cases} \mathrm{Cl}_{\mathcal{D}}(m_0^i, m_0^j) = \mathrm{Cl}_{\mathcal{D}}(m_1^i, m_1^j), & \text{or} \\ m_0^i = m_0^j \text{ and } m_1^i = m_1^j. \end{cases}$

For an adversary A, define its *IND-CLS-CPA advantage* against FSE as

$$\mathbf{Adv}^{\text{ind-cls-cpa}}_{\mathtt{FSE}}(A) \;=\; \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-0}}_{\mathtt{FSE}}(A) = 1\right].$$

We say that FSE is *indistinguishable under same-closeness-pattern chosen-plaintext attack* (IND-CLS-CPA-secure) if the IND-CLS-CPA advantage of any adversary against FSE is small.

It should be apparent that IND-CLS-CPA-security is optimal for EFSE schemes: revealing equality and closeness patterns of left/right queries is unavoidable as the (public) $\mathrm{Cl}_{\mathcal{R}}$ function leaks closeness and equality. Thus, to avoid the naive attack, we must outlaw queries that exploit this leakage.

5.4 Template Bucket-Tagging Construction for EFSE

In this somewhat technical section, we build up to a general construction of an EFSE scheme given a valid "bucketing function" on the desired closeness domain. In addition, we show that under certain conditions, the scheme is IND-CLS-CPA-secure. First, though, we define several primitives, along with relevant security notions, that will be components of the construction. The primitives are: efficient searchable encryption (ESE) schemes [3], privacy-preserving batch-tagging families, and closeness-preserving bucketing functions.

5.4.1 Efficient searchable encryption and security

The ESE scheme primitive, which was not formalized earlier, is formally defined in Appendix C.1. Intuitively, an ESE is like a standard encryption scheme except

that it "leaks equality," that is, there is a (public) way to tell if two ciphertexts are encryptions of the same message. The appropriate security notion for ESE was defined by [11] and is called *indistinguishability under distinct chosen plaintext attack* (IND-DCPA)—it is also recalled in Appendix C.1. The notion is identical to IND-CPA except that left/right-queries must have the same "equality pattern" (and so avoiding the obvious attack, as ESE leaks equality.) For blockcipher-based IND-DCPA-secure constructions of ESE schemes, see [3].

5.4.2 Privacy-preserving batch-tagging

We say that $\mathcal{F}_{\mathtt{Tag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T})$ is a *tagging family* on domain $\mathcal{D}$ and range $\mathcal{R}$ if $\mathcal{K}_{\mathcal{T}}$ outputs random keys and $\mathcal{T}$ takes a key and an element of $\mathcal{D}$ and outputs an element of $\mathcal{R}$ such that $\mathcal{T}(K_{\mathcal{T}}, \cdot)$ is a (deterministic) function from $\mathcal{D}$ to $\mathcal{R}$. We further say that $\mathcal{F}_{\mathtt{BTag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ is a *batch-tagging family* if $(\mathcal{K}_{\mathcal{T}}, \mathcal{T})$ is a tagging family and $\mathcal{B}$ takes a key $K_{\mathcal{T}}$ and a set of elements $M \subseteq \mathcal{D}$ and outputs $\{\mathcal{T}(K_{\mathcal{T}}, m) \mid m \in M\}$.

Given a tagging family $(\mathcal{K}'_{\mathcal{T}}, \mathcal{T}')$ it is easy to construct a batch-tagging family $(\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$: let $\mathcal{K}_{\mathcal{T}} = \mathcal{K}'_{\mathcal{T}}$ and $\mathcal{T} = \mathcal{T}'$, and define $\mathcal{B}(K_{\mathcal{T}}, \cdot)$ to take a set of messages, run $\mathcal{T}(K_{\mathcal{T}}, \cdot)$ on each, and return the set of results.

We say that a tagging family $(\mathcal{K}_{\mathcal{T}}, \mathcal{T})$ or a batch-tagging family $(\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ is *collision-free* if for any key $K_{\mathcal{T}}$, $\mathcal{T}(K_{\mathcal{T}}, \cdot)$ is one-to-one on $\mathcal{D}$.

Now, we introduce a security definition for batch-tagging families. Called *privacy-preserving under chosen batch-tag attack*, it is essentially the privacy-preserving notion from [11] generalized to objects of the batch-tagging primitive.

Let $\mathcal{F}_{\mathtt{BTag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ be a batch-tagging family on domain $\mathcal{D}$ and range $\mathcal{R}$. For an adversary A and $b \in \{0, 1\}$ consider the experiment defined in Figure 9, where it is required that, if $(M_0^1, M_1^1), \ldots, (M_0^q, M_1^q)$ are the queries that A makes to its $\mathcal{LR}$-batch-tagging oracle (note: each M_j^i is a *set* of elements of $\mathcal{D}$), for all $I \subseteq [q]$ we have $\left|\bigcap_{i \in I} M_0^i\right| = \left|\bigcap_{i \in I} M_1^i\right|$.

$$\frac{\textbf{Experiment } \textbf{Exp}^{\text{pp-cbt-}b}_{\mathcal{F}_{\texttt{BTag}}}(A)}{\begin{array}{l} K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}} \\ b' \xleftarrow{\$} A^{\mathcal{B}(K_{\mathcal{T}},\mathcal{LR}(\cdot,\cdot,b))} \\ \text{Return } b' \text{ ,} \end{array}}$$

Figure 9: The PP-CBT experiment.

For an adversary A, define its *PP-CBT advantage* against $\mathcal{F}_{\texttt{BTag}}$ as

$$\mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\texttt{BTag}}}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\texttt{BTag}}}(A) = 1\,\right] - \Pr\left[\,\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\texttt{BTag}}}(A) = 1\,\right].$$

We say that $\mathcal{F}_{\texttt{BTag}}$ is *privacy-preserving under chosen batch-tag attack* (PP-CBT-secure) if the PP-CBT advantage of any adversary against $\mathcal{F}_{\texttt{BTag}}$ is small.

Notice that the requirement rules out an obvious attack: suppose to the contrary that, without loss of generality, the adversary could query $(M_0^1, M_1^1), \ldots, (M_0^q, M_1^q)$ with $\left|\bigcap_{i\in[q]} M_0^i\right| > \left|\bigcap_{i\in[q]} M_1^i\right|$. If $\mathcal{T}(K_{\mathcal{T}}, \cdot)$ is collision-free, $\left|\bigcap_{i\in[q]} \mathcal{B}(K_{\mathcal{T}}, M_b^i)\right| = \left|\bigcap_{i\in[q]} \{\mathcal{T}(K_{\mathcal{T}}, m) \mid m \in M_b^i\}\right| = \left|\bigcap_{i\in[q]} M_b^i\right|$, so by computing $\left|\bigcap_{i\in[q]} \mathcal{B}(K_{\mathcal{T}}, M_b^i)\right|$ from the oracle responses the adversary can identify b.

Instantiating a secure, collision-free batch-tagging scheme. Naturally, we will want a PP-CBT-secure scheme for our constructions in this paper, so how can we construct one? In fact, a PP-CBT-secure batch-tagging scheme can be created straightforwardly out of a pseudorandom permutation (PRF). We explain the batch-tagging PRF-based construction in Appendix C.2.

However, as will soon become clear, what we actually need is a PP-CBT-secure *collision-free* batch-tagging scheme. Note that if we instantiate the above PRF-based batch-tagging scheme with a pseudorandom permutation (PRP) it will be both collision-free and PP-CBT-secure. Any blockcipher then will work to instantiate the PRP, possibly by augmenting the blockcipher into a variable-input-length blockcipher [12] as described in [52]. The details are left to the reader.

5.4.3 Closeness-preserving bucketing functions

Fix a closeness domain $\Lambda = (\mathcal{D}, \mathrm{Cl})$. Let B be a (finite or infinite) set and let $\mathtt{Bkts} : \mathcal{D} \to 2^B$ be a function assigning a subset of B to every domain element. We call $\mathtt{Bkts}$ a *closeness-preserving bucketing function* (CPBF) from Λ into B if for every $x, y \in \mathcal{D}$ with $\mathrm{Cl}(x, y) = \mathtt{close}$, there exists $b \in B$ such that $b \in \mathtt{Bkts}(x) \cap \mathtt{Bkts}(y)$; and for every $x, y \in \mathcal{D}$ with $\mathrm{Cl}(x, y) = \mathtt{far}$, $\mathtt{Bkts}(x) \cap \mathtt{Bkts}(y) = \emptyset$.

Further, a CPBF $\mathtt{Bkts}$ is *consistent* with respect to closeness domain Λ if for any message sets $\{m_0^1, \ldots, m_0^q\}$ and $\{m_1^1, \ldots, m_1^q\}$ having the same closeness pattern[2], we have $\left|\bigcap_{i\in[q]} \mathtt{Bkts}(m_0^i)\right| = \left|\bigcap_{i\in[q]} \mathtt{Bkts}(m_1^i)\right|$. Consistency can be understood intuitively as follows: whenever a set of messages has the same closeness pattern as another set of messages, each set should be found all together in the same number of buckets.

Examples of CPBFs are integral to our constructions and are several are introduced in the remainder of this paper.

5.4.4 Template bucket-tagging EFSE construction

We now provide a general "template" construction for an EFSE scheme given a closeness-preserving bucketing function $\mathtt{Bkts}$, batch-tagging family $\mathcal{F}_{\mathtt{BTag}}$, and ESE scheme $\mathtt{ESE}$. We remark that this template is a generalization of the technique used in [48], though we have expanded, formalized, and refined it significantly. All forthcoming EFSE constructions in this paper use this general construction as a template.

Let $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$ be a closeness domain, $\mathtt{Bkts}$ a function from Λ into a set B, $\mathcal{F}_{\mathtt{BTag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ a batch-tagging family on domain $\mathcal{D}_{\mathcal{T}} = B$ and range $\mathcal{R}_{\mathcal{T}}$, and $\mathtt{ESE} = (\mathcal{K}_{\mathtt{ESE}}, \mathcal{E}nc_{\mathtt{ESE}}, \mathcal{D}ec_{\mathtt{ESE}}, F, G)$ an ESE scheme on $\mathcal{D}$. Then we define a general *bucket-tagging* StructFSE scheme $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ in Figure 10.

Correctness. The following result establishes that the template construction is a

[2]That is, either $\mathrm{Cl}_{\mathcal{D}}(m_0^i, m_0^j) = \mathrm{Cl}_{\mathcal{D}}(m_1^i, m_1^j)$ or $(m_0^i = m_0^j$ and $m_0^i = m_0^j)$ for all $i, j \in [q]$.

$\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}] = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}}, \mathtt{makeDS}, \mathtt{fuzzyQ})$ where

- $\mathcal{K}$ runs $K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}}$ and $K_{\mathtt{ESE}} \xleftarrow{\$} \mathcal{K}_{\mathtt{ESE}}$, and returns $K_{\mathcal{T}} \| K_{\mathtt{ESE}}$.
- $\mathcal{E}nc(K_{\mathcal{T}} \| K_{\mathtt{ESE}}, m)$ runs $B_m \leftarrow \mathtt{Bkts}(m)$; $\mathtt{tags} \leftarrow \mathcal{B}(K_{\mathcal{T}}, B_m)$; $c_R \leftarrow \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, m)$, and returns $c \leftarrow \mathtt{tags} \| c_R$.
- $\mathcal{D}ec(K_{\mathcal{T}} \| K_{\mathtt{ESE}}, c)$ parses c as $\mathtt{tags} \| c_R$ and returns $\mathcal{D}ec_{\mathtt{ESE}}(K_{\mathtt{ESE}}, c_R)$.
- $\mathrm{Cl}_{\mathcal{R}}(c, c')$ parses $c = \mathtt{tags} \| c_R$ and $c' = \mathtt{tags}' \| c'_R$, and returns `eq` if $G(c_R) = G(c'_R)$; (otherwise) `close` if $\mathtt{tags} \cap \mathtt{tags}' \neq \emptyset$; otherwise `far`.
- $\mathtt{makeDS}(\mathbf{C})$ initializes an efficient self-balancing search tree T representing an associative array from elements of $\mathcal{R}_{\mathcal{T}}$ to ciphertexts. For each ciphertext $c \in \mathbf{C}$ parsed as $c = \mathtt{tags} \| c_R$, and for each $t \in \mathtt{tags}$, add the node $(t \mapsto c)$ to T. Output $\mathtt{DS}_{\mathbf{C}} \leftarrow T$.
- $\mathtt{fuzzyQ}_{\mathbf{C}, \mathtt{DS}_{\mathbf{C}}}(c)$ parses c as $\mathtt{tags} \| c_R$ and interprets $\mathtt{DS}_{\mathbf{C}}$ as search tree T. Let $Q = \emptyset$. For each $t \in \mathtt{tags}$, search T for elements indexed by t; for any $(t \mapsto c')$ that exist, add c' to Q. Return $\mathtt{nbsQ}^{c}_{\mathbf{C}} \leftarrow Q$.

Figure 10: General bucket-tagging construction of a StructFSE scheme given $\mathtt{Bkts}$, $\mathcal{F}_{\mathtt{BTag}}$, $\mathtt{ESE}$.

valid StructFSE scheme as long as bucketing function $\mathtt{Bkts}$ is closeness-preserving and batch-tagging function $\mathcal{F}_{\mathtt{BTag}}$ is collision-free.

Proposition 5.4.1. *If $\mathcal{F}_{\mathtt{BTag}}$ is collision-free and $\mathtt{Bkts}$ is closeness-preserving, then $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ is a StructFSE scheme on Λ.*

Proof. We first show $\mathrm{Cl}_{\mathcal{R}}$ is correct: for any $m, m' \in \mathcal{D}$ and key $K_{\mathcal{T}} \| K_{\mathtt{ESE}} = K \xleftarrow{\$} \mathcal{K}$,

- $m = m'$ implies $G(c_R) = G(\mathcal{E}nc(K_{\mathtt{ESE}}, m)) = F(K_{\mathtt{ESE}}, m) = F(K_{\mathtt{ESE}}, m') = G(\mathcal{E}nc(K_{\mathtt{ESE}}, m')) = G(c'_R)$, so $\mathrm{Cl}_{\mathcal{R}}(\mathcal{E}nc(K, m), \mathcal{E}nc(K, m'))$ returns `eq`.
- $\mathrm{Cl}_{\mathcal{D}}(m, m') = \mathtt{close}$ implies $\mathtt{tags} \cap \mathtt{tags}' \neq \emptyset$ since $\mathtt{Bkts}$ is a CPBF and $\mathcal{B}(K_{\mathcal{T}}, \cdot)$ is deterministic, so $\mathrm{Cl}_{\mathcal{R}}(\mathcal{E}nc(K, m), \mathcal{E}nc(K, m'))$ returns `close`.
- $\mathrm{Cl}_{\mathcal{D}}(m, m') = \mathtt{far}$ implies $\mathtt{tags} \cap \mathtt{tags}' = \emptyset$ since $\mathtt{Bkts}$ is a CPBF and $\mathcal{B}(K_{\mathcal{T}}, \cdot)$ is collision-free, so $\mathrm{Cl}_{\mathcal{R}}(\mathcal{E}nc(K, m), \mathcal{E}nc(K, m'))$ returns `far`.

Now we can show `fuzzyQ` is correct: it returns all ciphertexts in $\mathbf{C}$ whose set of $\mathcal{B}$-bucket values intersect with that of the query c, which is precisely the set of ciphertexts c' in $\mathbf{C}$ for which either $\mathrm{Cl}_{\mathcal{R}}(c, c') = \texttt{close}$. □

Now, given that the scheme satisfies the conditions of Proposition 5.4.1, it should be clear that $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ is an EFSE as long as $\mu = \max_m |\mathtt{Bkts}(m)|$ is small. To see this, suppose database $\mathbf{C}$ contains k ciphertexts, and assume $k \gg \mu$. Then tree T will have at most $k\mu$ nodes, and a single search for a tag in the tree takes $O(\log(k\mu)) \in O(\log(k))$ time. `fuzzyQ` performs $O(\mu)$ searches on T, so the running time of `fuzzyQ` is $O(\mu \log(k))$, which is sublinear in k.

5.4.5 Conditions for optimal security of the scheme

Now that we have established that the template construction is a valid EFSE scheme given an appropriate CPBF with small $\mu = \max_m |\mathtt{Bkts}(m)|$ and collision-free batch-tagging function, we state conditions under which the construction is IND-CLS-CPA-secure.

In the remainder of this section, fix a closeness domain $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$, and let `Bkts` be a CPBF from Λ into a set B, $\mathcal{F}_{\mathtt{BTag}}$ a collision-free batch-tagging family on B, and `ESE` an ESE scheme on $\mathcal{D}$, so that $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ is a valid StructFSE scheme by Proposition 5.4.1.

The following theorem, proved in Appendix C.3, shows that if `Bkts` is consistent and $\mu = \max_{m \in \mathcal{D}} |\mathtt{Bkts}(m)|$ is small, then IND-CLS-CPA-security of the template scheme depends on PP-CBT-security of $\mathcal{F}_{\mathtt{BTag}}$ and IND-DCPA-security of `ESE`.

Theorem 5.4.2. *If* `Bkts` *is consistent with respect to* Λ*, then for any adversary* A *there exist adversaries* E_A *and* F_A *such that*

$$\mathbf{Adv}^{\text{ind-cls-cpa}}_{\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]}(A) = \mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(E_A) + \mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A).$$

Further, let $\mu = \max_{m \in \mathcal{D}} |\mathtt{Bkts}(m)|$*, and suppose* A *submits* q *length-*2ℓ *queries to its oracle. Then*

- *E_A submits q queries to its oracle, each of length $\leq 4\mu\ell$; q queries to $\mathcal{E}nc_{\texttt{ESE}}$, each of length ℓ; and 2 queries to* `Bkts`*, each of length ℓ;*
- *F_A submits q queries to its oracle, each of length 2ℓ; q queries to $\mathcal{B}$, each of length $\leq 2\mu\ell$; and 1 query to* `Bkts`*, of length ℓ.*

Otherwise, A, E_A, and F_A have the same running time.

5.4.6 A condition for insecurity of the scheme

Finally, we establish that consistency of `Bkts` is necessary to guarantee IND-CLS-CPA-security of the template bucket-tagging scheme. The theorem is proved in Appendix C.4.

Theorem 5.4.3. *Let $\texttt{FSE}_{\texttt{BktTag}}[\texttt{Bkts}, \mathcal{F}_{\texttt{BTag}}, \texttt{ESE}]$ be a valid EFSE defined in the model of Figure 10 on closeness domain Λ. Suppose that CPBF* `Bkts` *is not consistent on closeness domain Λ. Then there exists an adversary submitting q queries to its oracle whose IND-CLS-CPA-advantage against $\texttt{FSE}_{\texttt{BktTag}}[\texttt{Bkts}, \mathcal{F}_{\texttt{BTag}}, \texttt{ESE}]$ is 1.*

Summing up, if CPBF `Bkts` is consistent, batch-tagging oracle $\mathcal{F}_{\texttt{BTag}}$ is PP-CBT-secure *and* collision-free, and ESE scheme `ESE` is IND-DCPA-secure, then we may conclude $\texttt{FSE}_{\texttt{BktTag}}[\texttt{Bkts}, \mathcal{F}_{\texttt{BTag}}, \texttt{ESE}]$ is a valid StructFSE scheme. Furthermore, it is an (optimally) IND-CLS-CPA-secure EFSE if $\mu = \max_m |\texttt{Bkts}(m)|$ is small.

5.5 Toward an Ideally Secure Scheme

We seek an EFSE scheme achieving the ideal level of security, IND-CLS-CPA, as defined in Section 5.3.2. First, we show that the only previously existing candidate is, in general, not IND-CLS-CPA-secure due to Theorem 5.4.3. Then, we construct the first IND-CLS-CPA-secure EFSE scheme using the template from Section 5.4. Finally, we show that in a sense, the space-inefficiency of the secure scheme is necessary to accommodate general closeness domains.

5.5.1 Analysis of an EFSE scheme similar to [48]

The only previously existing EFSE-type scheme is presented in [48]. As noted, the basic structure of our template batch-tagging scheme is a generalization of their method, so it is natural to define a batch-tagging scheme in our model that captures the essence of (and perhaps improves) the [48] scheme. Here we show that this scheme has poor space-efficiency (length of ciphertext linear in the number of close neighbors of a message) and yet fails to achieve IND-CLS-CPA-security. In contrast, the schemes we develop in later sections either achieve IND-CLS-CPA-security, or have much better space-efficiency.

In [48], the authors construct several variants of a fuzzy-searchable scheme; here we present a variant/generalization[3]. Let $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$ be a closeness domain. We define the *neighbor set* of an element m to be $\mathtt{Nb}_m = \{m' \in \mathcal{D} \mid m' \neq m, \mathrm{Cl}_{\mathcal{D}}(m, m') = 1\}$. Let $\mathcal{G}_\Lambda = (\mathcal{V}_\Lambda, \mathcal{E}_\Lambda)$ be the closeness graph of Λ. Define $\mathtt{BktNbs} : \mathcal{D} \to \mathcal{V}_\Lambda$ as

$$\mathtt{BktNbs}(m) = \mathtt{Nb}_m \cup \{m\}.$$

Note that if $\mathrm{Cl}_{\mathcal{D}}(m, m') = \mathtt{close}$ then $\mathtt{BktNbs}(m) \cap \mathtt{BktNbs}(m') \supseteq \{m, m'\} \neq \emptyset$, so $\mathtt{BktNbs}$ is a CPBF on Λ. Let $\mathcal{F}_{\mathtt{BTag}}$ be a collision-free batch-tagging family on $\mathcal{V}_\Lambda$ and $\mathtt{ESE}$ an ESE scheme on $\mathcal{D}$, and define $\mathtt{FSEtagNbs}$ to be $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{BktNbs}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ according to Figure 10. If Λ is defined so that $n = \max_{m \in \mathcal{D}} |\mathtt{Nb}_m|$ is small, $\mathtt{FSEtagNbs}$ is an EFSE. However, we see that the ciphertext size is linear in n.

We claim that $\mathtt{FSEtagNbs}$ is IND-CLS-CPA-insecure for the closeness domains considered by [48], as well as most other conceivably useful domains. Suppose, for example, that the closeness domain has two pairs of close messages with different

[3] There are minor differences—notably, $\mathtt{FSEtagNbs}$ uses an IND-DCPA-secure ESE rather than a (stronger) IND-CPA-secure scheme, but this is not an issue as [48] leaks equality already through its tagging strategy. Moreover, we could instantiate $\mathtt{FSEtagNbs}$ with an IND-CPA-secure scheme in place of $\mathtt{ESE}$ and the attack described would still work, since the attack exploits the $\mathcal{F}_{\mathtt{BTag}}$-tagged neighbors, not $\mathtt{ESE}$. Other differences in [48] are inconsequential to the analysis.

numbers of common close neighbors: i.e.,

$$\mathrm{Cl}_{\mathcal{D}}(m_0, m_2) = \mathrm{Cl}_{\mathcal{D}}(m_1, m_2) = \texttt{close}\,; \qquad |\texttt{Nb}_{m_0} \cap \texttt{Nb}_{m_2}| \neq |\texttt{Nb}_{m_1} \cap \texttt{Nb}_{m_2}|. \tag{3}$$

Then the condition of Theorem 5.4.3 is satisfied for $q = 2$, so that `FSEtagNbs` is IND-CLS-CPA-insecure for any domain having m_0, m_1, m_2 satisfying (3).

The schemes of [48] are, essentially, instantiations of `FSEtagNbs` on closeness domains defined in terms of keywords and edit distance. If δ is the threshold edit distance, take m_2 to be any message of length at least 2δ. Let m_0 be m_2 but with the first $\delta + 1$ letters changed. Let m_1 be m_0 but with the last $\delta - 1$ letters changed. Then m_0, m_1, m_2 satisfy (3) and hence `FSEtagNbs` is IND-CLS-CPA-insecure here.

5.5.2 Construction of the first secure EFSE scheme

We now improve on the scheme of Section 5.5.1 and construct an EFSE scheme that is IND-CLS-CPA-secure even on rigid closeness domains. Let $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$ be a closeness domain with $\mathcal{D}$ finite and fixed message length ℓ. (That is, assume every $m \in \mathcal{D}$ can be uniquely described as a string of length ℓ.) Let $\mathcal{G}_\Lambda = (\mathcal{V}_\Lambda, \mathcal{E}_\Lambda)$ be the closeness graph of Λ. For $m \in \mathcal{D}$, let $E_m = \{\{m, m'\} \in \mathcal{E}_\Lambda \mid m' \in \mathcal{V}_\Lambda\}$ be the set of incident edges to m in $\mathcal{G}_\Lambda$, and $\Delta_m = |E_m|$. Let $\Delta = \max_{m \in \mathcal{D}} \Delta_m$.

Construct a new graph $\mathcal{G}_{\texttt{dum}} = (\mathcal{V}_{\texttt{dum}}, \mathcal{E}_{\texttt{dum}})$ where $\mathcal{V}_{\texttt{dum}} = \mathcal{V}_\Lambda \cup \{w_1, \ldots, w_\Delta\}$, and $\mathcal{E}_{\texttt{dum}}$ consists of all edges in $\mathcal{E}_\Lambda$, plus for any $m \in \mathcal{V}_\Lambda$, if $\Delta - \Delta_m > 0$ then let $\mathcal{E}_{\texttt{dum}}$ also contain edges $\{m, w_1\}, \ldots, \{m, w_{\Delta - \Delta_m}\}$. We call these additional edges *dummy edges* and $w_1, \ldots, w_\Delta$ *dummy vertices*. $\mathcal{G}_{\texttt{dum}}$ is thus a graph in which every element of $\mathcal{V}_\Lambda \subset \mathcal{V}_{\texttt{dum}}$ has degree Δ.

Define $\texttt{BktEdges} : \mathcal{D} \to \mathcal{E}_{\texttt{dum}}$ as

$$\texttt{BktEdges}(m) = \{e \in \mathcal{E}_{\texttt{dum}} \mid m \in e\}.$$

Then if $\mathrm{Cl}_{\mathcal{D}}(m, m') = \texttt{close}$ then $\texttt{BktEdges}(m) \cap \texttt{BktEdges}(m') \supseteq \{\{m, m'\}\} \neq \emptyset$; and if $\mathrm{Cl}_{\mathcal{D}}(m, m') = \texttt{far}$ then $\texttt{BktEdges}(m) \cap \texttt{BktEdges}(m') = \emptyset$. So `BktEdges` is a CPBF.

Let $\mathcal{F}_{\texttt{BTag}}$ be a collision-free batch-tagging family on domain $\mathcal{E}_{\texttt{dum}}$ and some range $\mathcal{R}_{\mathcal{T}}$, and let ESE be an ESE scheme on $\mathcal{D}$. Define the StructFSE scheme FSEtagEdges as $\texttt{FSE}_{\texttt{BktTag}}[\texttt{BktEdges}, \mathcal{F}_{\texttt{BTag}}, \texttt{ESE}]$ according to Figure 10. Notice that for all $m \in \mathcal{D}$, $|\texttt{BktEdges}(m)| \leq \Delta$. So, if Λ is defined so that Δ is small, FSEtagEdges is efficient.

Now we provide the security guarantee of FSEtagEdges.

Theorem 5.5.1. *Let Λ be a closeness domain with fixed message length ℓ, and* FSEtagEdges *the scheme defined above. For any adversary A there exist adversaries E_A and F_A such that*

$$\mathbf{Adv}^{\text{ind-cls-cpa}}_{\texttt{FSEtagEdges}}(A) = \mathbf{Adv}^{\text{ind-dcpa}}_{\texttt{ESE}}(E_A) + \mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\texttt{BTag}}}(F_A).$$

Further, suppose A submits q length-2ℓ queries to its oracle. Then

- *E_A submits q queries to its oracle, each of length $\leq 4\Delta\ell$; q queries to $\mathcal{E}nc_{\texttt{ESE}}$, each of length ℓ; and 2 queries to* Bkts*, each of length ℓ;*
- *F_A submits q queries to its oracle, each of length 2ℓ; q queries to $\mathcal{B}$, each of length $\leq 2\Delta\ell$; and 1 query to* Bkts*, of length ℓ.*

Otherwise, A, E_A, and F_A have the same running time.

The proof is in Appendix C.5, and simply shows the condition of Theorem 5.4.2 (i.e., consistency of BktEdges) is satisfied in this case.

If $\mathcal{E}nc_{\texttt{ESE}}$ and $\mathcal{B}$ are implemented efficiently, then the efficiency of E_A and F_A are each bounded by essentially a factor Δ times the efficiency of A. Thus, if Δ is small, and if ESE is IND-DCPA-secure and $\mathcal{F}_{\texttt{BTag}}$ is PP-CBT-secure, then FSEtagEdges is IND-CLS-CPA-secure. Recall that certain blockcipher-based constructions (discussed earlier) satisfy the necessary efficiency, security, and functionality conditions for ESE and $\mathcal{F}_{\texttt{BTag}}$. Finally, the last missing piece to achieve our IND-CLS-CPA-secure scheme is that BktEdges should be efficiently constructible, which holds if Λ is defined so that

E_m for any message $m \in \mathcal{D}$ is predetermined or easily calculated on-the-fly. Thus, if we assume two conditions on the closeness domain Λ:

- $\Delta = \max_{m \in \mathcal{D}} |E_m|$ is small;
- E_m is predetermined or easily calculated on-the-fly;

then `FSEtagEdges` is an IND-CLS-CPA-secure EFSE scheme on Λ.

So, we have successfully created a IND-CLS-CPA-secure scheme, but at what cost? It is apparent that the ciphertexts in `FSEtagEdges` can be quite long, namely, their length is linear in Δ, the maximum number of close neighbors of a message in Λ (not to mention the fact that a large Δ weakens the security reduction in Theorem 5.5.1). Δ could certainly be quite large—for example, on a metric closeness domain, even a relatively small threshold causes each message to have many close neighbors, and Δ increases exponentially with dimension of a metric closeness domain.

However, in the next section we show that if we desire a general FSE construction to work on arbitrary closeness domains, such long ciphertexts are necessary.

5.5.3 Lower bound on ciphertext length of an arbitrary-domain FSE scheme

The following result demonstrates the existence of closeness domains Λ on which any FSE scheme must have ciphertext length linear in the maximum degree of $\mathcal{G}_\Lambda$. (This matches the space-efficiency of `FSEtagEdges` from the previous section, demonstrating that `FSEtagEdges` is "best-possible" for FSE on certain closeness domains.) As we will see, this is an informational theoretic requirement, and relies only on functionality, rather than security, of the scheme. The proof of the following theorem is in Appendix C.6.

Theorem 5.5.2. *For any $\Delta > 0$, there exists a rigid closeness domain Λ where $\mathcal{G}_\Lambda$ has maximum degree Δ such that (for correctness) any FSE scheme built on Λ must have ciphertext length $\Omega(\Delta)$.*

5.6 Space-Efficient Schemes

The result of Theorem 5.5.2 indicates that it is costly to afford IND-CLS-CPA-security on general rigid closeness domains. A natural question is whether we can do better for non-rigid closeness domains, where we have an extra freedom: namely, near message pairs may be sent to either (i) far ciphertext pairs or (ii) close ciphertext pairs. However, note that if an adversary has any probabilistic edge in guessing which near message pairs are sent to category (i) and which to category (ii), he can easily break IND-CLS-CPA-security. The only way to avoid this attack would be for all near message pairs to have uniform probability to end up in category (i) vs. category (ii). And this negates the flexibility of having near messages—we expect an EFSE scheme satisfying this uniformity condition on near pairs would be just as inefficient as the `FSEtagEdges` scheme. Thus, it appears that IND-CLS-CPA-security is too strong for more efficient EFSEs to achieve, even on non-rigid closeness domains. So for more efficient schemes, we need a new, weaker notion of security.

Intuitively, what information do we hope an EFSE scheme on a non-rigid closeness domain Λ will protect, given that some number of ciphertexts are known? Let H be the set of messages corresponding to known ciphertexts. For two messages in the same component of the induced nearness subgraph $\mathcal{G}_\Lambda^{\mathtt{N}}(H)$ (we say they are in the same *nearness component*) an EFSE is designed so that anyone might discover this fact by looking at their ciphertexts. So, by using EFSE we automatically give up a large amount of information about messages in the same nearness component (namely, their link through a chain of near pairs.) It is a natural step to consider allowing more information leakage for messages within the same nearness component, while protecting as much as possible about messages in different nearness components—a kind of "inter-nearness-component security."

With this goal in mind, we introduce a security notion that requires schemes to hide all information about plaintexts in different nearness components except for an

aspect of "local structure"—which can be intuitively understood as the least significant bits of messages. As we will see, this local structure will be characterized by a message's relationship to a pre-chosen fixed regular lattice $\mathcal{L}$ (that we call the *anchor lattice*) on the message space. The important implication is that nothing major (i.e., only "local structure") is revealed about the relationship between a pair of disconnected messages (i.e., messages that cannot be connected through a chain of near known corresponding ciphertext pairs). Hence, it is a sort of "macrostructure security" across disconnected nearness components. Also, since it is difficult to formalize these concepts on general closeness domains, we confine our view to metric closeness domains, and specifically, Euclidean distance on arbitrary-dimension real domains. Such closeness domains have many applications, as we shall see.

5.6.1 Macrostructure security on metric domains

Before defining the new notion of security, we first introduce the concept of an anchor lattice, which plays a role both in the security notion and in constructions. Intuitively, a short basis for the anchor lattice distinguishes the "local" scale (say, on the order of small-constant combinations of the short basis vectors) as opposed to the "macro" scale in the domain.

Anchor lattice. Let $\mathcal{L}$ be a regular lattice[4] in $\mathbb{R}^\ell$, that is, a set of vectors characterized as all integer combinations of some linearly independent *basis vectors* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\ell \in \mathbb{R}^\ell$. We call $\mathcal{L}$ an *anchor lattice*, and assume that a short basis for it is public. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^\ell$, we say that $\mathbf{x}$ and $\mathbf{y}$ are in the same *(translation-invariant) $\mathcal{L}$-class* if there exists $\mathbf{w} \in \mathcal{L}$ with $\mathbf{x} + \mathbf{w} = \mathbf{y}$; in this case, we say $\mathbf{w}$ is the *$\mathcal{L}$-witness* from $\mathbf{x}$ to $\mathbf{y}$.

Macrostructure security. Let $\Lambda = \left(\mathbb{R}^\ell, \mathcal{M}_d^{\delta^{\mathtt{C}}, \delta^{\mathtt{F}}}\right)$ be the metric closeness domain on $\mathbb{R}^\ell$, the Euclidean metric d, close threshold $\delta^{\mathtt{C}} > 0$, and far threshold $\delta^{\mathtt{F}} \geq \delta^{\mathtt{C}}$.

[4]Our use of lattices should not be confused with the techniques of lattice-based cryptography.

Let $\mathcal{L}$ be an anchor lattice on $\mathbb{R}^\ell$. We introduce a notion of security called *indistinguishability under same-nearness-component-and-$\mathcal{L}$-class chosen-plaintext attack*, which for simplicity we rename *macrostructure security with respect to anchor lattice* $\mathcal{L}$ as "same nearness component and $\mathcal{L}$ class" places no restrictions on the (disconnected) macro-scale structure of adversarial queries.

Let $\mathtt{FSE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}})$ be an EFSE scheme on Λ. For an adversary A and $b \in \{0,1\}$, let experiment $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-}b}_{\mathtt{FSE}}(A)$ be identical to IND-CPA experiment $\mathbf{Exp}^{\text{ind-cpa-}b}_{\mathtt{FSE}}(A)$ in Figure 1 but with the restriction: for left/right-queries (m_0^i, m_1^i), $i \in [q]$ made by the adversary, letting $H_0 = \{m_0^1, \ldots, m_0^q\}$ and $H_1 = \{m_1^1, \ldots, m_1^q\}$, require

1. $|m_0^i| = |m_1^i|$ for all $i \in [q]$;
2. $\forall i \in [q]$, m_0^i and m_1^i are in the same $\mathcal{L}$-class; furthermore, the $\mathcal{L}$-witness from m_0^i to m_1^i is also the $\mathcal{L}$-witness from m_0^j to m_1^j whenever m_0^i and m_0^j are in the same connected component of $\mathcal{G}^{\mathtt{N}}_{\Lambda}(H_0)$.

For an adversary A, define its *IND-NR$\mathcal{L}$-CPA advantage* against $\mathtt{FSE}$ as

$$\mathbf{Adv}^{\text{ind-nr}\mathcal{L}\text{-cpa}}_{\mathtt{FSE}}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}(A) = 1\,\right] - \Pr\left[\,\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-0}}_{\mathtt{FSE}}(A) = 1\,\right].$$

We say $\mathtt{FSE}$ is *indistinguishable under same-nearness-component-and-$\mathcal{L}$-class chosen-plaintext attack* (IND-NR$\mathcal{L}$-CPA-secure), or alternatively *macrostructure-secure with respect to anchor lattice* $\mathcal{L}$ (MacroStruct-$\mathcal{L}$-secure) if the IND-NR$\mathcal{L}$-CPA advantage of any adversary against $\mathtt{FSE}$ is small.

The second left/right-query requirement asks that a left-query component of $\mathcal{G}^{\mathtt{N}}_{\Lambda}(H_0)$ is a $\mathcal{L}$-translation (translation by a vector in $\mathcal{L}$) of the corresponding right-query component of $\mathcal{G}^{\mathtt{N}}_{\Lambda}(H_1)$. This implies that left and right queries have the same equality/closeness pattern:

- if $m_0^i = m_0^j$ then these messages are in the same nearness component (as they are the same vertex) so $\exists l \in \mathcal{L}$ with $m_1^i = m_0^i + l = m_0^j + l = m_1^j$;

- if $\mathrm{Cl}(m_0^i, m_0^j) \in \{\texttt{close}, \texttt{near}\}$ then these messages are in the same nearness component so $\exists l \in \mathcal{L}$ with $m_1^i = m_0^i + l$, $m_1^j = m_0^j + l$, implying $d(m_1^i, m_1^j) = d(m_1^i + l, m_1^j + l) = d(m_0^i, m_0^j)$, so $\mathrm{Cl}(m_1^i, m_1^j) = \mathrm{Cl}(m_0^i, m_0^j)$.

Thus, MacroStruct-$\mathcal{L}$-security is clearly weaker than IND-CLS-CPA-security.

Returning to the big picture, an MacroStruct-$\mathcal{L}$-secure scheme may leak how all messages in a nearness component lie with respect to nearby points in the anchor lattice. However, since the lattice itself is regular, no information is leaked about where those nearby lattice points actually are. Thus, for messages in different nearness components, an adversary learns nothing about the distance between them, or their approximate locations in the space, besides some quite-insignificant bits, as well as that the distance is above $\delta^{\mathtt{F}}$ (which is by design.)

Outline of remainder of section. Unless otherwise noted, let $\Lambda = \left(\mathbb{R}^\ell, \mathcal{M}_d^{\delta^{\mathtt{C}}, \delta^{\mathtt{F}}}\right)$ be the metric closeness domain on $\mathbb{R}^\ell$, the Euclidean metric d, close threshold $\delta^{\mathtt{C}} > 0$, and far threshold $\delta^{\mathtt{F}} \geq \delta^{\mathtt{C}}$. We aim to construct space-efficient EFSE schemes on Λ that meet our new notion of MacroStruct-$\mathcal{L}$-security for anchor lattices $\mathcal{L} \subset \mathbb{R}^\ell$. Before getting to specific schemes, we introduce the concept of an "anchor radius" on Λ and $\mathcal{L}$, and show that a valid anchor radius ρ implies an EFSE construction (based on the bucketing template of Section 5.4) on Λ that is MacroStruct-$\mathcal{L}$-secure. Next, we introduce a general problem of trying to improve space-efficiency and accommodate as-small-as-possible near thresholds of such EFSE schemes by choosing lattices $\mathcal{L}$ and anchor radii ρ wisely. We then introduce several specific schemes on particular closeness domains. Finally, we explain how to build an EFSE scheme on a joint closeness domain (a product of small-dimension closeness domains, with closeness defined conjunctively) as might be useful in biometric-data-matching applications.

5.6.2 Anchor radii and general macrostructure-secure construction

Fix anchor lattice $\mathcal{L}$ in $\mathbb{R}^\ell$. For $\rho > 0$, we say that ρ is an *anchor radius* on closeness domain Λ and anchor lattice $\mathcal{L}$ if

1. for any $\mathbf{m}, \mathbf{m}' \in \mathcal{D}$ with $d(\mathbf{m}, \mathbf{m}') \leq \delta^{\mathtt{C}}$, there exists $\mathbf{v} \in \mathcal{L}$ such that $d(\mathbf{v}, \mathbf{m}) \leq \rho$ and $d(\mathbf{v}, \mathbf{m}') \leq \rho$;
2. $\rho \leq \delta^{\mathtt{F}}/2$.

If ρ is an anchor radius on Λ and $\mathcal{L}$, then $\mathtt{BktsAnc}^{\rho}_{\mathcal{L}} : \mathbb{R}^\ell \to \mathcal{L}$ defined as

$$\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}) = \{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}, \mathbf{v}) \leq \rho\}$$

is a CPBF on Λ, as condition (1) implies that whenever $d(\mathbf{m}, \mathbf{m}') \leq \delta^{\mathtt{C}}$, there exists $\mathbf{v} \in \mathcal{L}$ such that $\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}) \cap \mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}') \supseteq \{\mathbf{v}\}$; and condition (2) implies $\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}) \cap \mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}') = \emptyset$ whenever $d(\mathbf{m}, \mathbf{m}') > \delta^{\mathtt{F}}$. Thus, if ρ is an anchor radius on Λ and $\mathcal{L}$, $\mathcal{F}_{\mathtt{BTag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ is a collision-free batch-tagging family on domain $\mathcal{D}_{\mathcal{T}} = \mathcal{L}$, and $\mathtt{ESE}$ is an ESE scheme on $\mathcal{D}$, then the scheme $\mathtt{FSEtagAnc}^{\rho}_{\mathcal{L}} = \mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$ is an EFSE by Section 5.4.

The following result, proved in Appendix C.7, shows that any EFSE scheme $\mathtt{FSEtagAnc}^{\rho}_{\mathcal{L}}$ defined in the above manner is MacroStruct-$\mathcal{L}$-secure provided $\mathtt{ESE}$ is IND-DCPA-secure and $\mathcal{F}_{\mathtt{BTag}}$ is PP-CBT-secure.

Theorem 5.6.1. *Let closeness domain Λ, anchor lattice $\mathcal{L}$, anchor radius ρ, and EFSE scheme $\mathtt{FSEtagAnc}^{\rho}_{\mathcal{L}}$ be defined as above. For any adversary A there exist adversaries E_A and F_A such that*

$$\mathbf{Adv}^{\text{ind-nr}\mathcal{L}\text{-cpa}}_{\mathtt{FSEtagAnc}^{\rho}_{\mathcal{L}}}(A) = \mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(E_A) + \mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A).$$

Further, let $\mu = \max_{\mathbf{m} \in \mathcal{D}} |\{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}, \mathbf{v}) \leq \rho\}|$, and if A submits q queries to its oracle, of total query length 2γ, then

- *E_A submits q queries to its oracle, of total query length 2γ, and also submits q queries to $\mathcal{B}$, of total query length at most $2\mu(\gamma + q\log_2 \rho)$;*
- *F_A submits q queries to its oracle, of total query length at most $4\mu(\gamma + \log_2 \rho)$, and also submits q queries to $\mathcal{E}nc_{\mathtt{ESE}}$, of total query length γ.*

Thus, if we can find an anchor radius ρ on closeness domain Λ and anchor lattice $\mathcal{L}$ then $\mathtt{FSEtagAnc}^{\rho}_{\mathcal{L}}$ as constructed above is an MacroStruct-$\mathcal{L}$-secure EFSE scheme on Λ.

5.6.3 On attaining space-efficiency and small nearness threshold

Suppose that we are given a close threshold $\delta^{\mathtt{C}} > 0$ and are asked to provide an EFSE scheme on $\Lambda = \left(\mathbb{R}^{\ell}, \mathcal{M}_d^{\delta^{\mathtt{C}},\delta^{\mathtt{F}}}\right)$ where $\delta^{\mathtt{F}}$ can be chosen as needed, with two objectives: first, minimize $\delta^{\mathtt{F}}$ (to accommodate stricter closeness domains), and second, minimize ciphertext length, which depends on the distribution of $|\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ for $\mathbf{m} \in \mathcal{D}$. Intuitively, once an anchor lattice $\mathcal{L}$ is fixed, this means choosing minimal ρ so that $\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}) \cap \mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m}') \neq \emptyset$ whenever $d(\mathbf{m}, \mathbf{m}') \leq \delta^{\mathtt{C}}$, and then setting $\delta^{\mathtt{F}} = 2\rho$ so that ρ is an anchor radius. A smaller ρ with respect to $\mathcal{L}$ will mean both a smaller ciphertext length and a smaller $\delta^{\mathtt{F}}$.

Faced with the challenge, though, it is unclear what anchor lattice $\mathcal{L}$ to start with. In fact, the best solution may depend on our relative valuation of minimizing $\delta^{\mathtt{F}}$ versus minimizing ciphertext length. A denser lattice would seem to increase ciphertext length while allowing for smaller anchor radius $\rho = \frac{1}{2}\delta^{\mathtt{F}}$; a sparser lattice would have the opposite effect. This is an interesting question and we pose the following open problem.

Problem 5.6.2. *Given a space $\mathbb{R}^{\ell}$ and close threshold $\delta^{\mathtt{C}} > 0$, minimize a function of $\max_{\mathbf{m}\in\mathbb{R}^{\ell}} |\{\mathbf{v} \mid d(\mathbf{m}, \mathbf{v}) \leq \rho|$ and ρ by selecting an appropriate anchor lattice $\mathcal{L} \subset \mathbb{R}^{\ell}$ and setting ρ to be the minimal constant such that every pair of close points (i.e., distance at most $\delta^{\mathtt{C}}$) in $\mathbb{R}^{\ell}$ are each within ρ of the same point in $\mathcal{L}$.*

A straightforward possibility is to pick an anchor lattice $\mathcal{L}$ so that balls of some radius ω around each lattice point cover the entire space, and to set $\rho = \delta^{\mathtt{C}} + \omega$. Then, a point $\mathbf{m}$ is distance ω from some $\mathbf{v} \in \mathcal{L}$, and $d(\mathbf{m}, \mathbf{m}') \leq \delta^{\mathtt{C}}$ implies $d(\mathbf{m}', \mathbf{v}) \leq \rho$ by the triangle inequality. But how to choose the anchor lattice and ω? And it is likely we can do better. In any case, addressing Problem 5.6.2 is beyond the scope of this work. Instead, in the next subsection we propose what seem to be "good" practical choices of $\mathcal{L}$ and ρ for various spaces.

5.6.4 Specific anchor-based schemes for various dimensions

We now introduce several specific constructions of tag-anchor EFSE schemes, each defined by selecting an appropriate anchor lattice and anchor radius. For simplicity, in each of these examples we assume close threshold $\delta^{\mathtt{C}} = 1$. (Other close thresholds are possible by scaling.) See Table 3 for a summary of the constructions.

Table 3: Summary of space-efficiency and minimum near threshold values for specific anchor-based EFSE schemes on real spaces with close threshold 1.

name	domain	anchor rad. ρ	$\lvert\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(\cdot)\rvert$ range	Minimum $\delta^{\mathtt{F}}$
Integer lattice	$\mathbb{R}^1$	1	$\{2, 3\}$	2
Triangular lattice	$\mathbb{R}^2$	$\sqrt{5}/2$	$\{3, 4, 5, 6, 7\}$	$\sqrt{5} \approx 2.24$
Rectangular grid	$\mathbb{R}^\ell$, $\ell \geq 1$	3/2	see Table 4	3

Integer lattice for $\mathbb{R}^1$. Let $\Lambda = \left(\mathbb{R}, \mathcal{M}^{1,\delta^{\mathtt{F}}}\right)$. Set $\mathcal{L} = \mathbb{Z}$, and set $\rho = 1$. Then ρ is a valid anchor radius: if $d(m, m') \leq 1$, then there exists an integer z in between m and m' such that $d(m, z) \leq 1$ and $d(m', z) \leq 1$.

Minimum near threshold: 2.

Space efficiency: $\lvert\mathtt{BktsAnc}^{\rho}_{\mathcal{L}}(m)\rvert \in \{2, 3\}$ for all $m \in \mathcal{D}$.

Triangular lattice for $\mathbb{R}^2$. Let $\Lambda = \left(\mathbb{R}^2, \mathcal{M}^{1,\delta^{\mathtt{F}}}\right)$. Set $\mathcal{L}$ to be the regular triangular lattice generated by the vectors $(1, 0)$ and $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$, and set $\rho = \sqrt{5}/2$. Then ρ is a valid anchor radius, by the following argument. Let $\mathbf{m} \in \mathbb{R}^2$, and let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathcal{L}$ be the vertices of (one of) the triangular region(s) T containing $\mathbf{m}$.

The union of three balls, each of radius $\frac{\sqrt{5}}{2}$ and centered at the three points $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, cover all of T as well as every point within distance 1 of a point of T. (A point on the border of this region that is closest to T is one at the intersection of two of these balls, which is distance $\sqrt{\left(\frac{\sqrt{5}}{2}\right)^2 - \left(\frac{1}{2}\right)^2} = 1$ from a midpoint of one of T's edges.) Thus, for any $\mathbf{m}' \in \mathbb{R}$ with $d(\mathbf{m}, \mathbf{m}') \leq 1$, we have $d(\mathbf{m}', \mathbf{v}_i) \leq \frac{\sqrt{5}}{2}$ for some $i \in \{1, 2, 3\}$.

Minimum near threshold: $\sqrt{5}$.

Space efficiency: $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})| \in \{3, 4, 5, 6, 7\}$ for all $\mathbf{m} \in \mathcal{D}$. (See Figure 13 in Appendix C.8.)

Rectangular grid for arbitrary dimension. Fix a dimension $\ell \geq 1$, and let $\Lambda = \left(\mathbb{R}^\ell, \mathcal{M}_d^{1,\delta^{\mathrm{F}}}\right)$. Set

$$\mathcal{L} = \frac{\mathbb{Z}^\ell}{\sqrt{\ell}} = \left\{ \left(\frac{z_1}{\sqrt{\ell}}, \ldots, \frac{z_\ell}{\sqrt{\ell}} \right) \mid z_1, \ldots, z_\ell \in \mathbb{Z} \right\}.$$

Set $\rho = 3/2$, and ρ is a valid anchor radius by the following argument. Let $\mathbf{m}, \mathbf{m}' \in \mathbb{R}^\ell$ with $d(\mathbf{m}, \mathbf{m}') \leq 1$. First, note that the points in $\mathbb{R}^\ell$ furthest from elements of $\mathcal{L}$ are the half-grid points, e.g. $\left(\frac{1}{2\sqrt{\ell}}, \ldots, \frac{1}{2\sqrt{\ell}}\right)$, which are distance $1/2$ from a grid point. So let $\mathbf{z} \in \mathcal{L}$ be a grid point that is within distance $1/2$ of $\mathbf{m}$. Then $d(\mathbf{m}', \mathbf{z}) \leq d(\mathbf{m}', \mathbf{m}) + d(\mathbf{m}, \mathbf{z}) \leq 1 + 1/2$. Hence $d(\mathbf{m}, \mathbf{z})$ and $d(\mathbf{m}', \mathbf{z})$ are both at most $3/2$.

Minimum near threshold: 3

Space efficiency: depends on ℓ. See Appendix C.8 for analysis: we argue that finding an easy formula for $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ is difficult or impossible, we derive a theoretical upper bound on $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ in terms of ℓ, and we report values from empirical tests. Table 4 (expanded in the Appendix) displays some results.

5.6.5 Conjunctive closeness for multiple attributes

In many fuzzy search applications, we may desire fuzziness to cover many attributes in a conjunctive (AND) manner. As a motivating example, consider a database of people's identities, indexed by biometrics such as height, weight, fingerprint data, and iris scan data. Such attributes are generally each 1- or 2-dimensional, but taken

Table 4: Analysis of the distribution of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ for $\mathbf{m} \in \mathcal{L}$ and various dimensions ℓ: a loose upper bound, empirically-computed value at a grid point, and empirical maxima and minima among 10000 random points.

ℓ	Loose upper bound	Value at grid point	Empirical lower bound	Empirical upper bound
1	5	3	3	4
2	39	13	12	16
3	340	81	68	81
4	3084	425	425	1023

together a person's set of biometric data defines a point in a high-dimensional space. A query may be a set of biometric measurements to search for a patient's health records, which are inaccurate by nature. In such an application, we desire closeness on each individual biometric measurement independently, that is, "heights should be close AND weights should be close AND ...," to achieve a match.

Let $\Lambda_1, \ldots, \Lambda_r$ be closeness domains, where $\Lambda_i = \left(\mathbb{R}^{\ell_i}, \mathcal{M}_{d_i}^{\delta^{\mathtt{C}}, \delta^{\mathtt{F}}}\right)$ for all $i \in [r]$. For $i \in [r]$, let $\mathcal{L}_i$ be a regular lattice on $\mathcal{D}_i$ and ρ_i a valid anchor radius on Λ_i, $\mathcal{L}_i$. Define conjunctive closeness domain $\Lambda = (\mathcal{D}, \mathrm{Cl})$ where $\mathcal{D} = \mathbb{R}^{\ell_1} \times \cdots \times \mathbb{R}^{\ell_r}$ and $\mathrm{Cl} : \mathcal{D} \to \{\texttt{close}, \texttt{near}, \texttt{far}\}$ is defined as

$$\mathrm{Cl}(\mathbf{m}, \mathbf{m}') = \begin{cases} \texttt{close} & \text{if } d_i(m_i, m_i') \le \delta^{\mathtt{C}} \text{ for all } i \in [r]; \\ \texttt{far} & \text{if } d_i(m_i, m_i') > \delta^{\mathtt{F}} \text{ for any } i \in [r]. \\ \texttt{near} & \text{otherwise.} \end{cases}$$

Let $\mathcal{L} = \mathcal{L}_1 \times \cdots \times \mathcal{L}_r$, and define $\texttt{BktsAnc}^{\rho_1,\ldots,\rho_r}_{\mathcal{L}} : \mathcal{D} \to \mathcal{L}$ as

$$\texttt{BktsAnc}^{\rho_1,\ldots,\rho_r}_{\mathcal{L}}(\mathbf{m}) = \{(l_1, \ldots, l_r) \mid d_i(m_i, l_i) \le \rho_i \forall i \in [r]\}.$$

Then $\texttt{BktsAnc}^{\rho_1,\ldots,\rho_r}_{\mathcal{L}}$ is a CPBF on Λ:

- if $\mathrm{Cl}(\mathbf{m}, \mathbf{m}') = \texttt{close}$ then $\mathrm{Cl}_i(m_i, m_i') = \texttt{close}$ for all $i \in [r]$, so for each $i \in [r]$ there exists $l_i \in \mathcal{L}_i$ such that $d_i(m_i, l_i) \le \rho_i$ and $d_i(m_i', l_i) \le \rho_i$ and thus $(l_1, \ldots, l_r) \in \texttt{BktsAnc}^{\rho_1,\ldots,\rho_r}_{\mathcal{L}}(\mathbf{m}) \cap \texttt{BktsAnc}^{\rho_1,\ldots,\rho_r}_{\mathcal{L}}(\mathbf{m}')$;

- if $\mathrm{Cl}(\mathbf{m}, \mathbf{m}') = \texttt{far}$ then $\mathrm{Cl}_i(m_i, m_i') = \texttt{far}$ for some $i \in [r]$, so $\nexists l_i \in \mathcal{L}_i$ with $d_i(m_i, l_i) \leq \rho_i$, $d_i(m_i', l_i) \leq \rho_i$, so $\texttt{BktsAnc}_{\mathcal{L}}^{\rho_1,\ldots,\rho_r}(\mathbf{m}) \cap \texttt{BktsAnc}_{\mathcal{L}}^{\rho_1,\ldots,\rho_r}(\mathbf{m}') = \emptyset$.

A similar argument to that of Theorem 5.6.1 then shows that the EFSE scheme $\texttt{FSEtagAnc}_{\mathcal{L}}^{\rho_1,\ldots,\rho_r} = \texttt{FSE}_{\texttt{BktTag}}[\texttt{BktsAnc}_{\mathcal{L}}^{\rho_1,\ldots,\rho_r}, \mathcal{F}_{\texttt{BTag}}, \texttt{ESE}]$ is MacroStruct-$\mathcal{L}$-secure provided $\texttt{ESE}$ is an IND-DCPA-secure ESE scheme on $\mathcal{D}$ and $\mathcal{F}_{\texttt{BTag}}$ is a PP-CBT-secure collision-free batch-tagging family on $\mathcal{L}$.

CHAPTER VI

CONCLUSION

I studied the cryptographic properties of order-preserving encryption (OPE) and efficient fuzzy-searchable encryption (EFSE). These are primitives that enable efficient (respectively) range query and error-tolerant query support on encrypted data, which is motivated by the desire to encrypt information on untrusted cloud storage servers without sacrificing efficient and flexible query support on the data.

Our cryptographic studies of OPE and EFSE represent the first provable-security study of each. For both, we defined appropriate primitives and optimal security notions, then developed provably-secure constructions. The OPE case demanded extra security analysis in the form of one-wayness bounds on our construction. For EFSE, the optimally-secure scheme was somewhat space-inefficient, so we proposed space-efficient schemes that are secure under a weaker, but practical, security notion.

Our analyses of OPE and EFSE are not comprehensive—they should be viewed as first steps, perhaps natural first steps, toward characterizing security of the primitives. For future work, it is possible that alternative security notions for OPE and space-efficient EFSE schemes may be useful, based on an application's needs, and new constructions should be built accordingly.

OPE and EFSE schemes existed before our research, but there was no compelling reason to use them as they had no guarantees of security. Now, armed with our provable-security guarantees, practitioners can implement our schemes with a greater knowledge of the risks avoided, and a greater awareness of what risks may not be avoided. Thus, our research allows practitioners to conscientiously implement efficient range or error-tolerant queries on encrypted data in a cloud storage setting.

APPENDIX A

OPE AND POPFS PROOFS

A.1 Proof of Theorem 3.2.1

Proof of Theorem 3.2.1. Let $M' = \lfloor M/2 \rfloor$. Consider the following IND-OCPA adversary A against $\mathcal{OPE}$:

Adversary $A^{\mathcal{E}nc(K,\mathcal{LR}(\cdot,\cdot,b))}$

$m_0 \xleftarrow{\$} [M']$, $m_1 \leftarrow M - m_0 + 1$

$c \leftarrow \mathcal{E}nc(K, \mathcal{LR}(m_0, m_1, b))$

$c_L \leftarrow 0$; $c_R \leftarrow N + 1$

If $m_0 > 1$:

$m_L \leftarrow m_0 - 1$, $m_R \leftarrow m_1 + 1$

$c_L \leftarrow \mathcal{E}nc(K, \mathcal{LR}(m_L, m_L, b))$

$c_R \leftarrow \mathcal{E}nc(K, \mathcal{LR}(m_R, m_R, b))$

Return 1 with probability $\frac{c-c_L}{c_R-c_L}$

Else return 0

The IND-OCPA correctness and efficiency claims of A should be clear from the construction.

Fix a key K, so that $\mathcal{E}nc(K, \cdot)$ is a well-defined order-preserving function from $[M]$ to $[N]$. For $m \in [M']$, let $X_m = \mathcal{E}nc(K, M-m+1) - \mathcal{E}nc(K, m)$ and $X_0 = N+1$. Let S be the set of messages m in $[M']$ such that $\frac{X_m}{X_{m-1}} \leq \frac{1}{c}$. Then if $|S| > M'/2$ we have

$$N + 1 = X_0 \geq \frac{X_0}{X_{M'}} = \prod_{m \in [M']} \frac{X_{m-1}}{X_m} > t^{M'/2} \geq t^{\lfloor M/4 \rfloor},$$

a contradiction to $N < t^{\lfloor M/4 \rfloor}$. Thus, $|S| \leq M'/2$ and so

$$\Pr\left[m \xleftarrow{\$} [M'] \;\middle|\; \frac{X_m}{X_{m-1}} \leq \frac{1}{t} \right] \leq \frac{1}{2}.$$

In the following, for given $m_0 \in [M']$ let $m_1, m_L, m_R, c_0, c_1, c_L, c_R$ be determined from m_0 as in A. Using Markov's inequality,

$$\begin{aligned}
\mathbf{Adv}^{\text{ind-ocpa}}_{\mathcal{OPE}}(A) &= \Pr\left[\mathbf{Exp}^{\text{ind-ocpa-1}}_{\mathcal{OPE}}(A) = 1 \right] - \Pr\left[\mathbf{Exp}^{\text{ind-ocpa-0}}_{\mathcal{OPE}}(A) = 1 \right] \\
&= \Pr\left[A^{\mathcal{E}nc(K,\mathcal{LR}(\cdot,\cdot,1))} = 1 \right] - \Pr\left[A^{\mathcal{E}nc(K,\mathcal{LR}(\cdot,\cdot,0))} = 1 \right] \\
&= \mathop{\mathbb{E}}_{m_0 \xleftarrow{\$} [M']} \left[\frac{c_1 - c_L}{c_R - c_L} \right] - \mathop{\mathbb{E}}_{m_0 \xleftarrow{\$} [M']} \left[\frac{c_0 - c_L}{c_R - c_L} \right] \\
&= \mathop{\mathbb{E}}_{m_0 \xleftarrow{\$} [M']} \left[\frac{c_1 - c_0}{c_R - c_L} \right] \\
&\geq \frac{1}{t} \Pr_{m_0 \xleftarrow{\$} [M']} \left[\frac{c_1 - c_0}{c_R - c_L} \geq \frac{1}{t} \right] \\
&= \frac{1}{t} \Pr_{m_0 \xleftarrow{\$} [M']} \left[\frac{X_m}{X_{m-1}} \geq \frac{1}{t} \right] \\
&> \frac{1}{2t}.
\end{aligned}$$

This completes the proof. □

A.2 Proof of Theorem 3.3.2

Proof of Theorem 3.3.2. Since we consider unbounded adversaries, we can ignore the inverse oracle in our analysis, since such an adversary can always query all points in the domain to learn all points in the image. Let $M = |\mathcal{D}|$, $N = |\mathcal{R}|$, $d = \min(\mathcal{D}) - 1$, and $r = \min(\mathcal{R}) - 1$. We will say that two functions $g, h : \mathcal{D} \to \mathcal{R}$ are *equivalent* if $g(m) = h(m)$ for all $m \in \mathcal{D}$. (Note that if $\mathcal{D} = \emptyset$, any two functions $g, h : \mathcal{D} \to \mathcal{R}$ are vacuously equivalent.) Let f be any function in $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$. To prove the theorem, it is enough to show that the function defined by $\mathbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f with probability $1/|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|$. We prove this using strong induction on M and N.

Consider the base case where $M = 1$, i.e., $\mathcal{D} = \{m\}$ for some m, and $N \geq M$. When it is first called, $\mathbf{LazySample}(\mathcal{D}, \mathcal{R}, m)$ will determine an element c uniformly

at random from $\mathcal{R}$ and enter it into $F[\mathcal{D}, \mathcal{R}, m]$, whereupon any future calls of $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, m)$ will always output $F[\mathcal{D}, \mathcal{R}, m] = c$. Thus, the output of $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, m)$ is always c, so $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f if and only if $c = f(m)$. Since c is chosen randomly from $\mathcal{R}$, $c = f(m)$ with probability $1/|\mathcal{R}|$. Thus, $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f in this case with probability $1/|\mathcal{R}| = 1/|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|$.

Now suppose $M > 1$, and $N \geq M$. As an induction hypothesis, assume that for all domains $\mathcal{D}'$ of size M' and ranges $\mathcal{R}'$ of size $N' \geq M'$, where either $M' < M$ or ($M' = M$ and $N' < N$), and for any function f' in $\mathsf{OPF}_{\mathcal{D}',\mathcal{R}'}$, $\textbf{LazySample}(\mathcal{D}', \mathcal{R}', \cdot)$ is equivalent to f' with probability $1/|\mathsf{OPF}_{\mathcal{D}',\mathcal{R}'}|$.

When it is first called, $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ sets $I[\mathcal{D}, \mathcal{R}, y]$ to be the value of $\mathrm{HGD}(M, N, y - r; cc)$, where $y = r + \lceil N/2 \rceil$, $r = \min(\mathcal{R}) - 1$. Henceforth, on this and future calls of $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, m)$, the algorithm sets $x = d + I[\mathcal{D}, \mathcal{R}, y - r]$ and runs $\textbf{LazySample}(\mathcal{D}_1, \mathcal{R}_1, m)$ if $m \leq x$, or runs $\textbf{LazySample}(\mathcal{D}_2, \mathcal{R}_2, m)$ if $m > x$, where $\mathcal{D}_1 = \{1, \ldots, x\}$, $\mathcal{R}_1 = \{1, \ldots, y\}$, $\mathcal{D}_2 = \{x + 1, \ldots, M\}$, $\mathcal{R}_2 = \{y + 1, \ldots, N\}$. Let f_1 be f restricted to the domain $\mathcal{D}_1$, and let f_2 be f restricted to the domain $\mathcal{D}_2$. Let x_0 be the unique integer in $\mathcal{D} \cup \{d\}$ such that $f(z) \leq y$ for all $z \in \mathcal{D}$ with $z \leq x_0$, and $f(z) > y$ for all $z \in \mathcal{D}$ with $z > x_0$. Note then that $\textbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f if and only if all three of the following events occur:

E_1: f restricted to range $\mathcal{R}_1$ stays within domain $\mathcal{D}_1$, and f restricted to range $\mathcal{R}_2$ stays within domain $\mathcal{D}_2$—that is, x is chosen to be x_0.

E_2: $\textbf{LazySample}(\mathcal{D}_1, \mathcal{R}_1, \cdot)$ is equivalent to f_1.

E_3: $\textbf{LazySample}(\mathcal{D}_2, \mathcal{R}_2, \cdot)$ is equivalent to f_2.

By the law of conditional probability, and since E_2 and E_3 are independent,

$$\begin{aligned}\Pr[\,E_1 \cap E_2 \cap E_3\,] &= \Pr[\,E_1\,]\,\Pr[\,E_2 \;\cap\; E_3 \mid E_1\,] \\ &= \Pr[\,E_1\,]\,\Pr[\,E_2 \mid E_1\,]\,\Pr[\,E_3 \mid E_1\,]\,.\end{aligned}$$

$\Pr[\,E_1\,]$ is the hypergeometric probability that $\mathrm{HGD}(M, N, y{-}r)$ will return $x_0{-}d$, so

$$\Pr[\,E_1\,] = P_{HGD}(x_0 - d; N, M, \lceil N/2 \rceil) = \frac{\binom{\lceil N/2 \rceil}{x_0 - d}\binom{N - \lceil N/2 \rceil}{M - (x_0 - d)}}{\binom{N}{M}}.$$

Assuming for the moment that neither $\mathcal{D}_1$ nor $\mathcal{D}_2$ are empty, notice that both $|\mathcal{R}_1|$ and $|\mathcal{R}_2|$ are strictly less than $|\mathcal{R}|$, and $|\mathcal{D}_1|$ and $|\mathcal{D}_2|$ are less than or equal to $|\mathcal{D}|$, so the induction hypothesis holds for each. That is, $\mathbf{LazySample}(\mathcal{D}_1, \mathcal{R}_1, \cdot)$ is equivalent to f_1 with probability $1/|\mathsf{OPF}_{\mathcal{D}_1,\mathcal{R}_1}| = 1/\binom{|\mathcal{R}_1|}{|\mathcal{D}_1|}$, and $\mathbf{LazySample}(\mathcal{D}_2, \mathcal{R}_2, \cdot)$ is equivalent to f_2 with probability $1/|\mathsf{OPF}_{\mathcal{D}_2,\mathcal{R}_2}| = 1/\binom{|\mathcal{R}_2|}{|\mathcal{D}_2|}$. Thus, we have that

$$\Pr[\,E_2 \mid E_1\,] = \frac{1}{\binom{\lceil N/2 \rceil}{x_0 - d}} \qquad \text{and} \qquad \Pr[\,E_3 \mid E_1\,] = \frac{1}{\binom{N - \lceil N/2 \rceil}{d + M - x_0}}.$$

Also, note that if $\mathcal{D}_1 = \emptyset$, then $\Pr[\,E_2 \mid E_1\,] = 1 = \frac{1}{\binom{\lceil N/2 \rceil}{x_0 - d}}$ since $x_0 = d$. Likewise, if $\mathcal{D}_2 = \emptyset$, then $\Pr[\,E_3 \mid E_1\,]$ will be the same as above. We conclude that

$$\Pr[\,E_1 \cap E_2 \cap E_3\,] = \frac{\binom{\lceil N/2 \rceil}{x_0 - d}\binom{N - \lceil N/2 \rceil}{M - (x_0 - d)}}{\binom{N}{M}} \cdot \frac{1}{\binom{\lceil N/2 \rceil}{x_0 - d}} \cdot \frac{1}{\binom{N - \lceil N/2 \rceil}{d + M - x_0}} = \frac{1}{\binom{N}{M}}.$$

Thus, $\mathbf{LazySample}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f with probability $\dfrac{1}{\binom{N}{M}} = \dfrac{1}{|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|}$. Since f was an arbitrary element of $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$, the result follows. □

A.3 Proof of Proposition 3.3.3

Proof of Proposition 3.3.3. For the average case bound, we use a result of Chvátal [22] that the tail of the hypergeometric distribution can be bounded so that

$$\sum_{i=k+1}^{M} P_{HGD}(i; N, M, c) \leq e^{-2t^2 M},$$

where t is a fraction such that $0 \leq t \leq 1 - c/N$, and $k = (c/N + t)M$. Taking $c = N/2$, this implies an upper bound on the probability of the hypergeometric distribution assigning our middle domain gap to an "outlying" domain gap:

$$\sum_{i \notin S} P_{HGD}(i; N, M, N/2) \leq 2e^{-2t^2 M} \tag{4}$$

where S is the subdomain $[(1/2 - t)M, (1/2 + t)M]$.

For $M < 12$, after at most 12 calls to **LazySample** we will reach a domain of size 1, and terminate. So suppose that $M \geq 12$. Taking $t = 1/4$ in (4) implies that **LazySample** assigns the middle ciphertext gap to a plaintext gap in the "middle subdomain" $[M/4, 3M/4]$ with probability at least $1 - 2e^{-2(1/4)^2 M} \geq 1 - 2e^{-3/2} > 1/2$. When a domain gap in S is chosen it shrinks the current domain by a fraction of at least $3/4$. So, picking in the middle subdomain $\log_{4/3} M = \frac{\log M}{\log 4/3} < 2.5 \log M$ times will shrink it to size less than 12. Since the probability to pick in the middle subdomain is greater than $1/2$ on each recursive call of **LazySample**, we expect at most $5 \log M$ recursive calls to reach domain size $M < 12$. Therefore, in total at most $5 \log M + 12$ recursive calls are needed on average to map an input domain point. □

A.4 Proof of Proposition 3.4.1

Proof of Proposition 3.4.1. We use a standard hybrid argument, changing the experiment where A has oracle $\texttt{TapeGen}(K, \cdot, \cdot)$ into one with oracle $\mathcal{O}_R(\cdot, \cdot)$ in two steps. First change the former oracle to on input ℓ, x output not $G(\ell, F(K, x))$ but $G(\ell, s)$ for a independent random $s \in \{0,1\}^k$. The change in A's advantage is bounded by $\mathbf{Adv}^{\text{prf}}_F(B_1)$, where B_1 is the PRF adversary against F that runs A, responding to a query ℓ, x by querying its own oracle with x to receive response y, and then returning $G(\ell, y)$ to A. Next change A's oracle to on input ℓ, x return $\mathcal{O}_R(\ell, x)$. This time the change in A's advantage is bounded by $\mathbf{Adv}^{\text{vol-prg}}_G(B_2)$, where B_2 is the VOL-PRG adversary against G that runs A, responding to a query ℓ, x with the response it receives to query ℓ to its own oracle, and the proposition follows. □

A.5 Proof of Proposition 3.4.2

Proof of Proposition 3.4.2. Consider the following adversary.

Adversary $B^{\mathcal{O}(\cdot,\cdot)}$

$i \xleftarrow{\$} [q]$; $\texttt{ctr} \leftarrow 0$

Define $\mathcal{P}$ as the oracle taking query 1^ℓ and running

$\texttt{ctr} \leftarrow \texttt{ctr} + 1$

If $\texttt{ctr} < i$: $s \xleftarrow{\$} \{0,1\}^k$; Return first ℓ bits of $E(s,\langle 1\rangle)\|E(s,\langle 2\rangle)\|\cdots$

If $\texttt{ctr} = i$: $s \xleftarrow{\$} \{0,1\}^k$; Return first ℓ bits of $\mathcal{O}(s,\langle 1\rangle)\|\mathcal{O}(s,\langle 2\rangle)\|\cdots$

If $\texttt{ctr} > i$: $r \xleftarrow{\$} \{0,1\}^\ell$; Return r

$b \xleftarrow{\$} A^{\mathcal{P}(\cdot)}$

Return b

In the PRF experiment, B's oracle $\mathcal{O}$ can be either the blockcipher E or a random function $R : \{0,1\}^k \times \{0,1\}^n \to \{0,1\}^n$. Note that B with oracle E and $i = 0$ emulates A with oracle $G[E]$; while B with oracle R and $i = q$ emulates A with oracle S, where S is the oracle that on input 1^ℓ returns a random string in $\{0,1\}^\ell$. Hence,

$$\begin{aligned} \Pr\left[\, A^{\mathcal{O}_{G[E]}(\cdot)} = 1 \,\right] &= \Pr\left[\, B^{E(\cdot,\cdot)} = 1 \mid i = 1 \,\right] \\ \Pr\left[\, A^{S(\cdot)} = 1 \,\right] &= \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \mid i = k \,\right] \end{aligned} \tag{5}$$

Also, notice that for all $j \in \{2,\ldots,q-1\}$, B with oracle E and $i = j$ has identical behavior to B with oracle R and $i = j - 1$. Thus,

$$\Pr\left[\, B^{E(\cdot,\cdot)} = 1 \mid i = j \,\right] = \Pr\left[\, B^{R(\cdot,\cdot)} \mid i = j - 1 \,\right] \qquad \text{for all } j \in \{2,\ldots,q\}. \tag{6}$$

Then,

$$
\begin{aligned}
&\mathbf{Adv}^{\text{vol-prg}}_{G[E]}(A) \\
&= \Pr\left[\, A^{\mathcal{O}_{G[E]}(\cdot)} = 1 \,\right] - \Pr\left[\, A^{S(\cdot)} = 1 \,\right] \\
&= \Pr\left[\, B^{E(\cdot,\cdot)} = 1 \mid i = 1 \,\right] - \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \mid i = k \,\right] && \text{[by (5)]} \\
&= \sum_{j=1}^{q} \Pr\left[\, B^{E(\cdot,\cdot)} = 1 \mid i = j \,\right] - \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \mid i = j \,\right] && \text{[by (6)]} \\
&= q \sum_{j=1}^{q} \Pr\left[\, B^{E(\cdot,\cdot)} = 1 \mid i = j \,\right] \Pr\left[\, i = j \,\right] - \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \mid i = j \,\right] \Pr\left[\, i = j \,\right] \\
&= q \sum_{j=1}^{q} \Pr\left[\, B^{E(\cdot,\cdot)} = 1 \ \cap \ i = j \,\right] - \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \ \cap \ i = j \,\right] \\
&\leq q \left(\Pr\left[\, B^{E(\cdot,\cdot)} = 1 \,\right] - \Pr\left[\, B^{R(\cdot,\cdot)} = 1 \,\right] \right) \\
&= q \mathbf{Adv}^{\text{prf}}_{E}(B)
\end{aligned}
$$

The efficiency claims should be clear from the definition of B. It is also clear that $G[E]$ is consistent: for $\ell' < \ell$, note that the first ℓ' bits of $G[E](s, 1^{\ell'})$ and $G[E](s, 1^{\ell})$ are the same, as they are just the first ℓ' bits of $E(s, \langle 1 \rangle) \| E(s, \langle 2 \rangle) \| \cdots$. □

A.6 Proof of Theorem 3.4.3

Proof of Theorem 3.4.3. Define adversary B as follows. Given an oracle for either `TapeGen` or a random function with corresponding inputs and output lengths, B runs A and replies to its oracle queries by simulating $\mathcal{E}nc^{\text{HGD}}$ and $\mathcal{D}ec^{\text{HGD}}$ algorithms using its oracle. B then outputs what A outputs. Note that only the procedure `TapeGen` used by these algorithms uses the secret key. B simulates it using its own oracle. We

have

$$\begin{aligned}
&\mathbf{Adv}^{\text{popf-cca}}_{\mathcal{OPE}^{\text{HGD}}[\texttt{TapeGen}]}(A) \\
&= \Pr\left[A^{\mathcal{E}nc^{\text{HGD}}(K,\cdot),\mathcal{D}ec^{\text{HGD}}(K,\cdot)} = 1\right] - \Pr\left[A^{g(\cdot),g^{-1}(\cdot)} = 1\right] \\
&= \Pr\left[A^{\mathcal{E}nc^{\text{HGD}}(K,\cdot),\mathcal{D}ec^{\text{HGD}}(K,\cdot)} = 1\right] - \Pr\left[A^{\textbf{LazySample}(\mathcal{D},\mathcal{R},\cdot),\textbf{LazySampleInv}(\mathcal{D},\mathcal{R},\cdot)} = 1\right] \\
&\leq \mathbf{Adv}^{\text{prf}}_{\texttt{TapeGen}}(B) \ + \ \lambda \ .
\end{aligned}$$

The first equation is by definition. The second equation is due to Theorem 3.3.2. The last inequality is justified by the construction of B, as it simulates $\mathcal{E}nc^{\text{HGD}}$ and $\mathcal{D}ec^{\text{HGD}}$ when given an oracle for `TapeGen` and simulates **LazySample** and **LazySampleInv** when given an oracle for a random function. Above, λ represents an "error term" due to the fact that the "exact" hypergeometric sampling algorithm of [42] technically requires infinite floating-point precision, which is not possible in the real world. One way to bound λ would be to bound the probability that an adversary can distinguish the used HGD sampling algorithm from the ideal (infinite precision) one. B's running time and resources are justified by observing the algorithms and their efficiency analysis. □

A.7 Proof of Theorem 3.5.1

Proof of Theorem 3.5.1. Fix $f \in \mathsf{OPF}_{\mathcal{D},\mathcal{R}}$. As in the proof of Theorem 3.3.2, it is enough to show that the function defined by $\textbf{LazySample}^{\star}(\mathcal{D},\mathcal{R},\cdot)$ is equivalent to f with probability $1/|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|$. We prove this using strong induction on M and N.

Consider the base case where $M = 1$, i.e., $\mathcal{D} = \{m\}$ for some m, and $N \geq M$. When it is first called, $\textbf{LazySample}^{\star}(\mathcal{D},\mathcal{R},m)$ will determine random coins cc, then enter the result of $\text{NHGD}(M,N,m-d;cc)$ into $I[\mathcal{D},\mathcal{R},m]$, whereupon this any future calls of $\textbf{LazySample}^{\star}(\mathcal{D},\mathcal{R},m)$ will always output $F[\mathcal{D},\mathcal{R},m] = c$. Note that by definition, $\text{NHGD}(M,N,m-d;cc)$ returns $f(m)$ with probability

$$P_{NHGD}(f(m)-r;N,1,1) = \frac{\binom{f(m)-r-1}{0}\cdot\binom{N-(f(m)-r)}{0}}{\binom{N}{1}} = \frac{1}{N} = \frac{1}{|\mathcal{R}|}.$$

Thus, the output of $\mathbf{LazySample}^\star(\mathcal{D}, \mathcal{R}, m)$ will always be $f(m)$ with probability $1/|\mathcal{R}|$, implying that $\mathbf{LazySample}^\star(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f in this case with probability $1/|\mathcal{R}| = 1/|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|$.

Now suppose $M > 1$, and $N \geq M$. As an induction hypothesis assume that for all domains $\mathcal{D}'$ of size M' and ranges $\mathcal{R}'$ of size $N' \geq M'$, where either $M' < M$ or $(M' = M$ and $N' < N)$, and for any function f' in $\mathsf{OPF}_{\mathcal{D}',\mathcal{R}'}$, $\mathbf{LazySample}^\star(\mathcal{D}', \mathcal{R}', \cdot)$ is equivalent to f' with probability $1/|\mathsf{OPF}_{\mathcal{D}',\mathcal{R}'}|$.

The first time it is called, $\mathbf{LazySample}^\star(\mathcal{D}, \mathcal{R}, \cdot)$ first computes $I[\mathcal{D}, \mathcal{R}, x]$ from $\mathrm{NHGD}(M, N, x - d; cc)$, where $x = d + \lceil M/2 \rceil$. Henceforth, on this and future calls of $\mathbf{LazySample}^\star(\mathcal{D}, \mathcal{R}, \cdot)$, the algorithm sets $y \leftarrow r + I[\mathcal{D}, \mathcal{R}, x]$, and follows one of three routes: if $x = m$, the algorithm terminates and returns y, if $m < x$ it will return the output of $\mathbf{LazySample}^\star(\mathcal{D}_1, \mathcal{R}_1, m)$, and if if $m > x$ it will return the output of $\mathbf{LazySample}^\star(\mathcal{D}_2, \mathcal{R}_2, m)$, where $\mathcal{D}_1 = \{1, \ldots, x-1\}$, $\mathcal{R}_1 = \{1, \ldots, y-1\}$, $\mathcal{D}_2 = \{x+1, \ldots, M\}$, $\mathcal{R}_2 = \{y+1, \ldots, N\}$. Let f_1 be f restricted to the domain $\mathcal{D}_1$, and let f_2 be f restricted to the domain $\mathcal{D}_2$. Note then that $\mathbf{LazySample}^\star(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f if and only if all three of the following events occur:

E_1: The invocation of $\mathrm{NHGD}(M, N, x - d; cc)$ returns the value $f(x) - r$.

E_2: $\mathbf{LazySample}^\star(\mathcal{D}_1, \mathcal{R}_1, \cdot)$ is equivalent to f_1.

E_3: $\mathbf{LazySample}^\star(\mathcal{D}_2, \mathcal{R}_2, \cdot)$ is equivalent to f_2.

As in the proof of Theorem 3.3.2,

$$\Pr[\, E_1 \cap E_2 \cap E_3 \,] = \Pr[\, E_1 \,] \Pr[\, E_2 \mid E_1 \,] \Pr[\, E_3 \mid E_1 \,] .$$

$\Pr[\, E_1 \,]$ is the negative hypergeometric probability that $\mathrm{NHGD}(M, N, x - d)$ will return $f(x) - r$, which is

$$\Pr[\, E_1 \,] = P_{NHGD}(f(x) - r; N, M, \lceil M/2 \rceil) = \frac{\binom{f(x)-r-1}{\lceil M/2 \rceil - 1}\binom{N-f(x)+r}{M-\lceil M/2 \rceil)}}{\binom{N}{M}}.$$

If E_1 holds, then f_1 is an element of $\mathsf{OPF}_{\mathcal{D}_1,\mathcal{R}_1}$ and f_2 is an element of $\mathsf{OPF}_{\mathcal{D}_2,\mathcal{R}_2}$. By definition, $|\mathcal{R}_1|, |\mathcal{R}_2| < |\mathcal{R}|$, and $|\mathcal{D}_1|, |\mathcal{D}_2| \leq |\mathcal{D}|$, so the induction hypothesis holds for each, and we have that

$$\Pr[\, E_2 \mid E_1 \,] = \frac{1}{\binom{|\mathcal{R}_1|}{|\mathcal{D}_1|}} = \frac{1}{\binom{f(x)-r-1}{\lceil M/2 \rceil - 1}}; \qquad \Pr[\, E_3 \mid E_1 \,] = \frac{1}{\binom{|\mathcal{R}_2|}{|\mathcal{D}_2|}} = \frac{1}{\binom{N-f(x)+r}{M-\lceil M/2 \rceil}}.$$

Thus,

$$\Pr[\, E_1 \cap E_2 \cap E_3 \,] = \frac{\binom{f(x)-r-1}{\lceil M/2 \rceil - 1}\binom{N-f(x)+r}{M-\lceil M/2 \rceil)}}{\binom{N}{M}} \frac{1}{\binom{f(x)-r-1}{\lceil M/2 \rceil - 1}} \frac{1}{\binom{N-f(x)+r}{M-\lceil M/2 \rceil}} = \frac{1}{\binom{N}{M}}.$$

Therefore, $\mathbf{LazySample}^{\star}(\mathcal{D}, \mathcal{R}, \cdot)$ is equivalent to f with probability $\frac{1}{\binom{N}{M}} = \frac{1}{|\mathsf{OPF}_{\mathcal{D},\mathcal{R}}|}$. Since f was an arbitrary element of $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}$, the result follows. □

APPENDIX B

ONE-WAYNESS OF POPFS PROOFS

B.1 Proving Theorem 4.4.1

Before proceeding, we recall certain probabilities relating to the hypergeometric distribution and their connection to random OPFs, as explained in Section 3.3 and Section 3.5. These probabilities will show up at several points in the analysis.

Let $N \geq M$, $0 \leq y \leq N$, $0 \leq x \leq M$. Recall hypergeometric and negative hypergeometric probabilities

$$P_{HGD}(N, M, y, x) = \frac{\binom{y}{x}\binom{N-y}{M-x}}{\binom{N}{M}} ;$$
$$P_{NHGD}(N, M, y, x) = \frac{\binom{y-1}{x-1}\binom{N-y}{M-x}}{\binom{N}{M}} \qquad (x, y \neq 0) .$$

For convenience, define a third, related probability:

$$P_{*}(N, M, y, x) = \frac{\binom{y-1}{x-1}\binom{N-y}{M-x}}{\binom{N-1}{M-1}} \qquad (x, y \neq 0) .$$

As explained in Section 3.5, random OPFs are naturally linked to negative hypergeometric probabilities. We will use the fact that for $(\mathcal{K}_{\mathrm{r}}, \mathcal{E}nc_{\mathrm{r}}, \mathcal{D}ec_{\mathrm{r}}) = \mathsf{ROPF}_{[M],[N]}$ and $m \in [M]$, $c \in [N]$,

$$\Pr_{K \leftarrow \mathcal{K}_{\mathrm{r}}}[\,\mathcal{E}nc_{\mathrm{r}}(K, m) = c\,] = P_{NHGD}(N, M, c, m) .$$

Now, we turn to the proof. The proof relies on two lemmas and a corollary to a third lemma, as follows.

Lemma B.1.1. *For window size r, challenge set size z, and any adversary A, there exists a OW-adversary A' such that*

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \leq z\mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z+1],[N-z+1]}}(A').$$

The proof is in Appendix B.1.1.

Lemma B.1.2. *For any adversary A,*

$$\mathbf{Adv}^{1,1\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \le \frac{1}{N}\sum_{c=1}^{N} P_*(N,M,c,m_c)\ ,$$

where $m_c = \left\lceil \frac{Mc}{N+1} \right\rceil$ for any $c \in [N]$.

The proof is in Appendix B.1.2.

Lemma B.1.3. *Let $N_0 \ge 2M_0$ be (positive) multiples of 2 and let $M = 2^q M_0$ and $N = 2^q N_0$ for integer $q \ge 1$. Define $\alpha_0 = P_*(N_0, M_0, N_0/2, m_{N_0/2})$. Then*

$$\frac{1}{N}\sum_{c=1}^{N} P_*(N,M,c,m_c) < \frac{2}{M} + \alpha_0 \frac{\pi e^{1/M_0}}{2^{q/2}}\ .$$

The proof is in Appendix B.1.3.

Corollary B.1.4. *If $N \ge 2M \ge 32$ and $m_c = \left\lceil \frac{Mc}{N+1} \right\rceil$ for any $c \in [N]$, then*

$$\frac{1}{N}\sum_{c=1}^{N} P_*(N,M,c,m_c) < \frac{4}{\sqrt{M}}\ .$$

Proof. Let $M_0 = 16$. Then $N_0 \ge 32$, and we have

$$\begin{aligned}
\alpha_0 &= P_*(N_0, M_0, N_0/2, M_0/2) \\
&= P_*(N_0, 16, N_0/2, 8) \\
&= \frac{\binom{N_0/2-1}{7}\binom{N_0/2}{8}}{\binom{N_0-1}{15}} \\
&= \frac{(N_0/2-1)\cdots(N_0/2-7)(N_0/2)\cdots(N_0/2-7)15!}{(N_0-1)\cdots(N_0-15)7!8!} \\
&= \frac{N_0(N_0-2)^2(N_0-4)^2\cdots(N_0/2-14)^2 15!}{(N_0-1)\cdots(N_0-15)2^{15}7!8!} \\
&= \frac{N_0(N_0-2)(N_0-4)\cdots(N_0/2-14)15!}{(N_0-1)(N_0-3)\cdots(N_0-15)2^{15}7!8!} \\
&= \left(1+\frac{1}{N_0-1}\right)\left(1+\frac{1}{N_0-3}\right)\cdots\left(1+\frac{1}{N_0-15}\right)\frac{15!}{2^{15}7!8!} \\
&\le \left(1+\frac{1}{31}\right)\left(1+\frac{1}{29}\right)\cdots\left(1+\frac{1}{17}\right)\frac{15!}{2^{15}7!8!} \\
&< 0.278.
\end{aligned}$$

Since $M = 2^q M_0 = 2^{q+4}$, we have $2^{q/2} = \frac{\sqrt{M}}{4}$. Thus,

$$\begin{aligned}\frac{2}{M} + \frac{\pi\alpha_0 e^{1/M_0}}{2^{q/2}} &< \frac{1/\sqrt{16}}{\sqrt{M}} + \frac{4\pi(0.278)e^{1/16}}{\sqrt{M}} \\ &< \frac{4}{\sqrt{M}}\,.\end{aligned}$$

The result then follows from Lemma B.1.3. □

Now, we are ready to prove the main, general result.

Proof of Theorem 4.4.1. Let $M' = M - z + 1$, $N' = N - z + 1$.

$$\begin{aligned}\mathbf{Adv}^{1,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) &\le z\mathbf{Adv}^{1,1\text{-wow}}_{\mathsf{ROPF}_{[M'],[N']}}(A) && \text{(Lemma B.1.1)} \\ &\le z\frac{1}{N'}\sum_{c=1}^{N'} P_*(N', M', c, m_c) && \text{(Lemma B.1.2)} \\ &< z\frac{4}{\sqrt{M'}}. && \text{(Corollary B.1.4)}\end{aligned}$$

In the final step, $N \ge 2M$ and $M \ge 15 + z$ imply $N - z + 1 \ge 2(M - z + 1) \ge 32$. □

B.1.1 Proving Lemma B.1.1

We first introduce a concept related to r, z-WOW security called *specified* r, z-WOW security. The proof then proceeds in two steps. First, we construct an adversary A' whose specified r, z-WOW advantage is at least a factor $1/z$ of the r, z-WOW advantage of A (which, in fact, works for general schemes). In the second step, we exhibit a bijection between OPFs on the space $[M], [N]$ that hit a fixed set $\mathbf{C} \subseteq [N]$ of size $z-1$, and OPFs on the space $[M-z+1], [N-z+1]$. This allows us to construct an efficient r, 1-WOW adversary against $\mathsf{ROPF}_{[M-z+1],[N-z+1]}$ using an efficient specified r, z-WOW adversary against $\mathsf{ROPF}_{[M],[N]}$, with the same advantage. Putting these constructions together yields the result.

An Intermediate Security Definition. The *specified* r, z*-window-one-wayness advantage* of adversary A with respect to scheme $\mathcal{SE}_{\mathcal{D},\mathcal{R}} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ is

$$\mathbf{Adv}^{s\text{-}r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) \;=\; \Pr\left[\,\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) = 1\,\right],$$

where the security experiment is as follows.

Experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A)$

$K \xleftarrow{\$} \mathcal{K}$; $\mathbf{M} \xleftarrow{\$} \mathrm{Cmb}_z^{[M]}$

$m_0 \xleftarrow{\$} \mathbf{M}$; $\mathbf{C} \leftarrow \mathcal{E}nc(K, \mathbf{M})$; $c_0 \leftarrow \mathcal{E}nc(K, m_0)$

$(m_L, m_R) \xleftarrow{\$} A(\mathbf{C}, c_0)$

Return 1 if $(m_R - m_L + 1 \bmod M) \leq r$ and either

$m_0 \in [m_L, m_R]$ or $(m_L > m_R$ and $m_0 \in [m_L, M] \cup [1, m_R])$;

Return 0 otherwise.

The only difference between this experiment and the standard r, z-WOW one is that here, the experiment demands that the adversary return an r-window containing the pre-image of the *specified* ciphertext $c_0 \in \mathbf{C}$ (rather than any ciphertext from $\mathbf{C}$.)

Reducing r, z-WOW Security to Specified r, z-WOW Security for Any Scheme. As our first step, we show that for any efficient r, z-WOW adversary against a general scheme $\mathcal{SE}$, there exists an efficient specified r, z-WOW adversary A' whose success probability is at least a factor of $1/z$ of that of A.

Lemma B.1.5. *For any scheme $\mathcal{SE}_{\mathcal{D},\mathcal{R}}$ and r, z, and any r, z-WOW adversary A, there exists an equally efficient specified r, z-WOW adversary A' such that*

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) \leq z\mathbf{Adv}^{s\text{-}r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A') \ .$$

Proof. Given A, let A' on input $(\mathbf{C}, c)$ run $(m_L, m_R) \xleftarrow{\$} A(\mathbf{C})$ and return (m_L, m_R). Whenever A outputs (m_L, m_R) such that $\exists m \in \mathbf{M}$ with $m \in [m_L, m_R]$ or $(m_L > m_R$ and $m \in [m_L, M] \cup [1, m_R])$, then A' wins if $m = m_0$. Since m_0 is random from $\mathbf{M}$, independent of the rest of the experiment, we conclude that A' wins the specified experiment at least $1/z$ of the times that A wins the standard experiment. The result follows. □

Reducing ROPF Specified r, z-WOW Security to $r, 1$-WOW Security. Now, fix scheme $\mathsf{ROPF}_{[M],[N]} = (\mathcal{K}_{\mathrm{r}}, \mathcal{E}nc_{\mathrm{r}}, \mathcal{D}ec_{\mathrm{r}}) =$ and r, z. It is left to reduce the

success probability of a specified r, z-adversary A against this scheme to that of an $r, 1$-WOW adversary against $\mathsf{ROPF}_{[M-z+1],[N-z+1]}$.

We first introduce a number of notations that will be useful in the proof. Let $z' = z - 1$. For orderable sets $\mathcal{D}, \mathcal{R}$, and $H \subset \mathcal{R}$, let $\mathsf{OPF}_{\mathcal{D},\mathcal{R}}(H)$ denote $\{f \in \mathsf{OPF}_{\mathcal{D},\mathcal{R}} \mid H \subset f(\mathcal{D})\}$, i.e., the set of OPFs from $\mathcal{D}$ to $\mathcal{R}$ with all elements of H in their range. Similarly, for a set U, $n \leq |U|$, and $H \subset U$ with $|H| \leq n$, let $\mathrm{Cmb}_n^U[H]$ denote the set of n-element subsets of U that contain H. For set S with elements $x_1 < x_2 < \ldots x_{|S|}$, and $x \in S$, $H \subseteq S$, $i \in [|S|]$, $I \subseteq [|S|]$, let

$$\mathrm{Idx}_x^S = j \text{ such that } x = x_j \,, \qquad \mathrm{Idxs}_H^S = \{j \mid x_j \in H\} \,,$$
$$\mathrm{Elt}_i^S = x_i \,, \qquad \mathrm{Elts}_I^S = \{x_i \mid i \in I\} \,.$$

Finally, for equal-sized orderable sets S_1, S_2, let $\mathrm{UqOPF}(S_1, S_2)$ be the unique OPF from S_1 to S_2.

The next lemma demonstrates the connection between OPFs in space $[M], [N]$ that hit a certain z'-element subset of $[N]$, and general OPFs in space $[M - z'], [N - z']$.

Lemma B.1.6. *Fix $\mathbf{C} \subseteq [M]$ with $|\mathbf{C}| = z'$. There is a chain of natural bijections between the following sets.*

$$\mathsf{OPF}_{[M],[N]}(\mathbf{C}) \overset{\beta_1}{\rightleftarrows} \mathrm{Cmb}_M^{[N]}[\mathbf{C}] \overset{\beta_2}{\rightleftarrows} \mathrm{Cmb}_{M-z'}^{[N]\setminus\mathbf{C}} \overset{\beta_3}{\rightleftarrows} \mathrm{Cmb}_{[M-z']}^{[N-z']} \overset{\beta_4}{\rightleftarrows} \mathsf{OPF}_{[M-z'],[N-z']}$$

Proof. The bijective functions and their inverses can be defined as follows:

$$\beta_1 : f \mapsto f([M]) \,; \qquad \beta_1^{-1} : S \mapsto \mathrm{UqOPF}([M], S)$$
$$\beta_2 : S \mapsto S \setminus \mathbf{C} \,; \qquad \beta_2^{-1} : S \mapsto S \cup \mathbf{C}$$
$$\beta_3 : S \mapsto \mathrm{Idxs}_S^{[N]\setminus\mathbf{C}} \,; \qquad \beta_3^{-1} : I \mapsto \mathrm{Elts}_I^{[N]\setminus\mathbf{C}}$$
$$\beta_4 : S \mapsto \mathrm{UqOPF}([M - z'], S) \,; \qquad \beta_4^{-1} : f \mapsto f([M - z'])$$

Since all functions are well-defined, the bijections are clear. See Figure 11 for a visual depiction of elements associated through the bijections. □

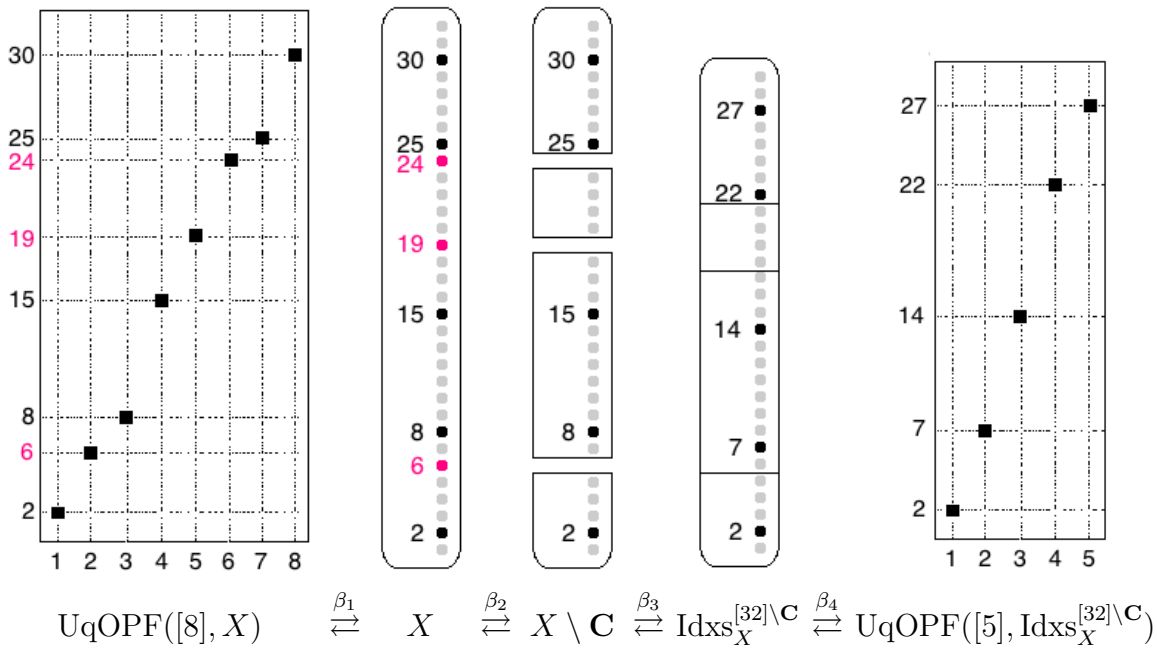

Figure 11: Example of associated elements in the chain of bijections from Lemma B.1.6. In the example, $N = 32$, $M = 8$, $\mathbf{C} = \{6, 19, 24\}$, and we are looking at the particular OPF $f \in \mathsf{OPF}_{[8],[32]}(\mathbf{C})$ with range $X = \{2, 6, 8, 15, 19, 24, 25, 30\}$.

Before we show the final reduction, we state and prove a small lemma.

Lemma B.1.7. *Let* $\mathsf{ROPF}_{[M],[N]} = (\mathcal{K}_{\mathrm{r}}, \mathcal{E}nc_{\mathrm{r}}, \mathcal{D}ec_{\mathrm{r}})$, *and* $z \geq 1$. *Then for any set* $\mathbf{C} \in \mathrm{Cmb}_z^{[N]}$,

$$\Pr_{\substack{K \xleftarrow{\$} \mathcal{K}_{\mathrm{r}} \\ \mathbf{M} \xleftarrow{\$} \mathrm{Cmb}_z^{[M]}}} [\, \mathbf{C} = \{\mathcal{E}nc_{\mathrm{r}}(K, m) \mid m \in \mathbf{M}\} \,] = 1/\binom{N}{z}.$$

Proof. The probability that some $\mathbf{C} \subseteq [N]$ is chosen as the encryptions of the elements of $\mathbf{M}$ is equal to the probability that $\mathcal{E}nc_{\mathrm{r}}(K, \cdot)$ sends *some* z plaintexts $\mathbf{M}' \subseteq [M]$ to $\mathbf{C}$, times the probability that the appropriate $\mathbf{M}$ was picked from $[M]$. The former probability is equal to the likelihood that $\mathbf{C}$ is a subset of a random M-element subset of N, or $\binom{N-z}{M-z}/\binom{N}{M}$. The latter probability is $1/\binom{M}{z}$. Hence, the desired probability

is

$$\frac{\binom{N-z}{M-z}}{\binom{N}{M}}\frac{1}{\binom{M}{z}} = \frac{(N-z)!M!(N-M)!z!(M-z)!}{(M-z)!(N-M)!N!M!}$$
$$= \frac{(N-z)!z!}{N!}$$
$$= 1/\binom{N}{z} .$$

□

Now, here is the second reduction.

Lemma B.1.8. *Fix r, z, M, and N. Let $z' = z - 1$. For any efficient specified r, z-WOW adversary A to scheme $\mathsf{ROPF}_{[M],[N]}$, there exists an efficient $r, 1$-WOW adversary A' to scheme $\mathsf{ROPF}_{[M-z'],[N-z']}$ such that*

$$\mathbf{Adv}^{s\text{-}r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \leq \mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A') .$$

Proof. Let A be an adversary in experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$. We construct a similarly efficient adversary A' to experiment $\mathbf{Exp}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$ using A as follows.

Adversary $A'(\{c'\})$
- $\mathbf{C}' \xleftarrow{\$} \mathrm{Cmb}^{[N]}_{z'}$; $c \leftarrow \mathrm{Elt}^{[N]\setminus\mathbf{C}'}_{c'}$; $\mathbf{C} \leftarrow \mathbf{C}' \cup \{c\}$
- $(m_L, m_R) \xleftarrow{\$} A(\mathbf{C}, c)$
- $z'_- \leftarrow |\{y \in \mathbf{C} \mid y < c\}|$
- $m'_L \leftarrow m_L - z'_-$ $\quad m'_R \leftarrow m_R - z'_-$
- Return (m'_L, m'_R).

Assume that c' is a random ciphertext in $[N - z']$ (as it is in $\mathbf{Exp}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$, by Lemma B.1.7). Then we must show that the input $(\mathbf{C}, c)$ to A accurately mimics the experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$. That is, it must be that $\mathbf{C}$ looks random from $\mathrm{Cmb}^{[N]}_z$ (recalling Lemma B.1.7 applied to the experiment's challenge sets), and c' looks random from $\mathbf{C}$. Note that c looks uniformly random among $[N] \setminus \mathbf{C}'$ because c' is a random index in $[N - z']$ and c is chosen as the (c')th largest element of $[N] \setminus \mathbf{C}$. Hence, A' accurately simulates the experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$.

Let $\beta_1, \ldots, \beta_4$ be as defined in Lemma B.1.6. For any OPF f from $[M]$ to $[N]$ with $\mathbf{C}$ in its range, let

$$\beta_f = (\beta_4 \circ \beta_3 \circ \beta_2 \circ \beta_1)(f)$$

be the associated (unique) OPF from $[M - z']$ to $[N - z']$. For fixed $\mathbf{C}$ and c, let $z'_- = |\{c' \in \mathbf{C} \mid c' < c\}|$. Then note that for any $m \in [M]$, if $f(m) = c \notin \mathbf{C}$ then $\beta_f(m - z'_-) = c - z'_-$ and vice versa.

Thus, if A correctly guesses a window m_L, m_R that succeeds in $\mathbf{Exp}^{s-r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$ when f is chosen as the random OPF, then the output m'_L, m'_R of A' succeeds in $\mathbf{Exp}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$ when β_f is chosen as the random OPF; and the converse is also true. Hence, A and A' have the same advantage in their respective experiments.

We also note that A' is efficient if A is efficient, as the extra steps of sampling an element of $\mathrm{Cmb}^{[N]}_{z'}$ and re-indexing c, m'_L, and m'_R are all efficient operations. □

We are now ready to prove the main lemma of this section.

Proof of Lemma B.1.1. For any r, z, and any efficient r, z-WOW adversary A, there exist efficient algorithms A'', A' such that

$$\begin{aligned}
\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) &\leq z\mathbf{Adv}^{s-r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A'') && \text{(Lemma B.1.5)} \\
&\leq z\mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z+1],[N-z+1]}}(A') && \text{(Lemma B.1.8)}
\end{aligned}$$

The result follows. □

B.1.2 Proving Lemma B.1.2

The proof uses two supporting lemmas. One has already been proved, as Lemma B.1.7 in the special case $z = 1$ establishes that the uniform choice of plaintext in the experiment ensures a uniformly distributed challenge ciphertext. The second lemma, stated next, allows us to calculate the most likely plaintext for a given ciphertext.

Lemma B.1.9. *For fixed $N, M, c \in \mathbb{N}$, $P_{NHGD}(N, M, c, \cdot)$ achieves its maximum over $[M]$ at some*

$$m_0 \in \left[\frac{Mc}{N+1}, \frac{Mc}{N+1} + 1 \right] .$$

In particular, if $N = tM$ for some positive integer t, then $P_{NHGD}(N, M, c, \cdot)$ achieves its maximum over $[M]$ at the unique point

$$m_0 = \lceil Mc/N \rceil = \lceil c/t \rceil .$$

Proof. Suppose that $P_{NHGD}(N, M, c, \cdot)$ achieves its maximum over $[M]$ at m_0. Then the function must have a local maximum there; that is,

$$P_{NHGD}(N, M, c, m_0 - 1) \leq P_{NHGD}(N, M, c, m_0) ,$$
$$P_{NHGD}(N, M, c, m_0) \geq P_{NHGD}(N, M, c, m_0 + 1) .$$

Notice that for $m \geq 2$,

$$\begin{aligned} \frac{P_{NHGD}(N, M, c, m)}{P_{NHGD}(N, M, c, m-1)} &= \frac{\binom{c-1}{m-1}\binom{N-c}{M-m}}{\binom{c-1}{(m-1)-1}\binom{N-c}{M-(m-1)}} \\ &= \frac{(M-(m-1))(c-(m-1))}{(m-1)(N-M-c+m)} \\ &= \frac{\frac{Mc}{m-1} - M - c + m - 1}{N-M-c+m} \end{aligned}$$

is at least 1 if and only if $\frac{Mc}{m-1} - 1 \geq N$, or $m \leq \frac{Mc}{N+1} + 1$; it is at most 1 if and only if $\frac{Mc}{m-1} - 1 \leq N$, or $m - 1 \geq \frac{Mc}{N+1}$. The former implies $m_0 \leq \frac{Mc}{N+1} + 1$; the latter implies $m_0 \geq \frac{Mc}{N+1}$.

So, the maximum value of $P_{NHGD}(N, M, c, m)$ occurs at either a unique point, or two adjacent points in $[M]$. Thus, these local maxima are global maxima, and the necessary condition $\frac{Mc}{N+1} \leq m_0 \leq \frac{Mc}{N+1} + 1$ is also sufficient for m_0 to be a global maximum.

To see the second property, note that $N = tM$ implies

$$\left\lceil \frac{Mc}{N+1} \right\rceil = \left\lceil \frac{c}{t}\left(\frac{N}{N+1}\right) \right\rceil = \left\lceil \frac{c}{t} - \frac{c}{t(N+1)} \right\rceil = \left\lceil \frac{c}{t} \right\rceil ,$$

where the last step is implied by the following: note that the fractional part of c/t is either 0 or at least $1/t$. In either case, subtracting $c/(t(N+1)) < 1/t$ from c/t will not change the value of its ceiling. Also note that in this case, $Mc/(N+1) = Mc/(tM+1)$ is not an integer and thus m_0 is unique. □

Corollary B.1.10. *Fix encryption scheme* $(\mathcal{K}_{\mathrm{r}}, \mathcal{E}nc_{\mathrm{r}}, \mathcal{D}ec_{\mathrm{r}}) = \mathsf{ROPF}_{[M],[N]}$*, and let* $c \in [N]$*. Then* m_c *is a most likely plaintext for* c *if and only if*

$$\frac{Mc}{N+1} \leq m_c \leq \frac{Mc}{N+1} + 1 \ .$$

In particular, if $N = tM$ *for some positive integer* t*, then* m_c *is unique for each* c *and*

$$m_c = \lceil Mc/N \rceil = \lceil c/t \rceil \ .$$

Proof. For any m, c, the probability that $\mathcal{E}nc_{\mathrm{r}}(K, m) = c$ over random $K \in \mathcal{K}_{\mathrm{r}}$ is $P_{NHGD}(N, M, c, m)$. Thus, the result follows directly from Lemma B.1.9. □

We are now ready to prove the lemma.

Proof of Lemma B.1.2. In the one-wayness experiment, notice that an adversary A is not allowed any oracle access, and in fact the only information A receives is the ciphertext c. Thus, given c, the adversary's best recourse is to output the most likely plaintext for c. By Lemma B.1.7, the c given to A is uniform from $[N]$, so the OW advantage of A is bounded above by the average probability (over all $c \in [N]$) that c is the image of its most likely plaintext m_c under random $f \in \mathsf{OPF}_{[M],[N]}$, knowing that c is the image of *some* plaintext under f.

Fix $c \in [N]$. Given that $c \in \{f(m) \mid m \in [M]\}$, the probability that $f(m_c) = c$ is equal to the number of OPFs going through (m_c, c), over the number of OPFs that have a point (x, c) for some $x \in [M]$, or

$$\frac{\binom{c-1}{m_c-1}\binom{N-c}{M-c}}{\binom{N-1}{M-1}} = P_*(N, M, c, m_c) \ .$$

Thus, in the one-wayness experiment, the probability that the (randomly determined) challenge ciphertext is the image of its most likely plaintext is the average of the above quantity for each value of c. □

B.1.3 Proving Lemma B.1.3

The proof proceeds in several steps. Here is an outline:

- Lemma B.1.11 relates the middle ciphertext's most likely plaintext's NHGD probability for a given plaintext/ciphertext space to that of a space twice the size, using an algebraic argument.
- Corollary B.1.12 iterates this result, producing a formula for the middle ciphertext's most likely plaintext's NHGD probability in a large space given the analogous value α_0 in a small space.
- Lemma B.1.13 and Lemma B.1.14 together relate any ciphertext's most likely plaintext's NHGD probability to that of the middle ciphertext in the space, using Stirling's approximation and certain bounds on the gamma function.
- Finally, the proof of Lemma B.1.3 ties these results together, approximating the sum of most likely plaintext NHGD probabilities over the ciphertext space in terms of that of the middle ciphertext, and hence to that of the middle ciphertext in a smaller space.

For readability, we introduce the following notation. For a, b, t positive integers such that $a > b$ and $t < a/b$, let

$$(a)_{[b]} = a(a-1)(a-2)\cdots(a-(b-1))\,;$$
$$(a)_{[b;t]} = a(a-t)(a-2t)\cdots(a-(b-1)t)\,.$$

Approximating most likely NHGD probabilities for the middle ciphertext. Set a domain size M and range size N, larger domain size $M^* = 2M$ and

range size $N^* = 2N$, and consider "middle ciphertexts" $c = N/2$ and $c^* = N^*/2 = N$. We show that if M and $N - M$ are large, then the relative most likely NHGD probabilities for c and c^* (knowing that the ciphertexts are hit) in their respective spaces is approximately equal to the constant $1/\sqrt{2}$.

Lemma B.1.11. *Let N, M be multiples of 2 such that $N \geq 2M$, let $M^* = 2M$ and $N^* = 2N$, and let $c = N/2$ and $c^* = N^*/2 = N$. For any $c \in [N]$, let $m_c = \left\lceil \frac{Mc}{N+1} \right\rceil$. If M is large, then*

$$\frac{P_*(N^*, M^*, c^*, m_{c^*})}{P_*(N, M, c, m_c)} \approx \frac{1}{\sqrt{2}} .$$

In particular,

$$\frac{1}{\sqrt{2}} \lessapprox \frac{P_*(N^*, M^*, c^*, m_{c^*})}{P_*(N, M, c, m_c)} < \frac{1}{\sqrt{2}} \cdot e^{1/(2M)} .$$

Proof. Set $M' = N - M$. Observe that

$$\begin{aligned}
\frac{P_*(N^*, M^*, c^*, m_{c^*})}{P_*(N, M, c, m_c)} &= \frac{\binom{N-1}{M-1}^2 \binom{N}{M}}{\binom{2N-1}{2M-1}\binom{N/2-1}{M/2-1}\binom{N/2}{M/2}} \\
&= \frac{\binom{N}{M}^3}{\binom{2N}{2M}\binom{N/2}{M/2}^2} \\
&= \frac{N!^3 (2M)!(2M')!(M/2)!^2 (M'/2)!^2}{M!^3 (M')!^3 (2N)!(N/2)!^2} \\
&= \frac{N!^2}{(N/2)!^2} \frac{N!}{(2N)!} \frac{(M/2)!^2}{M!^2} \frac{(2M)!}{M!} \frac{(M'/2)!^2}{(M')!^2} \frac{(2M')!}{(M')!} \\
&= \frac{((N)_{[N/2]})^2}{(2N)_{[N]}} \frac{(2M)_{[M]}}{((M)_{[M/2]})^2} \frac{(2M')_{[M']}}{((M')_{[M'/2]})^2} \\
&= \frac{2^N ((N)_{[N/2]})^2}{(2N)_{[N]}} \frac{(2M)_{[M]}}{2^M ((M)_{[M/2]})^2} \frac{(2M')_{[M']}}{2^{M'} ((M')_{[M'/2]})^2} \\
&= \frac{((2N)_{[N/2;2]})^2}{(2N)_{[N]}} \frac{(2M)_{[M]}}{((2M)_{[M/2;2]})^2} \frac{(2M')_{[M']}}{((2M')_{[M'/2;2]})^2} \\
&= \frac{(2N)_{[N/2;2]}}{(2N-1)_{[N/2;2]}} \frac{(2M-1)_{[M/2;2]}}{(2M)_{[M/2;2]}} \frac{(2M'-1)_{[M'/2;2]}}{(2M')_{[M'/2;2]}}
\end{aligned}$$

Define the above quantity to be α. Also, let

$$\beta = \frac{(2N-1)_{[N/2;2]}}{(2N-2)_{[N/2;2]}} \frac{(2M-2)_{[M/2;2]}}{(2M-1)_{[M/2;2]}} \frac{(2M'-2)_{[M'/2;2]}}{(2M'-1)_{[M'/2;2]}}$$
$$\beta' = \frac{(2N+1)_{[N/2;2]}}{(2N)_{[N/2;2]}} \frac{(2M)_{[M/2;2]}}{(2M+1)_{[M/2;2]}} \frac{(2M')_{[M'/2;2]}}{(2M'+1)_{[M'/2;2]}}$$

and notice that for large M and N,

$$\beta \lessapprox \alpha \lessapprox \beta'.$$

On the other hand,

$$\alpha\beta = \frac{2N}{N} \frac{M}{2M} \frac{(N-M)}{(2N-2M)} = 1/2\,,$$

$$\begin{aligned} \alpha\beta' &= \frac{2N+1}{N+1} \frac{M+1}{2M+1} \frac{(N-M+1)}{2N-2M+1} \\ &< 2\left(\frac{1}{2} + \frac{1/2}{2M+1}\right)\left(\frac{1}{2} + \frac{1/2}{2N-2M+1}\right) \\ &< \frac{1}{2}\left(1 + \frac{1}{2M}\right)\left(1 + \frac{1}{2(N-M)}\right) \\ &< \frac{1}{2} e^{1/(2M)+1/(2(N-M))} \\ &< \frac{e^{1/M}}{2}\,. \end{aligned}$$

Hence,

$$\alpha = \sqrt{\alpha^2} \gtrapprox \sqrt{\alpha\beta} = \frac{1}{\sqrt{2}}\,; \qquad \alpha = \sqrt{\alpha^2} \lessapprox \sqrt{\alpha\beta'} < \frac{e^{1/(2M)}}{\sqrt{2}}\,. \qquad \square$$

Now, we can easily approximate most likely NHGD probabilities for middle ciphertexts in large spaces, in the following manner.

Corollary B.1.12. *Let $N_0 \geq 2M_0$ be multiples of 2, let $M = 2^q M_0$ and $N = 2^q N_0$, and let $c = N/2$ and $c_0 = N_0/2$. Define* $\begin{cases} \alpha = P_*(N, M, c, m_c)\,; \\ \alpha_0 = P_*(N_0, M_0, c_0, m_{c_0})\,. \end{cases}$ *Then*

$$\frac{\alpha_0}{2^{q/2}} \lessapprox \alpha < \frac{\alpha_0}{2^{q/2}} e^{1/M_0}.$$

Proof. The left side of the statement directly follows from repeated application of Lemma B.1.11. Similarly, by the lemma,

$$\begin{aligned}\alpha &< \frac{\alpha_0}{2^{q/2}} \prod_{i=1}^{q} e^{1/(2^i M_0)} \\ &= \frac{\alpha_0}{2^{q/2}} e^{(1/M_0)\sum_{i=1}^{q} 2^{-i}} \\ &< \frac{\alpha_0}{2^{q/2}} e^{1/M_0} \ .\end{aligned}$$

□

Relating general most likely NHGD probabilities to that of the middle ciphertext. In this section we show how to approximate most likely NHGD probabilities for any ciphertext in a large space using the probability corresponding to the middle ciphertext.

Recall the definition of the gamma function: for x a real number,

$$\Gamma(x) = \int_0^\infty r^{x-1} e^{-r} \ dr.$$

The gamma function satisfies the following properties, for x real.

$$\Gamma(x+1) = x\Gamma(x) \ ; \qquad\qquad \Gamma(1) = 1 \ .$$

For notational convenience, we will let $\hat{\Gamma}(x) = \Gamma(x+1)$. The above properties imply that $\hat{\Gamma}(x)$ is an extension of the factorial function to real numbers. In particular, for positive integer n,

$$\hat{\Gamma}(n) = \Gamma(n+1) = n!$$

Also, Stirling's approximation applies to Γ: for real $x > 0$,

$$\hat{\Gamma}(x) = \Gamma(x+1) = \sqrt{2\pi x}(x/e)^x e^{\lambda_x} \ ,$$

where

$$\frac{1}{12x+1} < \lambda_x < \frac{1}{12x} \ .$$

We first prove a short lemma that will be used in the next proof.

Lemma B.1.13. *Let M, N be multiples of 2 and $N \geq 2M$. Let $k \in (0,1)$, and $k' = 1-k$. Let $M' = N - M$. Then*

$$\frac{\Gamma(kN)\,\hat{\Gamma}(k'N)\,\Gamma\left(\frac{M}{2}\right)\hat{\Gamma}\left(\frac{M}{2}\right)\hat{\Gamma}^2\left(\frac{M'}{2}\right)}{\Gamma(kM)\,\hat{\Gamma}(kM')\,\hat{\Gamma}(k'M)\,\hat{\Gamma}(k'M')\,\Gamma\left(\frac{N}{2}\right)\hat{\Gamma}\left(\frac{N}{2}\right)} \leq \frac{1}{2\sqrt{kk'}}.$$

Proof. Using Stirling's approximation,

$$\begin{aligned}
&\frac{\Gamma(kN)\,\hat{\Gamma}(k'N)\,\Gamma\left(\frac{M}{2}\right)\hat{\Gamma}\left(\frac{M}{2}\right)\hat{\Gamma}^2\left(\frac{M'}{2}\right)}{\Gamma(kM)\,\hat{\Gamma}(kM')\,\hat{\Gamma}(k'M)\,\hat{\Gamma}(k'M')\,\Gamma\left(\frac{N}{2}\right)\hat{\Gamma}\left(\frac{N}{2}\right)} \\
&= \frac{kM\frac{N}{2}}{kN\frac{M}{2}} \frac{\hat{\Gamma}(kN)\,\hat{\Gamma}(k'N)\,\hat{\Gamma}\left(\frac{M}{2}\right)\hat{\Gamma}\left(\frac{M}{2}\right)\hat{\Gamma}^2\left(\frac{M'}{2}\right)}{\hat{\Gamma}(kM)\,\hat{\Gamma}(kM')\,\hat{\Gamma}(k'M)\,\hat{\Gamma}(k'M')\,\hat{\Gamma}^2\left(\frac{N}{2}\right)} \\
&= e^{\lambda} \cdot \sqrt{\frac{kNk'N\left(\frac{M}{2}\right)^2\left(\frac{N-M}{2}\right)^2}{kMkM'k'Mk'M'\left(\frac{N}{2}\right)^2}} \cdot \frac{(kN)^{kN}(k'N)^{k'N}\left(\frac{M}{2}\right)^{2M/2}\left(\frac{M'}{2}\right)^{2M'/2}}{(kM)^{kM}(kM')^{kM'}(k'M)^{k'M}(k'M')^{k'M'}\left(\frac{N}{2}\right)^{2N/2}} \\
&= e^{\lambda} \cdot \frac{1}{2\sqrt{kk'}}\,,
\end{aligned}$$

where

$$\begin{aligned}
\lambda &= \lambda_{kN} + \lambda_{k'N} + 2\lambda_{M/2} + 2\lambda_{M'/2} - \lambda_{kM} - \lambda_{kM'} - \lambda_{k'M} - \lambda_{k'M'} - 2\lambda_{N/2} \\
&\approx \frac{1}{12}\left(\frac{1}{kN} + \frac{1}{k'N} + \frac{2}{M/2} + \frac{2}{M'/2}\right) - \frac{1}{12}\left(\frac{1}{kM} + \frac{1}{kM'} + \frac{1}{k'M} + \frac{1}{k'M'} + \frac{2}{N/2}\right) \\
&= \frac{1}{12}\left(\frac{1}{M} + \frac{1}{M'} - \frac{1}{N}\right)\left(4 - \frac{1}{k} - \frac{1}{k'}\right) \\
&< \frac{1}{6M}\left(4 - \frac{1}{kk'}\right) \\
&\leq 0\,,
\end{aligned}$$

since the maximum value of $kk' = k(1-k)$ for $k \in (0,1)$ is $1/4$. □

Now, we provide a bound on the ratio between the most likely plaintext probability of a ciphertext c, with $1/M \leq c \leq (M-1)/M$, versus that of the middle ciphertext.

Lemma B.1.14. *Let M, N be multiples of 2 and $N \geq 2M$. Let k be a multiple of $1/N$ such that $1/M \leq k \leq M - 1/M$. Then*

$$\frac{P_*(N, M, kN, m_{kN})}{P_*(N, M, N/2, m_{N/2})} \leq \frac{1}{\sqrt{k(1-k)}}\,.$$

Proof. Let $k' = 1-k$ and $M' = N-M$. We use the following bounds of D. Kershaw [45]: for $x > 0$ and $0 < s < 1$,

$$\left(x+\frac{s}{2}\right)^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < \left(x-\frac{1}{2}+\left(s+\frac{1}{4}\right)^{1/2}\right)^{1-s}$$

Rewriting the bounds, for $y > 1$ and $0 < \delta < 1$, we have

$$\frac{\Gamma(y)}{\Gamma(y-\delta)} < \left(y-\frac{3}{2}+\left(\frac{5}{4}-\delta\right)^{1/2}\right)^{\delta}; \qquad \frac{\Gamma(y-\delta)}{\Gamma(y)} < \left(y-\frac{\delta}{2}-\frac{1}{2}\right)^{-\delta}.$$

Let $\epsilon = \lceil kM \rceil - kM$. By Lemma B.1.10, $m_{kN} = \lceil kM \rceil = kM + \epsilon$. Then using Lemma B.1.13,

$$\begin{aligned}
&\frac{P_*(N,M,kN,m_{kN})}{P_*(N,M,\frac{N}{2},m_{\frac{N}{2}})} \\
&= \frac{\binom{kN-1}{kM+\epsilon-1}\binom{k'N}{k'M-\epsilon}}{\binom{N/2-1}{M/2-1}\binom{N/2}{M/2}} \\
&= \frac{\Gamma(kN)\hat{\Gamma}(k'N)\Gamma\left(\frac{M}{2}\right)\hat{\Gamma}\left(\frac{M}{2}\right)\hat{\Gamma}^2\left(\frac{M'}{2}\right)}{\Gamma(kM+\epsilon)\hat{\Gamma}(kM'-\epsilon)\hat{\Gamma}(k'M-\epsilon)\hat{\Gamma}(k'M'+\epsilon)\Gamma\left(\frac{N}{2}\right)\hat{\Gamma}\left(\frac{N}{2}\right)} \\
&\le \frac{1}{2\sqrt{kk'}}\frac{\Gamma(kM)}{\Gamma(kM+\epsilon)}\frac{\hat{\Gamma}(kM')}{\hat{\Gamma}(kM'-\epsilon)}\frac{\hat{\Gamma}(k'M)}{\hat{\Gamma}(k'M-\epsilon)}\frac{\hat{\Gamma}(k'M')}{\hat{\Gamma}(k'M'+\epsilon)} \\
&= \frac{1}{2\sqrt{kk'}}\left(\frac{\left(kM'-\frac{1}{2}+\sqrt{\frac{5}{4}-\epsilon}\right)\left(k'M-\frac{1}{2}+\sqrt{\frac{5}{4}-\epsilon}\right)}{\left(kM+\frac{\epsilon-1}{2}\right)\left(k'M'+\frac{\epsilon+1}{2}\right)}\right)^{\epsilon} \\
&= \frac{1}{2\sqrt{kk'}}\left(\frac{\left(k-\frac{1}{2M'}+\frac{1}{M'}\sqrt{\frac{5}{4}-\epsilon}\right)}{\left(k+\frac{\epsilon-1}{2M}\right)}\frac{\left(k'-\frac{1}{2M}+\frac{1}{M}\sqrt{\frac{5}{4}-\epsilon}\right)}{\left(k'+\frac{\epsilon+1}{2M'}\right)}\right)^{\epsilon} \\
&= \frac{1}{2\sqrt{kk'}}\left(1+\frac{\frac{1}{M'}\left(\sqrt{\frac{5}{4}-\epsilon}-\frac{1}{2}\right)+\frac{1-\epsilon}{2M}}{k+\frac{\epsilon-1}{2M}}\right)^{\epsilon}\left(1+\frac{\frac{1}{M}\left(\sqrt{\frac{5}{4}-\epsilon}-\frac{1}{2}\right)-\frac{\epsilon+1}{2M'}}{k'+\frac{\epsilon+1}{2M'}}\right)^{\epsilon} \\
&= \frac{1}{2\sqrt{kk'}}\left(1+\frac{\frac{M}{M'}\left(2\sqrt{\frac{5}{4}-\epsilon}-1\right)+1-\epsilon}{2Mk+\epsilon-1}\right)^{\epsilon}\left(1+\frac{\left(2\sqrt{\frac{5}{4}-\epsilon}-1\right)-\frac{M}{M'}(\epsilon+1)}{2Mk'+\frac{M}{M'}(\epsilon+1)}\right)^{\epsilon} \\
&< \frac{1}{2\sqrt{kk'}}\left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}-\epsilon}{2Mk+\epsilon-1}\right)^{\epsilon}\left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}}{2M(1-k)}\right)^{\epsilon} \\
&= \frac{g(\epsilon,k,M)}{2\sqrt{kk'}},
\end{aligned}$$

where $g(\epsilon,k,M) = \left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}-\epsilon}{2Mk+\epsilon-1}\right)^{\epsilon}\left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}}{2M(1-k)}\right)^{\epsilon}$. Notice that the bounds on k imply $Mk \geq 1$ and $M(1-k) \geq 1$, so

$$g(\epsilon,k,M) < f(\epsilon) = \left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}-\epsilon}{1+\epsilon}\right)^{\epsilon}\left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}}{2}\right)^{\epsilon},$$

which, for $\epsilon \in [0,1]$, is bounded by 2, as can be seen in Figure 12.

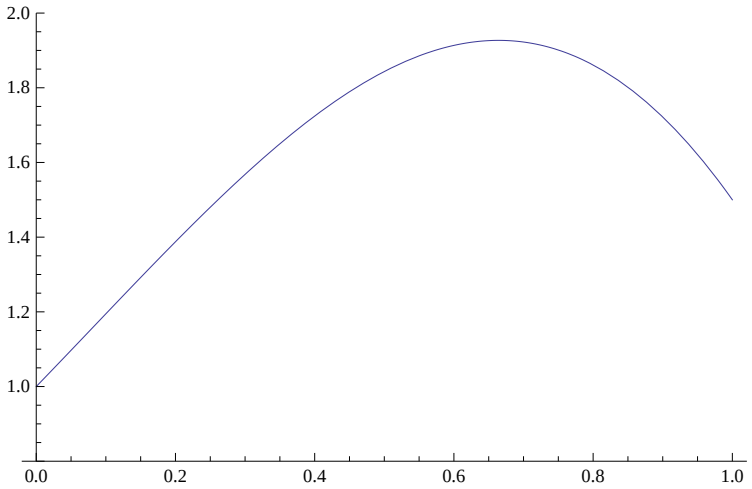

Figure 12: Graph of $f(\epsilon) = \left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}-\epsilon}{1+\epsilon}\right)^{\epsilon}\left(1+\frac{2\sqrt{\frac{5}{4}-\epsilon}}{2}\right)^{\epsilon}$ for $\epsilon \in [0,1]$.

Therefore,

$$\frac{P_*(N,M,kN,m_{kN})}{P_*(N,M,\frac{N}{2},m_{\frac{N}{2}})} < \frac{1}{\sqrt{k(1-k)}}.$$ □

As a side note, the maximum value of $f(\epsilon)$ for $\epsilon \in [0,1]$ in the above proof is slightly lower than 2. Mathematica finds a bound of approximately 1.927:

```
In[83]:= f[x]=(1+(2(1.25-x)^0.5 -x)/(1+x))^x (1+(2(1.25-x)^0.5)/2)^x;
         Maximize[{f[x],0<=x<=1},x]
Out[84]= {1.92692,{x->0.664124}}
```

Also, one could try to bound the value of $g(\epsilon,k,M)$ in the proof (which is more difficult to do computationally) to achieve a tighter bound; empirical evidence shows it can be made very close to 1. Mathematica can handle this if we place some lower bound on M. For instance, forcing $M > 1000000$:

```
In[85]:= g[x,k,M]=(1.0+(2(1.25-x)^(0.5)-x)/(2 k M+x-1))^x (1.0+(2(1.25-x)^(0.5))/(2(1-k) M))^x;
         FindMaximum[{g[x,k,M],{0<= x<= 1,M>1000000,1.0<= M k<= M-1}},{x,{k,0.5},M}]
Out[86]= {1.,{x->0.420798,k->0.500004,M->1.03175*10^6}}
```

We are content, however, to proceed with the looser upper bound of 2.

The preceding results can now be put together to prove the main lemma statement.

Proof of Lemma B.1.3. By Lemma B.1.14 and Corollary B.1.12,

$$\begin{aligned}
\frac{1}{N}\sum_{c=1}^{N} P_*(N,M,c,m_c) &< \frac{2}{M} + \frac{1}{N}\sum_{c=N/M}^{N-N/M} P_*(N,M,c,m_c) \\
&< \frac{2}{M} + \frac{P_*(N,M,N/2,m_{N/2})}{N}\sum_{c=N/M}^{N-N/M} \frac{1}{\sqrt{(c/N)(1-c/N)}} \\
&< \frac{2}{M} + P_*(N,M,N/2,m_{N/2}) \cdot \int_0^1 \frac{1}{\sqrt{x(1-x)}}\, dx \\
&= \frac{2}{M} + P_*(N,M,N/2,m_{N/2}) \arcsin(2x-1)\Big|_0^1 \\
&= \frac{2}{M} + P_*(N,M,N/2,m_{N/2}) \cdot \pi \\
&< \frac{2}{M} + \alpha_0 \frac{\pi e^{1/M_0}}{2^{q/2}}\ .
\end{aligned}$$

□

B.2 Comparing tight and simple bounds

In Table 5, we compare the tight bound of Lemma B.1.2 and the simple bound of Lemma B.1.3 for several values of M and N and see that the results are close. We have separated the constant factor 2 in each simple bound to illustrate how close the bounds would be if the factor 2 were improved to 1 (as described after the proof of Lemma B.1.14.)

Table 5: Sample evaluation of tight vs. simple bounds. For the simple bounds, $M_0 = 2^6$.

M	N	Tight	Simple
2^8	2^{16}	0.077	$0.087 \cdot 2$
2^9	2^{17}	0.055	$0.060 \cdot 2$
2^{10}	2^{18}	0.039	$0.042 \cdot 2$

B.3 Proving Theorem 4.4.2

The first half of the result,

$$\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \geq \mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A),$$

is obvious, as giving the adversary more challenge ciphertexts can only help it win. It is left to prove the bound on $\mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$.

We use the following notation for the tail probabilities of the hypergeometric distribution.

$$H_+(c,N,M,m_0) = \sum_{m=m_0}^{M} P_{HGD}(N,M,c,m),$$
$$H_-(c,N,M,m_0) = \sum_{m=0}^{m_0} P_{HGD}(N,M,c,m).$$

The proof of the theorem appears after a lemma.

Lemma B.3.1. *Let M, N, $c \in [N]$, and $r \in [M]$ be given. Let $\delta = \frac{r-1}{2M}$, and let $m_L, m_R \in [M]$ be defined as* $\begin{cases} m_L = \max\{m_c - \lfloor \delta M \rfloor, 1\}, \\ m_R = \min\{m_c + \lfloor \delta M \rfloor, M\}, \end{cases}$ *where $m_c = \left\lceil \frac{Mc}{N+1} \right\rceil$. Then*

$$\sum_{m=1}^{m_L-1} P_*(N,M,c,m) \leq e^{-2\delta^2(M-1)} \quad \text{and} \quad \sum_{m=m_R+1}^{M} P_*(N,M,c,m) \leq e^{-2\delta^2(M-1)}.$$

Proof. We will use a bound by Chvátal [22] on the upper tail of the hypergeometric distribution:

$$H_+\left(c,N,M,\left(\frac{c}{N}+d\right)M\right) \leq e^{-2d^2M}.$$

Chvátal's upper tail bound implies a similar lower tail bound:

$$\begin{aligned}
H_-\left(c, N, M, \left(\frac{c}{N} - d\right) M\right) &= \sum_{i=0}^{(c/N-d)M} P_{HGD}(N, M, c, i) \\
&= \sum_{i=0}^{(c/N-d)M} P_{HGD}(N, M, N-c, M-i) \\
&= \sum_{j=M-(c/N-d)M}^{M} P_{HGD}(N, M, N-c, j) \\
&= H_+\left(N-c, N, M, \left(\frac{N-c}{N} + d\right) M\right) \\
&\leq e^{-2d^2 M}.
\end{aligned}$$

Notice that $m_R \geq \frac{cM}{N-1} + \delta M \geq \left(\frac{c-1}{N-1} + \delta\right)(M-1)$. So

$$\begin{aligned}
\sum_{m=m_R+1}^{M} P_*(N, M, c, m) &= \sum_{m=m_R+1}^{M} \frac{\binom{c-1}{m-1}\binom{N-c}{M-m}}{\binom{N-1}{M-1}} \\
&= \sum_{m=m_R}^{M-1} \frac{\binom{c-1}{m}\binom{N-c}{M-1-m}}{\binom{N}{M}} \\
&= H_+(c-1, N-1, M-1, m_R) \\
&\leq H_+\left(c-1, N-1, M-1, \left(\frac{c-1}{N-1} + \delta\right)(M-1)\right) \\
&\leq e^{-2\delta^2(M-1)}.
\end{aligned}$$

Similarly, $m_L - 2 \leq \frac{cM}{N-1} - \delta M - 1 \leq \left(\frac{c-1}{N-1} - \delta\right)(M-1)$. So

$$\begin{aligned}
\sum_{m=1}^{m_L-1} P_*(N, M, c, m) &= \sum_{m=1}^{m_L-1} \frac{\binom{c-1}{m-1}\binom{N-c}{M-m}}{\binom{N-1}{M-1}} \\
&= \sum_{m=0}^{m_L-2} \frac{\binom{c-1}{m}\binom{N-c}{M-1-m}}{\binom{N}{M}} \\
&= H_-(c-1, N-1, M-1, m_L-2) \\
&\leq H_-\left(c-1, N-1, M-1, \left(\frac{c-1}{N-1} - \delta\right)(M-1)\right) \\
&\leq e^{-2\delta^2(M-1)}.
\end{aligned}$$

□

We now prove the theorem.

Proof of Theorem 4.4.2. As already mentioned, the first inequality of the theorem is trivially true. It is left to prove the second inequality.

Consider the following r, 1-WOW adversary A.

Adversary $A(\{c\})$

$m_c \leftarrow \lceil \frac{Mc}{N+1} \rceil$

$\delta \leftarrow \frac{r-1}{2M}$

$m_L \leftarrow \max\{m_c - \lfloor \delta M \rfloor, 1\}$

$m_R \leftarrow \min\{m_c + \lfloor \delta M \rfloor, M\}$

Return (m_L, m_R)

(m_L, m_R) is a legal response in the r, 1-WOW experiment since the associated window has size $m_R - m_L + 1 \leq 2\delta M + 1 \leq r$. The probability that the adversary succeeds is the probability that $c \in [m_L, m_R]$, or

$$\sum_{m=m_L}^{m_R} P_*(N, M, c, m) \geq 1 - 2e^{-\frac{(r-1)^2}{2}\frac{(M-1)}{M^2}},$$

where the inequality follows from Lemma B.3.1. Since A only performs efficient operations, the result follows. □

B.4 Proving Theorem 4.4.3

The proof of the theorem parallels that of Theorem 4.4.1. As such, it requires several intermediate results that we now state.

Lemma B.4.1. *For window size r, challenge set size z, and any adversary A, there exists a OW-adversary A' such that*

$$\mathbf{Adv}^{r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \leq z(z-1)\mathbf{Adv}^{r,2\text{-wdow}}_{\mathsf{ROPF}_{[M-z+2],[N-z+2]}}(A') \ .$$

The proof is in Appendix B.4.1.

Lemma B.4.2. *For any adversary A,*

$$\mathbf{Adv}^{1,2\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \le \frac{1}{N-1}\sum_{w=1}^{N-1} P_*(N-1, M-1, w, d_w) \ ,$$

where $d_w = \left\lceil \frac{(M-1)w}{N} \right\rceil$.

The proof is in Appendix B.5.1.

Notice that the bound in Lemma B.4.2 is precisely the bound in Lemma B.1.2, only with parameters $M-1$, $N-1$ instead of M, N. Thus, we will be able to use the simple bound from Corollary B.1.4.

The proof of the theorem now easily follows.

Proof of Theorem 4.4.3. Let $M' = M - z + 2$, $N' = N - z + 2$.

$$\begin{aligned}
\mathbf{Adv}^{1,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A) &\le z(z-1)\mathbf{Adv}^{1,2\text{-wow}}_{\mathsf{ROPF}_{[M'],[N']}}(A) && \text{(Lemma B.4.1)}\\
&\le z(z-1)\frac{1}{N'-1}\sum_{c=1}^{N'-1} P_*(N'-1, M'-1, w, d_w) && \text{(Lemma B.4.2)}\\
&< z(z-1)\frac{4}{\sqrt{M'-1}} && \text{(Corollary B.1.4)}\\
&= z(z-1)\frac{4}{\sqrt{M-z+1}}.
\end{aligned}$$

In the third step, $N \ge 2M$ and $M \ge 15 + z$ imply $N' - 1 \ge 2(M'-1) \ge 32$. □

B.4.1 Proving Lemma B.4.1

Define *specified r, z-window-distance-one-wayness advantage* of adversary A with respect to scheme $\mathcal{SE}_{\mathcal{D},\mathcal{R}} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ as

$$\mathbf{Adv}^{s\text{-}r,z\text{-wdow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) \ = \ \Pr\left[\mathbf{Exp}^{s\text{-}r,z\text{-wdow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) = 1\right] ,$$

where the security experiment is as follows.

Experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A)$

$K \xleftarrow{\$} \mathcal{K}$; $\mathbf{M} \xleftarrow{\$} \mathrm{Cmb}_z^{[M]}$

$m_0 \xleftarrow{\$} \mathbf{M}$; $m_1 \xleftarrow{\$} \mathbf{M} \setminus \{m_0\}$

$\mathbf{C} \leftarrow \mathcal{E}nc(K, \mathbf{M})$; $(c_0, c_1) \leftarrow \mathcal{E}nc(K, (m_0, m_1))$

$(d_L, d_R) \xleftarrow{\$} A(\mathbf{C}, c_0, c_1)$

Return 1 if $d_2 - d_1 + 1 \leq r$ and $m_1 - m_0 \bmod M \in [d_1, d_2]$;

Return 0 otherwise.

Lemma B.4.3. *For any scheme $\mathcal{SE}_{\mathcal{D},\mathcal{R}}$ and r, z, and any r, z-WDOW adversary A, there exists an equally efficient specified r, z-WDOW adversary A' such that*

$$\mathbf{Adv}^{r,z\text{-wdow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A) \leq z(z-1)\mathbf{Adv}^{s\text{-}r,z\text{-wdow}}_{\mathcal{SE}_{\mathcal{D},\mathcal{R}}}(A').$$

Proof. Given A, let A' on input $(\mathbf{C}, c_0, c_1)$ simply run $(d_L, d_R) \xleftarrow{\$} A(\mathbf{C})$ and return (d_L, d_R). Whenever A outputs legal (d_L, d_R) such that $\exists m'_0, m'_1 \in [M]$ with $m'_1 - m'_0 \bmod M \in [d_1, d_2]$, then A' wins if $m_0 = m'_0$ and $m_1 = m'_1$. Since m_0 is random in $[M]$ and m_1 is random in $[M] \setminus \{m_0\}$, independent of the rest of the experiment, we conclude that A' wins the specified experiment at least $\frac{1}{z(z-1)}$ of the times that A wins the standard experiment. The lemma follows. □

Lemma B.4.4. *Fix r, z, M, and N. Let $z' = z - 2$. For any efficient specified r, z-WDOW adversary A to scheme $\mathsf{ROPF}_{[M],[N]}$, there exists an efficient $r, 2$-WDOW adversary A' to scheme $\mathsf{ROPF}_{[M-z'],[N-z']}$ such that*

$$\mathbf{Adv}^{s\text{-}r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) \leq \mathbf{Adv}^{r,2\text{-wdow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A').$$

Proof. Let A be an adversary to experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$. We construct an adversary A' to experiment $\mathbf{Exp}^{r,2\text{-wdow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$ using A as follows.

Adversary $A'(\{c'_0, c'_1\})$

$\mathbf{C}' \xleftarrow{\$} \mathrm{Cmb}_{z'}^{[N]}$

$c_i \leftarrow \mathrm{Elt}_{c'_i}^{[N]\setminus\mathbf{C}'}$ for $i = 0, 1$

$\mathbf{C} \leftarrow \mathbf{C}' \cup \{c_0, c_1\}$

$(d_L, d_R) \xleftarrow{\$} A(\mathbf{C}, c_0, c_1)$

$z'_{\mathrm{bt}} \leftarrow |\{y \in \mathbf{C} \mid c_0 < y < c_1\}|$

$d'_L \leftarrow d_L - z'_{\mathrm{bt}} \quad d'_R \leftarrow d_R - z'_{\mathrm{bt}}$

Return (d'_L, d'_R).

Assume that c'_0, c'_1 are random (distinct) ciphertexts in $[N - z']$ (as they are in the experiment $\mathbf{Exp}^{r,2\text{-wdow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$, by Lemma B.1.7). Then we must show that the input $(\mathbf{C}, c_0, c_1)$ to A accurately mimics the experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$. That is, it must be that $\mathbf{C}$ looks random from $\mathrm{Cmb}_z^{[N]}$ (recalling Lemma B.1.7 applied to the experiment's challenge sets), and $\{c_0, c_1\}$ looks random from $\mathrm{Cmb}_2^{\mathbf{C}}$. Note that c_0, c_1 are uniformly random distinct elements of $[N] \setminus \mathbf{C}$ because c'_0, c'_1 are random distinct indices in $[N - z']$ and c_i is chosen as the (c'_i)th largest element of $[N] \setminus \mathbf{C}$ for $i = 0, 1$. Hence, $\mathbf{C}$ looks random from $\mathrm{Cmb}_z^{[N]}$ and $\{c_0, c_1\}$ looks random from $\mathrm{Cmb}_2^{\mathbf{C}}$. Thus, A' accurately simulates the experiment $\mathbf{Exp}^{s\text{-}r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$.

Let $\beta_1, \ldots, \beta_4$ be as defined in Lemma B.1.6. For any OPF f from $[M]$ to $[N]$ with $\mathbf{C}$ in its range, let

$$\beta_f = (\beta_4 \circ \beta_3 \circ \beta_2 \circ \beta_1)(f)$$

be the associated (unique) OPF from $[M - z']$ to $[N - z']$. Let $z'_{\mathrm{bt}} = |\{c' \in \mathbf{C} \mid c_0 < c' < c_1\}|$. Then note that for any $m_0, m_1 \in [M]$, if $f(m_0) = c_0 \notin \mathbf{C}$ and $f(m_1) = c_1 \notin \mathbf{C}$ then $m_1 - m_0 = \beta_f(m_1) - \beta_f(m_0) + z'_{\mathrm{bt}}$ and vice versa.

Thus, if A correctly guesses a window d_L, d_R that succeeds in $\mathbf{Exp}^{s\text{-}r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A)$ when f is chosen as the random OPF, then the output d'_L, d'_R of A' succeeds in $\mathbf{Exp}^{r,2\text{-wdow}}_{\mathsf{ROPF}_{[M-z'],[N-z']}}(A')$ when β_f is chosen as the random OPF; and the converse is also true. Hence, A and A' have the same advantage in their respective experiments.

We also note that A' is efficient if A is efficient, as the extra steps of sampling an element of $\mathrm{Cmb}_{z'}^{[N]}$ and re-indexing c_0, c_1, d'_L, and d'_R are all efficient operations. $\square$

We are now ready to prove the main lemma of this section.

Proof of Lemma B.4.1. For any r, z, and any efficient r, z-WDOW adversary A, there exist efficient algorithms A'', A' such that

$$\begin{aligned}
\mathbf{Adv}^{r,z\text{-wdow}}_{\mathsf{ROPF}_{[M],[N]}}(A) &\leq z\mathbf{Adv}^{s-r,z\text{-wow}}_{\mathsf{ROPF}_{[M],[N]}}(A'') && \text{(Lemma B.4.3)} \\
&\leq z\mathbf{Adv}^{r,1\text{-wow}}_{\mathsf{ROPF}_{[M-z+1],[N-z+1]}}(A') && \text{(Lemma B.4.4)}
\end{aligned}$$

The result follows. $\square$

B.5 Proof of Theorem 4.4.4

Proof of Theorem 4.4.4. As in Theorem 4.4.2, the first inequality is trivially true. It is left to prove the second inequality, which we do by constructing an $r, 2$-WDOW adversary A as follows.

Adversary $A(\{c_1, c_2\})$

$w \leftarrow c_2 - c_1 \bmod N$

$d_w \leftarrow \left\lceil \frac{(M-1)w}{N} \right\rceil$

$\delta \leftarrow \frac{r-1}{2(M-1)}$

$d_L \leftarrow \max\{d_w - \lfloor \delta(M-1) \rfloor, 1\}$

$d_R \leftarrow \min\{d_w + \lfloor \delta(M-1) \rfloor, M-1\}$

Return (d_L, d_R).

(d_L, d_R) is a legal response in the $r, 2$-WDOW experiment since the associated window has size $d_R - d_L + 1 \leq 2\delta(M-1) + 1 \leq r$.

Note that $d_w = \left\lceil \frac{(M-1)w}{N} \right\rceil$ is the most likely plaintext distance between c_1 and c_2 by Corollary B.5.2. The probability that the adversary succeeds in the $r, 2$-WDOW experiment is the probability that $m_2 - m_1 \bmod M = d \in [d_L, d_R]$, or

$$\sum_{d=d_L}^{d_R} P_*(N-1, M-1, w, d) \geq 1 - 2e^{-\frac{(r-1)^2}{2}\frac{(M-2)}{(M-1)^2}},$$

by Lemma B.3.1. Since A only performs efficient operations, the result follows. $\square$

B.5.1 Proving Lemma B.4.2

A formula for the most likely plaintext distance. In Corollary B.5.2 below, we derive a formula for the most likely plaintext distance between two given ciphertexts. But first, the following lemma determines the probability that a given ciphertext pair corresponds to a given plaintext distance, which is used to prove the corollary.

Lemma B.5.1. *For $c_1, c_2 \in [N]$, let $\mathsf{OPF}_{[M],[N]}{}^* = \{f \in \mathsf{OPF}_{[M],[N]} : c_1, c_2 \in f([M])\}$. Then for any $d \in [M-1]$,*

$$\Pr_{f \xleftarrow{\$} \mathsf{OPF}_{[M],[N]}{}^*} \left[f^{-1}(c_2) - f^{-1}(c_1) \bmod M = d \right] = P_*(N-1, M-1, w, d) \ ,$$

where $w = c_2 - c_1 \bmod N$.

Proof. Let $c_1, c_2 \in [N]$. If $c_1 = c_2$, the result easily follows, so suppose $c_1 \neq c_2$. c_1 and c_2 partition the rest of the ciphertext space into two sets S and S':

$$S = \begin{cases} c_1+1, c_1+2, \ldots, c_2-1 & c_1 \leq c_2 \\ c_1+1, c_1+2, \ldots, N, 1, 2, \ldots, c_2-1 & c_1 > c_2 \end{cases}$$

$$S' = [M] \setminus (S \cup \{c_1, c_2\}) \ .$$

Let $w = c_2 - c_1 \bmod N$. Then $1 \leq w \leq N-1$, and note that $|S| = w-1$ and $|S'| = N-w-1$.

The probability, over random $f \in \mathsf{OPF}_{[M],[N]}*$, that $f^{-1}(c_2) - f^{-1}(c_1) \bmod M = d$ is equal to the number of OPFs g on $[M], [N]$ such that $\{c_1, c_2 \in g([M]), |g([M]) \cap S| = d-1$, and $|g([M]) \cap S'| = M-d-1\}$, over the number of OPFs g such that $c_1, c_2 \in g([M])$, or

$$\frac{\binom{w-1}{d-1}\binom{N-w-1}{M-d-1}}{\binom{N-2}{M-2}} = P_*(N-1, M-1, w, d).$$

$\square$

In particular, for fixed c_1, c_2, d, letting $w = c_2 - c_1 \bmod N$ the lemma says that

$$\Pr_{K \xleftarrow{\$} \mathcal{K}_{\mathrm{r}}} [\, \mathcal{D}ec_{\mathrm{r}}(K, c_2) - \mathcal{D}ec_{\mathrm{r}}(K, c_1) \bmod M = d \mid c_1, c_2 \in \mathcal{E}nc_{\mathrm{r}}(K, [M]) \,]$$
$$= P_*(N-1, M-1, w, d).$$

Now, we can locate the most likely plaintext distance for c_1, c_2.

Corollary B.5.2. *Let $c_1, c_2 \in [N]$ with $c_1 < c_2$, and $w = c_2 - c_1 \bmod N$. Then in $\mathsf{ROPF}_{[M],[N]}$, d_{c_1,c_2} is a most likely plaintext distance from c_1 to c_2 if and only if*

$$\frac{(M-1)w}{N} \leq d_{c_1,c_2} \leq \frac{(M-1)w}{N} + 1.$$

Proof. By Lemma B.1.9, for N, M, w fixed, $P_{NHGD}(N-1, M-1, w, \cdot)$ has a maximum at $d_0 \in [M-1]$ where

$$\frac{(M-1)w}{N} \leq d_0 \leq \frac{(M-1)w}{N} + 1.$$

Therefore, $P_*(N-1, M-1, w, \cdot)$ also has a maximum at d_0, so the result follows from Lemma B.5.1. □

Note in particular that d_{c_1,c_2} depends only on the difference $w = c_2 - c_1 \bmod N$. Thus, for $w \in [N-1]$, we define d_w to be the *most likely plaintext distance for w* and $d_w = d_{c_1,c_2}$ for all $c_1, c_2 \in [N]$ with $w = c_2 - c_1 \bmod N$.

The plaintext distance is uniformly random. Here we establish that no plaintext distance (from 1 to $M-1$) is more or less likely than any other, if the challenge plaintexts are uniformly random and distinct.

Lemma B.5.3. *For any $w \in [N-1]$, over $K \xleftarrow{\$} \mathcal{K}_{\mathrm{r}}$ and $\{m_1, m_2\} \xleftarrow{\$} \mathrm{Cmb}_2^{[M]}$,*

$$\Pr_{K, m_1, m_2} [\, \mathcal{E}nc_{\mathrm{r}}(K, m_2) - \mathcal{E}nc_{\mathrm{r}}(K, m_1) \bmod N = w \,] = \frac{1}{N-1}\,.$$

Proof. In the following, we consider addition and subtraction of ciphertexts to be taken mod N.

$$
\begin{aligned}
&\Pr_{K,m_1,m_2}[\,\mathcal{E}nc_{\mathrm{r}}(K,m_2)-\mathcal{E}nc_{\mathrm{r}}(K,m_1)=w\,]\\
&=\sum_{c\in[N]}\Pr_{K,m_1,m_2}[\,\mathcal{E}nc_{\mathrm{r}}(K,m_1)=c\ \cap\ \mathcal{E}nc_{\mathrm{r}}(K,m_2)=c+w\,]\\
&=\sum_{c\in[N]}\Pr_{K}[\,c,c+w\in\mathcal{E}nc_{\mathrm{r}}(K,[M])\,]\cdot\\
&\qquad\Pr_{m_1,m_2}[\,\mathcal{E}nc_{\mathrm{r}}(K,m_1)=c\ \cap\ \mathcal{E}nc_{\mathrm{r}}(K,m_2)=c+w\mid c,c+w\in\mathcal{E}nc_{\mathrm{r}}(K,[M])\,]\\
&=\sum_{c\in[N]}\frac{\binom{N-2}{M-2}}{\binom{N}{M}}\frac{1}{M}\frac{1}{M-1}\\
&=\frac{1}{N-1}.
\end{aligned}
$$

□

We are now ready to prove the lemma.

Proof of Lemma B.4.2. In the DOW experiment, since the adversary A is given only the challenge ciphertexts c_1, c_2, the adversary will have highest probability to win the game if it outputs the most likely plaintext distance for c_1, c_2. By Lemma B.5.3, $w = c_2 - c_1 \bmod N$ is uniform from $[N-1]$, so the DOW advantage of A is bounded above by the average probability (over all $w \in [N-1]$) that $d_w = \mathcal{D}ec_{\mathrm{r}}(K, c_2) - \mathcal{D}ec_{\mathrm{r}}(K, c_2)$, where K is a random key output by $\mathcal{K}_{\mathrm{r}}$ such that $c_1, c_2 \in \mathcal{E}nc_{\mathrm{r}}(K, [M])$. Thus, the result follows from Lemma B.5.1. □

B.6 Proof of Proposition 4.4.5

Proof of Proposition 4.4.5. Let $t = (N-1)/(M-1)$. Let b be a fixed value (less than $\sqrt{M-1}$) to be determined later. Define $\beta = \frac{2tb\sqrt{M-1}}{t-2}$.

By Lemma B.5.3, w is uniformly random in $[N-1]$, so

$$\Pr_{K,m_1,m_2}[\,w < \beta + 1\,] \leq \frac{\beta+1}{N-1}. \tag{7}$$

Recall from Corollary B.5.2 that $d_w = \left\lceil \frac{(M-1)w}{N} \right\rceil$ is the most likely plaintext distance of w. Let $\delta = \frac{b}{\sqrt{M-1}}$, and define

$$d_R = \min\{d_w + \lfloor \delta(M-1) \rfloor, M-1\}.$$

Then note that whenever $w \geq \beta + 1$,

$$\begin{aligned}
d_R &\leq d_w + \lfloor \delta(M-1) \rfloor \\
&\leq \frac{w}{t} + 1 + \delta(M-1) \\
&= \frac{w}{t} + 1 + b\sqrt{M-1} \\
&\leq \frac{2b\sqrt{M-1}}{t-2} + b\sqrt{M-1} + 1 && \text{(since } w > \beta\text{)} \\
&= \frac{b\sqrt{M-1}(2+t-2)}{t-2} + 1 \\
&= \frac{tb\sqrt{M-1}}{t-2} + 1 \\
&= \beta/2 + 1 \\
&< w/2 && \text{(since } w \geq \beta + 1.\text{)}
\end{aligned}$$

Hence,

$$\begin{aligned}
\Pr_{K,m_1,m_2}[\, 2d > w \mid w \geq \beta + 1 \,] &\leq \Pr_{K,m_1,m_2}[\, d > d_R \mid w \geq \beta + 1 \,] \\
&= \sum_{d=d_R+1}^{M-1} P_*(N-1, M-1, w, d) \\
&\leq e^{-2\delta^2(M-2)} \quad \text{(by Lemma B.3.1)} \\
&= e^{-2b^2 \frac{M-2}{M-1}} \\
&< e^{-b^2} \,. && (8)
\end{aligned}$$

Now, putting equations (7) and (8) together,

$$\begin{aligned}\Pr_{K,m_1,m_2}[\,2d > w\,] &\leq \Pr_{K,m_1,m_2}[\,2d > w \mid w \geq \beta+1\,] + \Pr_{K,m_1,m_2}[\,w < \beta+1\,] \\ &\leq e^{-b^2} + \frac{\beta+1}{N-1} \\ &= e^{-b^2} + \frac{2tb\sqrt{M-1}}{(t-2)(N-1)} + \frac{1}{N-1} \\ &= e^{-b^2} + \frac{2b}{(t-2)\sqrt{M-1}} + \frac{1}{N-1}.\end{aligned}$$

We may now select a value for b, say $b = \sqrt{\ln M}$. Then this bound becomes

$$\begin{aligned}\Pr_{K,m_1,m_2}[\,2d > w\,] &\leq 1/M + \frac{2\sqrt{\log M}}{(t-2)\sqrt{M-1}} + 1/(N-1) \\ &< 2/M + \frac{2}{t-2}\frac{1}{\sqrt{(M-1)/\ln M}} \\ &< \frac{3}{t}\frac{1}{\sqrt{(M-1)/\ln M}}\,,\end{aligned}$$

assuming $t \geq 7$. □

B.7 Proof of Theorem 4.5.1

Proof of Theorem 4.5.1. Let $\mathcal{D}$ be the domain, and suppose $A = (A_1, A_2)$ is an adversary with nontrivial IND-CCPA advantage against $\mathcal{CEOE}$. We construct an IND-CPA adversary B against $\mathcal{SE}$. B has access to O, a left-right encryption oracle for $\mathcal{SE}$ under a random secret key.

B runs A_1 to receive $\mathcal{M}_0, \mathcal{M}_1, \sigma$. Let l be the lengths of $|\mathcal{M}_0|, |\mathcal{M}_1|$. After sorting (separately) the elements of $\mathcal{M}_0$ and $\mathcal{M}_1$, B assigns label m_b^i to the ith smallest element of $\mathcal{M}_b$, for $i = 1, \ldots, l$ and $b = 0, 1$. B queries its left-right $\mathcal{SE}$-encryption oracle with matched pairs of these messages: $c_i' \leftarrow O(m_0^i, m_1^i)$ for $i = 1, \ldots, l$. Note that each pair consists of messages of equal length. Then, B prepends indices $c_i = i \| c_i'$ for $i = 1, \ldots, l$. Finally, it runs $A_2(\sigma, c_1, \ldots, c_l)$ to receive d, and outputs d.

It is clear that B's communication with A perfectly mimics the IND-OCPA experiment, and thus the IND-CPA advantage of B is equal to the IND-CCPA advantage of A. Clearly, B is efficient, since it only needs to sort the elements of $|\mathcal{M}_0|, |\mathcal{M}_1|$. □

B.8 Proof of Proposition 4.5.2

Proof of Proposition 4.5.2. Let $V_{\mathbf{m}}$ be the set of r-windows in $[M]$ that contain an element of $\mathbf{m}$. Notice that $|V_{\mathbf{m}}| \leq rz$, as each element of the challenge set is contained in at most r windows. Also, the total number of r-windows in $[M]$ is M. An adversary wins if it outputs an element in $V_{\mathbf{m}}$. Since A_{rand} outputs a random r-window, it is clear that $\mathbf{Adv}^{r,z\text{-wow}}_{\mathsf{RMOPF}_{[M],[N]}}(A_{\text{rand}}) \leq rz/M$.

Fix a function $f \in \mathsf{OPF}_{[M],[N]}$ and challenge set $\mathbf{c}$. Let $f^{-1}(\mathbf{c}) = \{x \in [M] \mid f(x) \in \mathbf{c}\}$. Let S be the set of modular intervals $I' \subseteq [M]$ such that $I' \cap f^{-1}(\mathbf{c}) \neq \emptyset$, and let $n = |S|$. For offset j, an adversary wins if it picks $I = (m_L, m_R)$ such that the interval $I + j = (m_L + j \bmod M, m_R + j \bmod M)$ is in S. For each I, note that there are precisely n values for $j \in [M]$ for which $I + j \in S$, and precisely $M - n$ for which $I + j \notin S$. Thus, over the choice of j, each interval I has the same probability of winning (namely, n/M.) Hence, a random choice of interval has the same probability of success as any other choice of interval. This is true for any function f and challenge set $\mathbf{c}$, so the result follows. □

APPENDIX C

EFSE PRIMITIVES AND PROOFS

C.1 Efficiently searchable encryption formal definition and security notion

As defined in [3], we say that $\texttt{ESE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, F, G)$ is an *efficient searchable encryption* (ESE) scheme on domain $\mathcal{D}$ if $(\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec)$ is a symmetric encryption scheme on $\mathcal{D}$ and F, G are deterministic functions such that for every $m \in \mathcal{D}$ and efficient randomized algorithm A that outputs distinct messages $m_0, m_1 \in \mathcal{D}$,

$$\Pr_{K \xleftarrow{\$} \mathcal{K}} [\, F(K, m) = G(\mathcal{E}nc(K, m)) \,] = 1, \quad \text{and}$$

$$\Pr_{\substack{K \xleftarrow{\$} \mathcal{K} \\ (m_0, m_1) \xleftarrow{\$} A}} [\, F(K, m_0) = G(\mathcal{E}nc(K, m_1)) \,] \quad \text{is sufficiently small.}$$

Notice that an ESE scheme leaks equality, as if c_1, c_2 are both encryptions of m under key K, then $G(c_1) = F(K, m) = G(c_2)$, and this happens with low probability if c_1 and c_2 are encryptions of distinct messages.

Since ESE schemes leak equality, the following notion called *indistinguishability under distinct chosen-plaintext attacks* [11] is appropriate to evaluate their security. For $b \in \{0, 1\}$, ESE scheme $\texttt{ESE}$, and adversary A, let $\mathbf{Exp}^{\text{ind-cpa-}b}_{\texttt{ESE}}(A)$ be identical to IND-CPA experiment $\mathbf{Exp}^{\text{ind-cpa-}b}_{\texttt{ESE}}(A)$ (see Chapter 2) but with the restriction that left/right-queries have the same *equality pattern*. That is, for left/right-query pairs (m_0, m_1) and (m'_0, m'_1), we have $m_0 = m'_0$ if and only if $m_1 = m'_1$. For an adversary A, define its *IND-DCPA advantage* against $\texttt{FSE}$ as

$$\mathbf{Adv}^{\text{ind-dcpa}}_{\texttt{ESE}}(A) \;=\; \Pr\left[\, \mathbf{Exp}^{\text{ind-dcpa-1}}_{\texttt{ESE}}(A) = 1 \,\right] - \Pr\left[\, \mathbf{Exp}^{\text{ind-dcpa-0}}_{\texttt{ESE}}(A) = 1 \,\right].$$

We say that $\texttt{ESE}$ is *indistinguishable under distinct chosen-plaintext attacks* (IND-DCPA-secure) if the IND-DCPA advantage of any adversary against $\texttt{ESE}$ is small.

C.2 Construction of a batch-tagging family on $\mathcal{D}$ given a PRF on $\mathcal{D}$

Let $\mathtt{PRF} = (\mathcal{K}_{\mathtt{PRF}}, \mathcal{F}_{\mathtt{PRF}})$ be a function family on domain $\mathcal{D}$ to some range $\mathcal{R}$. Let $\mathcal{F}_{\mathtt{BTag}} = (\mathcal{K}_{\mathcal{T}}, \mathcal{T}, \mathcal{B})$ where $\mathcal{K}_{\mathcal{T}} = \mathcal{K}_{\mathtt{PRF}}$, $\mathcal{T} = \mathcal{F}_{\mathtt{PRF}}$, and $\mathcal{B}$ is defined in the standard way using $\mathcal{T}$ as described above. We claim that if $\mathtt{PRF}$ is a PRF, then $\mathcal{F}_{\mathtt{BTag}}$ is PP-CBT-secure.

Proposition C.2.1. *For $\mathcal{F}_{\mathtt{BTag}}$ constructed as above out of function family $\mathtt{PRF}$, and any adversary A, there exist PRF adversaries F_0 and F_1 such that*

$$\mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(A) = \mathbf{Adv}^{\text{prf}}_{\mathtt{PRF}}(F_0) + \mathbf{Adv}^{\text{prf}}_{\mathtt{PRF}}(F_1).$$

Further, if A submits queries of total length γ to its oracle, then F_1 and F_2 each submit queries of total length γ to their oracles as well.

Proof. Let A be a PP-CBT adversary against $\mathcal{F}_{\mathtt{BTag}}$. For $\alpha \in \{0,1\}$, construct PRF adversary F_α against $\mathtt{PRF}$ as follows.

Adversary $F_\alpha^{\mathcal{O}(\cdot)}$

Let $\mathcal{P}_\alpha$ be the oracle that on input (M_0, M_1), runs:

Let $M_\alpha = \{m_\alpha^1, \ldots, m_\alpha^q\}$

$c_i \leftarrow \mathcal{O}(m_\alpha^i)$ for $i \in [q]$

Return $C = \{c_1, \ldots, c_q\}$

$b' \xleftarrow{\$} A^{\mathcal{P}_\alpha(\cdot,\cdot)}$

Return b'

The query-length claims on F_1 and F_2 should be clear from the construction.

Now, we claim that

$$\begin{aligned}
\mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\texttt{BTag}}}(A) &= \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\texttt{BTag}}}(A)=1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\texttt{BTag}}}(A)=1\right] \\
&= \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\texttt{BTag}}}(A)=1\right] - \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[F_1^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}=1\right] && \text{[I]} \\
&\quad + \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[F_1^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}=1\right] - \Pr_{f \xleftarrow{\$} \mathsf{Func}_{\mathcal{D},\mathcal{R}}}\left[F_1^{f(\cdot)}=1\right] && \text{[II]} \\
&\quad + \Pr_{f \xleftarrow{\$} \mathsf{Func}_{\mathcal{D},\mathcal{R}}}\left[F_1^{f(\cdot)}=1\right] - \Pr_{f' \xleftarrow{\$} \mathsf{Func}_{\mathcal{D},\mathcal{R}}}\left[F_0^{f'(\cdot)}=1\right] && \text{[III]} \\
&\quad + \Pr_{f \xleftarrow{\$} \mathsf{Func}_{\mathcal{D},\mathcal{R}}}\left[F_0^{f(\cdot)}=1\right] - \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[F_0^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}=1\right] && \text{[IV]} \\
&\quad + \Pr_{K \xleftarrow{\$} \mathcal{K}}\left[F_0^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}=1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\texttt{BTag}}}(A)=1\right] && \text{[V]} \\
&= 2\mathbf{Adv}^{\text{prf}}_{\mathcal{F}}(A)
\end{aligned}$$

Note that [II] and [IV] evaluate to $\mathbf{Adv}^{\text{prf}}_{\texttt{PRF}}(F_1)$ and $\mathbf{Adv}^{\text{prf}}_{\texttt{PRF}}(F_0)$, respectively. It is left to show that [I], [III], and [V] evaluate to zero.

[I] is zero: By construction of $\mathcal{F}_{\texttt{BTag}}$, the oracle $\mathcal{P}_\alpha(\cdot)$ constructed by $F_1^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}$ mimics the oracle in the experiment $\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\texttt{BTag}}}(A)$ on $\mathcal{F}_{\texttt{BTag}}$. So A is given equivalent oracles in both cases, and $F_1^{\mathcal{F}_{\texttt{PRF}}(K,\cdot)}$ outputs the result that A outputs.

[III] is zero: We show there is a bijection $f \leftrightarrow f'$ between functions in $\mathsf{Func}_{\mathcal{D},\mathcal{R}}$ such that $\mathcal{P}_1$ in adversary $F_1^{f(\cdot)}$ is equivalent to $\mathcal{P}_0$ in adversary $F_0^{f'(\cdot)}$. Then, since A is given equivalent oracles in either case, and either adversary outputs the output of A, the result follows.

Suppose $(M_0^1, M_1^1), \ldots, (M_0^q, M_1^q)$ are the queries A makes to its oracle in the PP-CBT experiment. Then by the PP-CBT restriction, for all $I \subseteq [q]$ we have $\left|\bigcap_{i\in I} M_0^i\right| = \left|\bigcap_{i\in I} M_1^i\right|$. Intuitively, this means that if we draw two Venn diagrams, one of the sets M_0^i for $i \in [q]$ and the other of the sets M_1^i for $i \in [q]$, the number of elements in corresponding (same-index) regions is identical in both diagrams. This implies that there exists a bijection $\phi : \bigcup_{i\in[q]} M_0^i \to \bigcup_{i\in[q]} M_1^i$ such that $m \in M_0^i$ if and only if $\phi(m) \in M_1^i$, for all $i \in [q]$.

Given $f \in \mathsf{Func}_{\mathcal{D},\mathcal{R}}$ fixed, let f' be the function that is the same as f, except that for any $m \in \bigcup_{i\in[q]} M_0^i$ corresponding to $\phi(m) \in \bigcup_{i\in[q]} M_1^i$, f' sends $m \mapsto f(\phi(m))$ and $\phi(m) \mapsto f(m)$. This indicates a bijection between functions f and f' in $\mathsf{Func}_{\mathcal{D},\mathcal{R}}$ where $\{f(m) \mid m \in M_0^i\} = \{f'(m) \mid m \in M_1^i\}$. Hence, for corresponding f, f' indicated by the bijection, $\mathcal{P}_1$ in adversary $F_1^{f(\cdot)}$ is equivalent to $\mathcal{P}_0$ in adversary $F_0^{f'(\cdot)}$.

[V] is zero: Analogous reasoning to [I]. □

C.3 Proof of Theorem 5.4.2

Proof of Theorem 5.4.2. Let $\mathtt{FSE} = \mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathcal{F}_{\mathtt{BTag}}, \mathtt{ESE}]$. Let A be an efficient IND-CLS-CPA adversary to $\mathtt{FSE}$. We construct a PP-CBT adversary E_A against $\mathcal{F}_{\mathtt{BTag}}$ and an IND-DCPA adversary F_A against $\mathtt{ESE}$, as follows.

Adversary $E_A^{\mathcal{B}(K_{\mathcal{T}},\mathcal{LR}(\cdot,\cdot,b))}$

$K_{\mathtt{ESE}} \xleftarrow{\$} \mathcal{K}_{\mathtt{ESE}}$

Define oracle $\mathcal{P}(m_0, m_1)$:

$B_0 \leftarrow \mathtt{Bkts}(m_0)$

$B_1 \leftarrow \mathtt{Bkts}(m_1)$

$\mathtt{tags} \leftarrow \mathcal{B}(K_{\mathcal{T}}, \mathcal{LR}(B_0, B_1, b))$

$c_R \leftarrow \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, m_0)$

Return $\mathtt{tags} \| c_R$

$b' \xleftarrow{\$} A^{\mathcal{P}(\cdot,\cdot)}$

Return b'

Adversary $F_A^{\mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}},\mathcal{LR}(\cdot,\cdot,b))}$

$K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}}$

Define oracle $\mathcal{Q}(m_0, m_1)$:

$B_1 \leftarrow \mathtt{Bkts}(m_1)$

$\mathtt{tags} \leftarrow \mathcal{B}(K_{\mathcal{T}}, B_1)$

$c_R \leftarrow \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, \mathcal{LR}(m_0, m_1, b))$

Return $\mathtt{tags} \| c_R$

$b' \xleftarrow{\$} A^{\mathcal{Q}(\cdot,\cdot)}$

Return b'

First, the efficiency claims about E_A and F_A should be clear from the definitions of oracles $\mathcal{P}$ and $\mathcal{Q}$ and the fact that each adversary runs A once while simulating A's oracle efficiently.

Now, we show that

$$
\begin{aligned}
&\mathbf{Adv}^{\text{ind-cls-cpa}}_{\mathtt{FSE}}(A) \\
&= \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-0}}_{\mathtt{FSE}}(A) = 1\right] \\
&= \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}(F_A) = 1\right] && \text{[I]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}(F_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}(F_A) = 1\right] && \text{[II]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}(F_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A) = 1\right] && \text{[III]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A) = 1\right] && \text{[IV]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-cls-cpa-0}}_{\mathtt{FSE}}(A) = 1\right] && \text{[V]} \\
&= \mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(F_A) + \mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A).
\end{aligned}
$$

Note that [II] evaluates to $\mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(F_A)$ and [IV] evaluates to $\mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(E_A)$. It is left to show that [I], [III], and [V] evaluate to zero.

[I] is zero: Knowing A is a valid adversary to experiment $\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}$, we claim F_A is a valid adversary to $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$. Within F_A, suppose $(m_0^1, m_1^1), \ldots, (m_0^q, m_1^q)$ are the queries A makes to $\mathcal{Q}(\cdot,\cdot)$. Then for any $i, j \in [q]$, $|m_0^i| = |m_1^i|$, and $m_0^i = m_0^j$ if and only if $m_1^i = m_1^j$, since A satisfies the restrictions of $\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}$. Thus, F_A satisfies the restriction of $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$.

In experiment $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$, the oracle $\mathcal{Q}(\cdot,\cdot)$ constructed by F_A simulates A's oracle in the experiment $\mathbf{Exp}^{\text{ind-cls-cpa-1}}_{\mathtt{FSE}}(A)$, and F_A outputs the result that A outputs.

[III] is zero: The oracle $\mathcal{P}(\cdot,\cdot)$ constructed by E_A in experiment $\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}$ and the oracle $\mathcal{Q}$ constructed by F_A in $\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}$ are functionally equivalent: after keys $K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}}$ and $K_{\mathtt{ESE}} \xleftarrow{\$} \mathcal{K}_{\mathtt{ESE}}$ are selected, both oracles take input (m_0, m_1) and output

$$\mathcal{B}(K_{\mathcal{T}}, B_1) \| \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, m_0).$$

So A is given equivalent oracles in the two cases, and each adversary outputs A's output.

[V] is zero: Knowing A is a valid adversary to experiment $\mathbf{Exp}^{\text{ind-cls-cpa-0}}_{\mathtt{FSE}}$, we claim E_A is a valid adversary to $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$. Suppose $(m_0^1, m_1^1), \ldots, (m_0^q, m_1^q)$ are the queries A makes to $\mathcal{P}(\cdot,\cdot)$. Then for any $i, j \in [q]$, either $\mathrm{Cl}_{\mathcal{D}}(m_0^i, m_0^j) = \mathrm{Cl}_{\mathcal{D}}(m_1^i, m_1^j)$ or $m_0^i = m_0^j$ and $m_1^i = m_1^j$. Fix $I \subseteq [q]$. Then since $\mathtt{Bkts}$ is consistent,

$$\left|\bigcap_{i\in I} \mathtt{Bkts}(m_0^i)\right| = \left|\bigcap_{i\in I} \mathtt{Bkts}(m_1^i)\right|.$$

Thus E_A satisfies the restriction of $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$. In experiment $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$, the oracle $\mathcal{P}(\cdot,\cdot)$ constructed by E_A simulates A's oracle in the experiment $\mathbf{Exp}^{\text{ind-cls-cpa-0}}_{\mathtt{FSE}}(A)$, and E_A outputs the result that A outputs, and the result follows. □

C.4 Proof of Theorem 5.4.3

Proof of Theorem 5.4.3. Suppose $\mathtt{Bkts}$ is not consistent. Thus, there exist $q > 1$ and message sets $\{m_0^1, \ldots, m_0^q\}$ and $\{m_1^1, \ldots, m_1^q\}$ having the same closeness pattern such that $\left|\bigcap_{i\in[q]} \mathtt{Bkts}(m_0^i)\right| \neq \left|\bigcap_{i\in[q]} \mathtt{Bkts}(m_1^i)\right|$.

We construct an adversary A against the IND-CLS-CPA security of the scheme $\mathtt{FSE}_{\mathtt{BktTag}}[\mathtt{Bkts}, \mathtt{PRTag}, \mathtt{ESE}]$ as follows. For $i \in [q]$, A submits queries (m_0^i, m_1^i) to its oracle, receiving ciphertexts $c_i = \mathtt{tags}_i \| c_i' = \mathcal{E}nc(K, m_b^i)$. A then compares $\left|\bigcap_{i\in[q]} \mathtt{tags}_i\right|$ with $\left|\bigcap_{i\in[q]} \mathtt{Bkts}_1\right|$. If they are equal, A outputs 1, and otherwise 0.

The attack is valid, as the corresponding messages have the same closeness pattern. Also, it is clear that A makes q oracle queries. Further, note that $\left|\bigcap_{i\in[q]} \mathtt{tags}_i\right| = \left|\bigcap_{i\in[q]} \mathtt{Bkts}_b\right| \neq \left|\bigcap_{i\in[q]} \mathtt{Bkts}_{1-b}\right|$ since $\mathcal{T}(K_{\mathcal{T}}, \cdot)$ is deterministic and collision-free for a given $\mathcal{F}_{\mathtt{BTag}}$-key $K_{\mathcal{T}}$. So A always succeeds, and the result follows. □

C.5 Proof of Theorem 5.5.1

Proof of Theorem 5.5.1. The result will follow easily if we simply show that `Bkts` satisfies the condition of Theorem 5.4.2.

Let $\{m_0^1, \ldots, m_0^q\}$ and $\{m_1^1, \ldots, m_1^q\}$ be sets of messages having the same closeness pattern. That is, either $\mathrm{Cl}_{\mathcal{D}}(m_0^i, m_0^j) = \mathrm{Cl}_{\mathcal{D}}(m_1^i, m_1^j)$ or $(m_0^i = m_0^j$ and $m_0^i = m_0^j)$ for all $i, j \in [q]$.

For $i \in [q]$, $\alpha \in \{0, 1\}$, let $E_\alpha^i = \{e \in \mathcal{E}_{\mathtt{dum}} \mid m_\alpha^i \in e\}$. Let $I = [q]$. Three cases arise:

1. Suppose $\{m_0^i \mid i \in I\}$ contains at least three messages, say $m_0^\alpha \neq m_0^\beta \neq m_0^\gamma$ for $\alpha \neq \beta \neq \gamma$. Then by the equality condition, $m_1^\alpha, m_1^\beta, m_1^\gamma$ are all distinct. Three (or more) distinct vertices cannot all share the same edge, so we conclude in this case that $\left|\bigcap_{i\in I} E_0^i\right| = 0 = \left|\bigcap_{i\in I} E_1^i\right|$.

2. Suppose $\{m_0^i \mid i \in I\}$ contains exactly two distinct messages, say m_0^α and m_0^β. Let $I_\alpha = \{i \in I \mid m_0^i = m_0^\alpha\}$ and $I_\beta = \{i \in I \mid m_0^i = m_0^\beta\}$; then I_α and I_β are nonempty. Let $\eta = \mathrm{Cl}_{\mathcal{D}}(m_0^\alpha, m_0^\beta) \in \{\mathtt{close}, \mathtt{near}, \mathtt{far}\}$.

   (a) For $i, j \in I_\alpha$ (or $i, j \in I_\beta$), $m_0^i = m_0^j$, so $m_1^i = m_1^j$;

   (b) For $i \in I_\alpha$, $j \in I_\beta$, $\mathrm{Cl}_{\mathcal{D}}(m_1^i, m_1^j) = \mathrm{Cl}_{\mathcal{D}}(m_0^i, m_0^j) = \eta$.

   Hence, if $\eta = \mathtt{close}$, $\{m_0^i \mid i \in I\}$ contains exactly two distinct close messages and $\{m_1^i \mid i \in I\}$ contains exactly two distinct close messages, so $\left|\bigcap_{i\in I} E_0^i\right| = 1 = \left|\bigcap_{i\in I} E_1^i\right|$.

   On the other hand, if $\eta = \mathtt{far}$, $\{m_0^i \mid i \in I\}$ contains exactly two distinct far messages and $\{m_1^i \mid i \in I\}$ contains exactly two distinct far messages, so $\left|\bigcap_{i\in I} E_0^i\right| = 0 = \left|\bigcap_{i\in I} E_1^i\right|$.

3. Finally, suppose $\{m_0^i \mid i \in I\}$ contains only one distinct message, say m_0^α. Then by the equality condition, $\{m_1^i \mid i \in I\}$ also contains only one distinct message.

So $\left|\bigcap_{i\in I} E_0^i\right| = \Delta = \left|\bigcap_{i\in I} E_1^i\right|$.

Thus, `Bkts` satisfies the restriction of Theorem 5.4.2. □

C.6 Proof of Theorem 5.5.2

Proof of Theorem 5.5.2. Let $\Delta \geq 2$ be fixed. Define $\Lambda = (\mathcal{D}, \mathrm{Cl}_{\mathcal{D}})$ as the rigid closeness domain with closeness graph $\mathcal{G}_\Lambda$ as follows.

Let K_Δ be the complete graph on Δ vertices, with vertices labeled $v_1, \ldots, v_\Delta$. Let $\sigma = 2^{\Delta(\Delta-1)/2}$, and let $\mathcal{G}_1, \ldots, \mathcal{G}_\sigma$ be the distinct subgraphs of K_Δ with the same order of labeled vertices for each: $v_1^i, \ldots, v_\Delta^i$ for $i \in [\sigma]$. Now, define $\mathcal{G}_\Lambda$ to be the union of $\mathcal{G}_1, \ldots, \mathcal{G}_\sigma$ (with no edges between any of the subgraphs). It should be clear that Δ is indeed the maximum degree of this graph.

Let $\mathtt{FSE} = (\mathcal{K}, \mathcal{E}nc, \mathcal{D}ec, \mathrm{Cl}_{\mathcal{R}})$ be a perfect FSE encryption scheme on Λ. We may assume that the ciphertext space of $\mathtt{FSE}$ is $\mathcal{R} = \{0,1\}^r$, and suppose to the contrary that $r < (\Delta - 1)/2$. Fix a key $K \xleftarrow{\$} \mathcal{K}$, and let $c_i^j = \mathcal{E}nc(K, v_i^j)$ for $i \in [\Delta]$, $j \in [\sigma]$. Further, let $\mathbf{c}^j = (c_1^j, \ldots, c_\Delta^j)$ for $j \in \sigma$.

Notice that each $\mathbf{c}^j$ consists of Δr bits, so there are $2^{\Delta r}$ possible values for $\mathbf{c}^j$, for each $j \in [\sigma]$. However, $\sigma = 2^{\Delta(\Delta-1)/2} > 2^{\Delta r}$, so by the pigeonhole principle we have $\mathbf{c}^i = \mathbf{c}^j$ for some $i \neq j \in [\sigma]$. Since $\mathcal{G}^i$ and $\mathcal{G}^j$ are different subgraphs of K_Δ, without loss of generality suppose edge $v_x^i v_y^i$ is in $\mathcal{G}^i$ but $v_x^j v_y^j$ is not in $\mathcal{G}^j$. Thus, $\mathrm{Cl}_{\mathcal{D}}(v_x^i, v_y^i) = \mathtt{close}$ and $\mathrm{Cl}_{\mathcal{D}}(v_x^j, v_y^j) = \mathtt{far}$, so $\mathtt{close} = \mathrm{Cl}_{\mathcal{R}}(c_x^i, c_y^i) = \mathrm{Cl}_{\mathcal{R}}(c_x^j, c_y^j) = \mathtt{far}$, a contradiction.

We conclude that $r \geq (\Delta - 1)/2$ and thus $r \in \Omega(\Delta)$. □

C.7 Proof of Theorem 5.6.1

Proof of Theorem 5.6.1. Let A be an IND-NR$\mathcal{L}$-CPA adversary to $\mathtt{FSEtagAnc}_{\mathcal{L}}^{\rho}$. We construct a PP-CBT adversary F_A against $\mathcal{F}_{\mathtt{BTag}}$ and an IND-DCPA adversary E_A against $\mathtt{ESE}$, as follows.

Adversary $E_A^{\mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}},\mathcal{LR}(\cdot,\cdot,b))}$

$K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}}$

Define oracle $\mathcal{Q}(\mathbf{m}_0, \mathbf{m}_1)$:

$\mathtt{Anc}_1 \leftarrow \{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}_1, \mathbf{v}) \le \rho\}$

$\mathtt{tags} \leftarrow \mathcal{B}(K_{\mathcal{T}}, \mathtt{Anc}_1)$

$c_R \leftarrow \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, \mathcal{LR}(\mathbf{m}_0, \mathbf{m}_1, b))$

Return $\mathtt{tags} \| c_R$

$b' \xleftarrow{\$} A^{\mathcal{Q}(\cdot,\cdot)}$

Return b'

Adversary $F_A^{\mathcal{B}(K_{\mathcal{T}},\mathcal{LR}(\cdot,\cdot,b))}$

$K_{\mathtt{ESE}} \xleftarrow{\$} \mathcal{K}_{\mathtt{ESE}}$

Define oracle $\mathcal{P}(\mathbf{m}_0, \mathbf{m}_1)$:

$\mathtt{Anc}_0 \leftarrow \{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}_0, \mathbf{v}) \le \rho\}$

$\mathtt{Anc}_1 \leftarrow \{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}_1, \mathbf{v}) \le \rho\}$

$\mathtt{tags} \leftarrow \mathcal{B}(K_{\mathcal{T}}, \mathcal{LR}(\mathtt{Anc}_0, \mathtt{Anc}_1, b))$

$c_R \leftarrow \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, \mathbf{m}_0)$

Return $\mathtt{tags} \| c_R$

$b' \xleftarrow{\$} A^{\mathcal{P}(\cdot,\cdot)}$

Return b'

First, since we have a short basis for $\mathcal{L}$, it is easy to find $\mathtt{Anc}_j$ for message $\mathbf{m}_j$, for $j \in \{0, 1\}$. Notice that if m_j is described with γ_j bits, the messages in $\mathtt{Anc}_j$ can be described with at most $\gamma_j + \log_2 \rho$ bits. Then, the efficiency claims on E_A and F_A are clear from the definitions of oracles $\mathcal{P}$ and $\mathcal{Q}$ and the fact that each adversary runs A once while simulating A's oracle efficiently.

Now, we show that

$$
\begin{aligned}
&\mathbf{Adv}^{\text{ind-nr}\mathcal{L}\text{-cpa}}_{\mathtt{FSE}}(A) \\
&= \Pr\left[\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-0}}_{\mathtt{FSE}}(A) = 1\right] \\
&= \Pr\left[\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}(A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}(E_A) = 1\right] && \text{[I]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}(E_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}(E_A) = 1\right] && \text{[II]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}(E_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A) = 1\right] && \text{[III]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A) = 1\right] && \text{[IV]} \\
&\quad + \Pr\left[\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A) = 1\right] - \Pr\left[\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-0}}_{\mathtt{FSE}}(A) = 1\right] && \text{[V]} \\
&= \mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(E_A) + \mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A).
\end{aligned}
$$

Note that [II] evaluates to $\mathbf{Adv}^{\text{ind-dcpa}}_{\mathtt{ESE}}(E_A)$ and [IV] evaluates to $\mathbf{Adv}^{\text{pp-cbt}}_{\mathcal{F}_{\mathtt{BTag}}}(F_A)$. It

is left to show that [I], [III], and [V] evaluate to zero.

[I] is zero: Knowing A is a valid adversary to experiment $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}$, we claim E_A is a valid adversary to $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$. Within E_A, suppose $(\mathbf{m}_0^1, \mathbf{m}_1^1), \ldots, (\mathbf{m}_0^q, \mathbf{m}_1^q)$ are the queries A makes to $\mathcal{Q}(\cdot,\cdot)$. Then for any $i, j \in [q]$, $|m_0^i| = |m_1^i|$, and $m_0^i = m_0^j$ if and only if $m_1^i = m_1^j$, since A satisfies the restrictions of $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}$. Thus, E_A satisfies the restriction of $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$.

In experiment $\mathbf{Exp}^{\text{ind-dcpa-1}}_{\mathtt{ESE}}$, the oracle $\mathcal{Q}(\cdot,\cdot)$ constructed by E_A simulates A's oracle in the experiment $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-1}}_{\mathtt{FSE}}(A)$, and E_A outputs the result that A outputs.

[III] is zero: The oracle $\mathcal{P}(\cdot,\cdot)$ constructed by F_A in experiment $\mathbf{Exp}^{\text{pp-cbt-1}}_{\mathcal{F}_{\mathtt{BTag}}}$ and the oracle $\mathcal{Q}$ constructed by E_A in $\mathbf{Exp}^{\text{ind-dcpa-0}}_{\mathtt{ESE}}$ are functionally equivalent: after keys $K_{\mathcal{T}} \xleftarrow{\$} \mathcal{K}_{\mathcal{T}}$ and $K_{\mathtt{ESE}} \xleftarrow{\$} \mathcal{K}_{\mathtt{ESE}}$ are selected, both oracles take input $(\mathbf{m}_0, \mathbf{m}_1)$ and output

$$\mathcal{B}(K_{\mathcal{T}}, \mathtt{Anc}_1) \| \mathcal{E}nc_{\mathtt{ESE}}(K_{\mathtt{ESE}}, \mathbf{m}_0).$$

So A is given equivalent oracles in the two cases, and each adversary outputs A's output.

[V] is zero: Knowing A is a valid adversary to experiment $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-0}}_{\mathtt{FSE}}$, we claim F_A is a valid adversary to $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$. Suppose $(\mathbf{m}_0^1, \mathbf{m}_1^1), \ldots, (\mathbf{m}_0^q, \mathbf{m}_1^q)$ are the queries A makes to $\mathcal{P}(\cdot,\cdot)$. Then for any $i, j \in [q]$, either $\mathrm{Cl}_{\mathcal{D}}(\mathbf{m}_0^i, \mathbf{m}_0^j) = \mathrm{Cl}_{\mathcal{D}}(\mathbf{m}_1^i, \mathbf{m}_1^j)$ or $\mathbf{m}_0^i = \mathbf{m}_0^j$ and $\mathbf{m}_1^i = \mathbf{m}_1^j$.

For $i \in [q]$, $\alpha \in \{0,1\}$, let $\mathtt{Anc}_\alpha^i = \{\mathbf{v} \in \mathcal{L} \mid d(\mathbf{m}_\alpha^i, \mathbf{v}) \leq \rho\}$. Fix $I \subseteq [q]$. Two cases arise:

1. Suppose $\exists i \neq j \in I$ such that $d(\mathbf{m}_0^i, \mathbf{m}_0^j) > \delta^{\mathtt{F}}$. Since A is a valid IND-NR$\mathcal{L}$-CPA adversary, its left/right-queries have the same equality/closeness

pattern, so this means $d(\mathbf{m}_1^i, \mathbf{m}_1^j) > \delta^{\mathtt{F}}$ as well. Then by the construction of $\mathtt{BktsAnc}_{\mathcal{L}}^{\rho}$, $\mathtt{Anc}_0^i \cap \mathtt{Anc}_0^j = \emptyset = \mathtt{Anc}_1^i \cap \mathtt{Anc}_1^j$ and thus

$$\left|\bigcap_{i\in I} \mathtt{Anc}_0^i\right| = 0 = \left|\bigcap_{i\in I} \mathtt{Anc}_1^i\right|.$$

2. Suppose $d(\mathbf{m}_0^i, \mathbf{m}_0^j) \leq \delta^{\mathtt{F}}$ for all $i \neq j \in I$. Then messages $\{\mathbf{m}_0^i \mid i \in I\}$ are all in the same nearness component of $\mathcal{G}_{\Lambda}^{\mathtt{N}}(H_0)$ where $H_0 = \{\mathbf{m}_0^1, \ldots, \mathbf{m}_0^q\}$. Since A is a valid IND-NR$\mathcal{L}$-CPA adversary, $\{\mathbf{m}_1^i \mid i \in I\}$ are all in the same nearness component of $\mathcal{G}_{\Lambda}^{\mathtt{N}}(H_1)$ where $H_1 = \{\mathbf{m}_0^1, \ldots, \mathbf{m}_0^q\}$, and there exists some $\mathbf{v} \in \mathcal{L}$ such that $m_0^i + \mathbf{v} = m_1^i$ for all $i \in I$. Note that since $\mathcal{L}$ is regular, this means $\mathtt{Anc}_0^i + \mathbf{v} = \{\mathbf{w} + \mathbf{v} \mid \mathbf{w} \in \mathtt{Anc}_1^i\} = \mathtt{Anc}_1^i$ for all $i \in I$. Thus, for $\mathbf{w} \in \mathbb{R}^{\ell}$,

$$\begin{aligned}
\mathbf{w} \in \bigcap_{i\in I} \mathtt{Anc}_0^i &\Leftrightarrow \mathbf{w} \in \mathtt{Anc}_0^i \text{ for all } i \in I \\
&\Leftrightarrow \mathbf{w} + \mathbf{v} \in \mathtt{Anc}_0^i + \mathbf{v} \text{ for all } i \in I \\
&\Leftrightarrow \mathbf{w} + \mathbf{v} \in \mathtt{Anc}_1^i \text{ for all } i \in I \\
&\Leftrightarrow \mathbf{w} + \mathbf{v} \in \bigcap_{i\in I} \mathtt{Anc}_1^i
\end{aligned}$$

Hence, there is a bijection $\mathbf{w} \leftrightarrow \mathbf{w} + \mathbf{v}$ between $\bigcap_{i\in I} \mathtt{Anc}_0^i$ and $\bigcap_{i\in I} \mathtt{Anc}_1^i$, implying $\left|\bigcap_{i\in I} \mathtt{Anc}_0^i\right| = \left|\bigcap_{i\in I} \mathtt{Anc}_1^i\right|$.

Therefore, F_A satisfies the restriction of $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$.

In experiment $\mathbf{Exp}^{\text{pp-cbt-0}}_{\mathcal{F}_{\mathtt{BTag}}}$, the oracle $\mathcal{P}(\cdot,\cdot)$ constructed by F_A simulates A's oracle in the experiment $\mathbf{Exp}^{\text{ind-nr}\mathcal{L}\text{-cpa-0}}_{\mathtt{FSE}}(A)$, and F_A outputs the result that A outputs, and the result follows. □

C.8 Space-efficiency analysis of efficient EFSE constructions

Here, we briefly analyze the space-efficiency of the schemes of Section 5.6.

Space-efficiency for the triangular lattice scheme. Notice that every triangular region of the lattice has the same pattern of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ values. For instance, a point in the middle of a triangular region is always within ρ of only three lattice points, while a point in the corner of a triangular region is always within ρ of seven lattice points. Figure 13 indicates these numbers for points in various sectors of a triangular region, and this pattern holds for all such regions. We conclude that $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})| \in \{3, 4, 5, 6, 7\}$ for all $\mathbf{m} \in \mathbb{R}^2$.

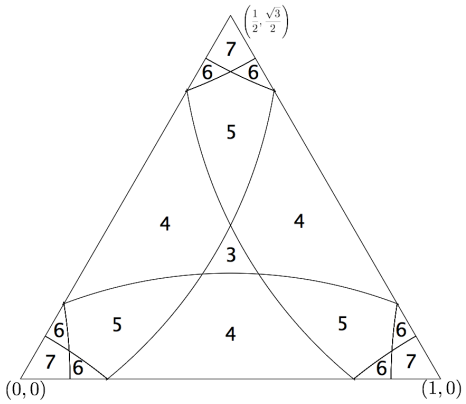

Figure 13: $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ values according to $\mathbf{m}$ location in a region of the triangular lattice

Space-efficiency for the rectangular grid scheme.

Notice that $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ equals the number of points in $\frac{\mathbb{Z}^\ell}{\sqrt{\ell}}$ whose distance from $\mathbf{m}$ is at most $3/2$. Equivalently, this is the number of integer-valued points whose distance from $\mathbf{m}\sqrt{\ell}$ is at most $3\sqrt{\ell}/2$.

We informally conjecture that, in general, it is difficult to describe the distribution of this number. In fact, in $\mathbb{R}^4$ if there were a general formula for the number of integer points in the ball of radius $x > 0$ centered at $\mathbf{0} = (0, 0, 0, 0)$, then we could efficiently factor integers of the form $n = pq$ with p, q prime: Appendix C.9 explains this fact. Though this is a different scenario (as factoring is difficult only for p, q large) it contributes evidence against the existence of a general function describing the number of integer points within an ℓ-dimensional ball centered at some point.

Thus, to describe space efficiency of the rectangular grid scheme, we resort to empirical findings and the following loose theoretical bound. Let $B(\mathbf{x}, r)$ denote the ℓ-dimensional ball of radius r centered at $\mathbf{x}$. For each integer-valued point $\mathbf{x}$, let $H_{\mathbf{x}}$ be the hypercube $\mathbf{x}+[0,1]^{\ell}$, and let $\mathcal{H} = \{H_{\mathbf{x}} \mid \mathbf{x} \in B(\mathbf{m}\sqrt{\ell}, 3\sqrt{\ell}/2)\}$. Since each such hypercube has volume 1, $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ equals the total volume of the hypercubes in $\mathcal{H}$. Now, we claim that $\bigcup_{H\in\mathcal{H}} H \subset B(\mathbf{x}, 5\sqrt{\ell}/2)$. This follows from noting that any point in a hypercube $H_{\mathbf{x}}$ is at most $\sqrt{\ell}$ from a point in $B(\mathbf{x}, 3\sqrt{\ell}/2)$, and using the triangle inequality. Hence, we have the loose upper bound $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})| \leq \mathrm{Vol}(B(\mathbf{x}, 5\sqrt{\ell}/2))$.

Table 6 in contains information about $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ for small dimensions ℓ. The first column evaluates the loose upper bound. The second column is the empirically-computed value of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ at a grid point. The third and fourth columns give empirical lower and upper bounds on the value of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ among 10000 points randomly selected in the space.

Table 6: An analysis of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ values for $\mathbf{m} \in \mathcal{L}$, for various dimensions ℓ. The first column evaluates the loose upper bound. The second column is the empirically-computed value of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ at a grid point. The third and fourth columns give empirical maxima and minima of $|\texttt{BktsAnc}^{\rho}_{\mathcal{L}}(\mathbf{m})|$ among 10000 points randomly selected in the space.

ℓ	Loose upper bound	Value at grid point	Empirical lower bound	Empirical upper bound
1	5	3	3	4
2	39	13	12	16
3	340	81	68	81
4	3084	425	425	1023
5	28736	2463		
6	272516	12277		
7	2616999	69779		
8	25366951	469457		
9	247667506	2634777		
10	2432025947	14763893		
11	23994113427	81598773		
12	237648085570	578480129		

C.9 Connection between the number of integer points in a 4-ball to factoring

Suppose we have a formula $f(x)$, valid for $x > 0$, for the number of integer points in $\mathbb{R}^4$ contained in the ball $B(\mathbf{0}, x)$ of radius x centered at $\mathbf{0} = (0, 0, 0, 0)$. Then for integer n, $f(n) - f(\sqrt{n^2 - 1})$ gives the number of integer points on the boundary of $B(\mathbf{0}, n)$, as for any four integers $x_1, \ldots, x_4$, if $\sum_{i=1}^4 x_i^2 \leq n^2$ then either $\sum_{i=1}^4 x_i^2 = n^2$ or $\sum_{i=1}^4 x_i^2 \leq n^2 - 1$.

Suppose that $n = pq$ for p, q odd primes. By Jacobi's four-square theorem, the number of ways to represent n as the sum of four squares is eight times the sum of the divisors of n, or $8(1 + p + q + n)$.

Thus, the number of ways n can be written as a sum of four squares $\sum_{i=1}^4 x_i^2$ is

$$8(1 + p + q + n) = f(n) - f(\sqrt{n^2 - 1}).$$

So, knowing $n = pq$, we can easily determine p and q.

REFERENCES

[1] Acharya, M., Girao, J., and Westhoff, D., "Secure comparison of encrypted data in wireless sensor networks," in *WiOpt*, pp. 47–53, IEEE Computer Society, 2005.

[2] Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y., "Order-preserving encryption for numeric data," in *SIGMOD Conference* (Weikum, G., König, A. C., and Dessloch, S., eds.), pp. 563–574, ACM, 2004.

[3] Amanatidis, G., Boldyreva, A., and O'Neill, A., "Provably-secure schemes for basic query support in outsourced databases.," in *DBSec* (Barker, S. and Ahn, G.-J., eds.), vol. 4602 of *Lecture Notes in Computer Science*, pp. 14–30, Springer, 2007.

[4] Bauer, F. L., *Decrypted Secrets: Methods and Maxims of Cryptology*. Springer, 2006.

[5] Belazzougui, D., Boldi, P., Pagh, R., and Vigna, S., "Monotone minimal perfect hashing: Searching a sorted table with $o(1)$ accesses," in *SODA* (Mathieu, C., ed.), pp. 785–794, SIAM, 2009.

[6] Bellare, M. and Namprempre, C., "Authenticated encryption: Relations among notions and analysis of the generic composition paradigm," in *ASIACRYPT* (Okamoto, T., ed.), vol. 1976 of *Lecture Notes in Computer Science*, pp. 531–545, Springer, 2000.

[7] Bellare, M., "New proofs for NMAC and HMAC: Security without collision-resistance," in *CRYPTO* (Dwork, C., ed.), vol. 4117 of *Lecture Notes in Computer Science*, pp. 602–619, Springer, 2006.

[8] Bellare, M., Boldyreva, A., Knudsen, L. R., and Namprempre, C., "Online ciphers and the Hash-CBC construction," in *CRYPTO* (Kilian, J., ed.), vol. 2139 of *Lecture Notes in Computer Science*, pp. 292–309, Springer, 2001.

[9] Bellare, M., Boldyreva, A., and O'Neill, A., "Deterministic and efficiently searchable encryption.," in *CRYPTO* (Menezes, A., ed.), vol. 4622 of *Lecture Notes in Computer Science*, pp. 535–552, Springer, 2007.

[10] Bellare, M., Fischlin, M., O'Neill, A., and Ristenpart, T., "Deterministic encryption: Definitional equivalences and constructions without random oracles," in *CRYPTO* (Wagner, D., ed.), vol. 5157 of *Lecture Notes in Computer Science*, pp. 360–378, Springer, 2008.

[11] Bellare, M., Kohno, T., and Namprempre, C., "Breaking and provably repairing the SSH authenticated encryption scheme: A case study of the Encode-then-Encrypt-and-MAC paradigm," *ACM Trans. Inf. Syst. Secur.*, vol. 7, pp. 206–241, May 2004.

[12] Bellare, M. and Rogaway, P., "On the construction of variable-input-length ciphers," in *FSE* (Knudsen, L. R., ed.), vol. 1636 of *Lecture Notes in Computer Science*, pp. 231–244, Springer, 1999.

[13] Bellare, M. and Rogaway, P., "The security of triple encryption and a framework for code-based game-playing proofs," in *EUROCRYPT*, Lecture Notes in Computer Science, pp. 409–426, Springer, 2006.

[14] Boldyreva, A. and Chenette, N., "Efficient fuzzy search on encrypted data." Submitted to ASIACRYPT, Jan. 2012.

[15] Boldyreva, A., Chenette, N., Lee, Y., and O'Neill, A., "Order-preserving symmetric encryption," in *EUROCRYPT* (Joux, A., ed.), vol. 5479 of *Lecture Notes in Computer Science*, pp. 224–241, Springer, 2009.

[16] Boldyreva, A., Chenette, N., and O'Neill, A., "Order-preserving encryption revisited: Improved security analysis and alternative solutions," in *CRYPTO* (Rogaway, P., ed.), vol. 6841 of *Lecture Notes in Computer Science*, pp. 578–595, 2011.

[17] Boldyreva, A., Fehr, S., and O'Neill, A., "On notions of security for deterministic encryption, and efficient constructions without random oracles.," in *CRYPTO* (Wagner, D., ed.), vol. 5157 of *Lecture Notes in Computer Science*, pp. 335–359, Springer, 2008.

[18] Boneh, D. and Waters, B., "Conjunctive, subset, and range queries on encrypted data," in *TCC* (Vadhan, S. P., ed.), vol. 4392 of *Lecture Notes in Computer Science*, pp. 535–554, Springer, 2007.

[19] Brakerski, Z., Gentry, C., and Vaikuntanathan, V., "(leveled) fully homomorphic encryption without bootstrapping," in *ITCS* (Goldwasser, S., ed.), pp. 309–325, ACM, 2012.

[20] Brinkman, R., Feng, L., Doumen, J., Hartel, P. H., and Jonker, W., "Efficient tree search in encrypted data," *Information Systems Security*, vol. 13, no. 3, pp. 14–21, 2004.

[21] Chang, Y.-C. and Mitzenmacher, M., "Privacy preserving keyword searches on remote encrypted data," in *Applied Cryptography and Network Security* (Ioannidis, J., Keromytis, A., and Yung, M., eds.), vol. 3531 of *Lecture Notes in Computer Science*, pp. 442–455, Springer, 2005.

[22] Chvátal, V., "The Tail of the Hypergeometric Distribution," *Discrete Mathematics*, vol. 25, pp. 285–287, 1979.

[23] Curtmola, R., Garay, J. A., Kamara, S., and Ostrovsky, R., "Searchable symmetric encryption: Improved definitions and efficient constructions," in *ACM Conference on Computer and Communications Security* (Juels, A., Wright, R. N., and di Vimercati, S. D. C., eds.), pp. 79–88, ACM, 2006.

[24] Damiani, E., di Vimercati, S. D. C., Jajodia, S., Paraboschi, S., and Samarati, P., "Balancing confidentiality and efficiency in untrusted relational DBMSs," in *ACM Conference on Computer and Communications Security* (Jajodia, S., Atluri, V., and Jaeger, T., eds.), pp. 93–102, ACM, 2003.

[25] Erkin, Z., Piva, A., Katzenbeisser, S., Lagendijk, R. L., Shokrollahi, J., Neven, G., and Barni, M., "Protection and retrieval of encrypted multimedia content: When cryptography meets signal processing," *EURASIP J. Information Security*, 2007.

[26] Fishman, G., *Discrete-Event Simulation: Modeling, Programming, and Analysis*. Springer Series in Operations Research, Springer, 2001.

[27] Fox, E. A., Chen, Q. F., Daoud, A. M., and Heath, L. S., "Order-preserving minimal perfect hash functions and information retrieval," *ACM Trans. Inf. Syst.*, vol. 9, pp. 281–308, July 1991.

[28] Gentle, J., *Random Number Generation and Monte Carlo Methods*. Statistics and Computing, Springer, 2003.

[29] Gentry, C., "Fully homomorphic encryption using ideal lattices," in *STOC* (Mitzenmacher, M., ed.), pp. 169–178, ACM, 2009.

[30] Goh, E.-J., "Secure indexes," *IACR Cryptology ePrint Archive*, 2003.

[31] Goldreich, O., Goldwasser, S., and Micali, S., "How to construct random functions," *J. ACM*, vol. 33, pp. 792–807, Aug. 1986.

[32] Goldreich, O., Goldwasser, S., and Nussboim, A., "On the implementation of huge random objects," *SIAM J. Comput.*, vol. 39, no. 7, pp. 2761–2822, 2010.

[33] Goldreich, O. and Ostrovsky, R., "Software protection and simulation on oblivious rams," *J. ACM*, vol. 43, no. 3, pp. 431–473, 1996.

[34] Golle, P., Staddon, J., and Waters, B. R., "Secure conjunctive keyword search over encrypted data," in *ACNS* (Jakobsson, M., Yung, M., and Zhou, J., eds.), vol. 3089 of *Lecture Notes in Computer Science*, pp. 31–45, Springer, 2004.

[35] Granboulan, L. and Pornin, T., "Perfect block ciphers with small blocks," in *FSE* (Biryukov, A., ed.), vol. 4593 of *Lecture Notes in Computer Science*, pp. 452–465, Springer, 2007.

[36] Hacigümüs, H., Iyer, B. R., Li, C., and Mehrotra, S., "Executing SQL over encrypted data in the database-service-provider model," in *SIGMOD Conference*, pp. 216–227, 2002.

[37] Hacigümüs, H., Iyer, B. R., and Mehrotra, S., "Efficient execution of aggregation queries over encrypted relational databases," in *DASFAA* (Lee, Y.-J., Li, J., Whang, K.-Y., and Lee, D., eds.), vol. 2973 of *Lecture Notes in Computer Science*, pp. 125–136, Springer, 2004.

[38] Hore, B., Mehrotra, S., and Tsudik, G., "A privacy-preserving index for range queries," in *VLDB* (Nascimento, M. A., Özsu, M. T., Kossmann, D., Miller, R. J., Blakeley, J. A., and Schiefer, K. B., eds.), pp. 720–731, Morgan Kaufmann, 2004.

[39] Indyk, P., Motwani, R., Raghavan, P., and Vempala, S., "Locality-preserving hashing in multidimensional spaces," in *STOC* (Leighton, F. T. and Shor, P. W., eds.), pp. 618–625, ACM, 1997.

[40] Iwata, T. and Kurosawa, K., "OMAC: One-key CBC MAC," in *FSE* (Johansson, T., ed.), vol. 2887 of *Lecture Notes in Computer Science*, pp. 129–153, Springer, 2003.

[41] Iyer, B. R., Mehrotra, S., Mykletun, E., Tsudik, G., and Wu, Y., "A framework for efficient storage security in RDBMS," in *EDBT* (Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., and Ferrari, E., eds.), vol. 2992 of *Lecture Notes in Computer Science*, pp. 147–164, Springer, 2004.

[42] Kachitvichyanukul, V. and Schmeiser, B., "Computer generation of hypergeometric random variates," *Statistical Computation and Simulation*, vol. 22, pp. 127–145, 1985.

[43] Kachitvichyanukul, V. and Schmeiser, B. W., "Algorithm 668: H2PEC: sampling from the hypergeometric distribution.," *ACM Trans. Math. Softw.*, vol. 14, no. 4, pp. 397–398, 1988.

[44] Kantarcioglu, M. and Clifton, C., "Security issues in querying encrypted data," in *DBSec* (Jajodia, S. and Wijesekera, D., eds.), vol. 3654 of *Lecture Notes in Computer Science*, pp. 325–337, Springer, 2005.

[45] Kershaw, D., "Some extensions of W. Gautschi's inequalities for the gamma function," *Mathematics of Computation*, vol. 41, no. 164, pp. 607–611, 1983.

[46] Kurosawa, K. and Ohtaki, Y., "UC-secure searchable symmetric encryption," in *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, Springer, 2012.

[47] Kuzu, M., Islam, M. S., and Kantarcioglu, M., "Efficient similarity search over encrypted data," in *ICDE*, IEEE, 2012.

[48] Li, J., Wang, Q., Wang, C., Cao, N., Ren, K., and Lou, W., "Fuzzy keyword search over encrypted data in cloud computing," in *INFOCOM*, pp. 441–445, IEEE, 2010.

[49] Li, J. and Omiecinski, E., "Efficiency and security trade-off in supporting range queries on encrypted databases.," in *DBSec* (Jajodia, S. and Wijesekera, D., eds.), vol. 3654 of *Lecture Notes in Computer Science*, pp. 69–83, Springer, 2005.

[50] Linial, N. and Sasson, O., "Non-expansive hashing," in *STOC* (Miller, G. L., ed.), pp. 509–518, ACM, 1996.

[51] López-Blázquez, F. and Salamanca-Miño, B., "Exact and approximated relations between negative hypergeometric and negative binomial probabilities.," *Communications in Statistics - Theory and Methods*, vol. 30, no. 5, pp. 957–967, 2001.

[52] Naor, M. and Reingold, O., "On the construction of pseudorandom permutations: Luby-rackoff revisited," *J. Cryptology*, vol. 12, no. 1, pp. 29–66, 1999.

[53] Özsoyoglu, G., Singer, D. A., and Chung, S. S., "Anti-tamper databases: Querying encrypted databases," in *DBSec* (di Vimercati, S. D. C., Ray, I., and Ray, I., eds.), pp. 133–146, Kluwer, 2003.

[54] Pinkas, B. and Reinman, T., "Oblivious RAM revisited," in *CRYPTO* (Rabin, T., ed.), vol. 6223 of *Lecture Notes in Computer Science*, pp. 502–519, Springer, 2010.

[55] Rivest, R., Adleman, L., and Dertouzos, M., "On data banks and privacy homomorphisms," *Foundations of Secure Computation*, pp. 169–177, 1978.

[56] Rogaway, P. and Shrimpton, T., "Deterministic authenticated-encryption: A provable-security treatment of the key-wrap problem," *IACR Cryptology ePrint Archive*, 2006.

[57] Say, A. C. C. and Nircan, A. K., "Random generation of monotonic functions for Monte Carlo solution of qualitative differential equations," *Automatica*, vol. 41, no. 5, pp. 739–754, 2005.

[58] Sedghi, S., van Liesdonk, P., Doumen, J., Hartel, P. H., and Jonker, W., "Computationally efficient searchable symmetric encryption," in *Secure Data Management* (Jonker, W. and Petkovic, M., eds.), vol. 6358 of *Lecture Notes in Computer Science*, pp. 87–100, Springer, 2010.

[59] Shi, E., Bethencourt, J., Chan, H. T.-H., Song, D. X., and Perrig, A., "Multi-dimensional range query over encrypted data," in *IEEE Symposium on Security and Privacy*, pp. 350–364, IEEE Computer Society, 2007.

[60] Smart, N. and Vercauteren, F., "Fully homomorphic encryption with relatively small key and ciphertext sizes." Cryptology ePrint Archive, Report 2009/571, 2009.

[61] Song, D. X., Wagner, D., and Perrig, A., "Practical techniques for searches on encrypted data," in *IEEE Symposium on Security and Privacy*, pp. 44–55, 2000.

[62] van Dijk, M., Gentry, C., Halevi, S., and Vaikuntanathan, V., "Fully homomorphic encryption over the integers," in *EUROCRYPT* (Gilbert, H., ed.), vol. 6110 of *Lecture Notes in Computer Science*, pp. 24–43, Springer, 2010.

[63] Walker, A. J., "An efficient method for generating discrete random variables with general distributions," *ACM Transactions on Mathematical Software*, vol. 3, no. 3, pp. 253–256, 1977.

[64] Wang, W. H. and Lakshmanan, L. V. S., "Efficient secure query evaluation over encrypted XML databases," in *VLDB* (Dayal, U., Whang, K.-Y., Lomet, D. B., Alonso, G., Lohman, G. M., Kersten, M. L., Cha, S. K., and Kim, Y.-K., eds.), pp. 127–138, ACM, 2006.

[65] Xu, J., Fan, J., Ammar, M. H., and Moon, S. B., "Prefix-preserving IP address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," in *ICNP*, pp. 280–289, IEEE Computer Society, 2002.

[66] Yum, D. H. and Lee, P. J., "On the average cost of order-preserving encryption based on hypergeometric distribution," *Inf. Process. Lett.*, vol. 111, no. 19, pp. 956–959, 2011.

VITA

Nathan L. Chenette was born on April 10, 1985, in Grinnell, IA. After graduating as valedictorian of his class at Grinnell High School, he attended college at Harvey Mudd College (HMC) in Claremont, CA. During undergraduate years, he studied abroad at the Budapest Semesters in Mathematics program, took part in a summer REU in linear algebra at Iowa State University in Ames, IA, and was project manager of a HMC Math Clinic research team that partnered with Hewlett-Packard Labs to study a problem in color science. Nathan obtained his BS in mathematics and computer science, with honors, from HMC in 2007. He then enrolled in the PhD program in Algorithms, Combinatorics, and Optimization at Georgia Tech in Atlanta, GA. After a brief stint in graph theory research under Robin Thomas, he began cryptography research with mentor Alexandra Boldyreva, which led to accepted papers and invited presentations at top cryptography conferences EUROCRYPT and CRYPTO. While a PhD student, Nathan spent a summer working at the National Security Agency, a summer at Center for Communications Research in La Jolla, CA, and a semester teaching at Clemson University in Clemson, SC. He plans to obtain his PhD degree in Algorithms, Combinatorics, and Optimization in Summer 2012, after which he will take a Visiting Assistant Professor position at Clemson University, where his wife Heather is a doctoral student.

Symmetric schemes for
efficient range and error-tolerant search
on encrypted data

Nathan L. Chenette

159 Pages

Directed by Professor Alexandra Boldyreva

Large-scale data management systems rely more and more on cloud storage, where the need for efficient search capabilities clashes with the need for data confidentiality. Encryption and efficient accessibility are naturally at odds, as for instance strong encryption necessitates that ciphertexts reveal nothing about underlying data. Searchable encryption is an active field in cryptography studying encryption schemes that provide varying levels of efficiency, functionality, and security, and efficient searchable encryption focuses on schemes enabling sub-linear (in the size of the database) search time. I present the first cryptographic study of efficient searchable symmetric encryption schemes supporting two types of search queries, range queries and error-tolerant queries.

The natural solution to accommodate efficient range queries on ciphertexts is to use order-preserving encryption (OPE). I propose a security definition for OPE schemes, construct the first OPE scheme with provable security, and further analyze security by characterizing one-wayness of the scheme. Efficient error-tolerant queries are enabled by efficient fuzzy-searchable encryption (EFSE). For EFSE, I introduce relevant primitives, an optimal security definition and a (somewhat space-inefficient, but in a sense efficient as possible) scheme achieving it, and more efficient schemes that achieve a weaker, but practical, security notion.

In all cases, I introduce new appropriate security definitions, construct novel schemes, and prove those schemes secure under standard assumptions. The goal

of this line of research is to provide constructions and provable security analysis that should help practitioners decide whether OPE or FSE provides a suitable efficiency-security-functionality tradeoff for a given application.