Fingerprinting Codes and Separating Hash Families

Submitted by

Penying Rochanakul

for the degree of Doctor of Philosophy

of the

Royal Holloway, University of London

2013

Declaration of Authorship

I, Penying Rochanakul, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed.....(Penying Rochanakul)

Date:

Abstract

The thesis examines two related combinatorial objects, namely fingerprinting codes and separating hash families.

Fingerprinting codes are combinatorial objects that have been studied for more than 15 years due to their applications in digital data copyright protection and their combinatorial interest. Four well-known types of fingerprinting codes are studied in this thesis; traceability, identifiable parent property, secure frameproof and frameproof. Each type of code is named after the security properties it guarantees. However, the power of these four types of fingerprinting codes is limited by a certain condition. The first known attempt to go beyond that came out in the concept of two-level traceability codes, introduced by Anthapadmanabhan and Barg (2009). This thesis extends their work to the other three types of fingerprinting codes, so in this thesis four types of two-level fingerprinting codes are defined. In addition, the relationships between the different types of codes are studied. We propose some first explicit non-trivial constructions for two-level fingerprinting codes and provide some bounds on the size of these codes.

Separating hash families were introduced by Stinson, van Trung, and Wei as a tool for creating an explicit construction for frameproof codes in 1998. In this thesis, we state a new definition of separating hash families, and mainly focus on improving previously known bounds for separating hash families in some special cases that related to fingerprinting codes. We improve upper bounds on the size of frameproof and secure frameproof codes under the language of separating hash families.

Acknowledgement

I would like to thank my supervisor, Professor Simon Blackburn, for his continuous support and guidance throughout my time as his student. His wisdom and kindness never ceased to amaze and educate me. He is always available to give helpful advices and critical comments. I am deeply grateful in his patience in proofreading and editing, which contributes immensely to the completion of this thesis.

I would also like to thank my parents for always supporting and encouraging me. Their never-fading faith in me and their unconditional love are my biggest sources of inspiration.

I am thankful to my dear colleagues and friends in the Information Security Group, especially those who are sharing the same office, McCrea355. My life in Egham can never be better without their friendship and accompanying. Special thanks to Amizah Malip for her greatest friendship, coffee and hugs.

I am grateful to Saif Al-Kuwari for providing the latex thesis template and to Marcelo Carlos for helping me overcome many latex problems.

I want to give a big thanks to Viet Pham for being so supportive and spending his time proofreading my meaningless thesis countless of times. His love and understanding put my mind at ease. I am indebted to him for implementing my algorithm in C even though I handed it to him in the last minute.

I would also like to extend my gratitude to all my thai friends who made me realised that the distance can never make us apart. Special credit and best wishes to my best foe, Sukrit Sucharitakul, who constantly maintains our friendship and exchanges encouragement in various forms, including cartoon, games and sarcasm.

Finally, I would like to thank the Development and Promotion of Science and Technology Talents Project (Royal Government of Thailand scholarship) for providing the funding which allowed me to complete my study at university level.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Protection against digital copyright infringement is important, but extremely difficult, especially in the internet age. Digital content, such as music, movies, documents, e-books, games or software, can be copied and distributed easily, resulting in a vast increase in illegal redistribution of data, in other words, *piracy*.

Hardware techniques that can prevent *intellectual property* from being copied freely can be employed. However, these techniques might also restrict an authorised user from doing something legitimate, such as making backup copies of CDs or DVDs, lending materials out through a library, or using copyrighted materials for research and education purposes under fair use laws. As an alternative, we are interested in techniques which, by providing unique identification of data in a certain manner, allow the gathering of evidence against illegal redistribution of data copies. These techniques are commonly known as *fingerprinting*. We are interested in the combinatorial properties behind such techniques, *fingerprinting codes*.

1.1 Motivation

Similar to human fingerprints, which are unique and can be used to identify their owner in the case of a criminal act, digital fingerprints uniquely identify a piece of digital data

and allow the content to be traced to their rightful owner. The fingerprints in each system can vary from a single digit to a cryptographic key.

Even though Wagner [38] first gave a taxonomy of fingerprints and suggested a use for computer software in 1983, the idea of fingerprinting is not something new. It has been used for several hundred years. For example, to protect a logarithm table [14], the publisher made intended errors on the least significant digits of $\log x$ for some values of x. A different choice of error values was made in each copy, so each copy is unique. We call contents in each position that differs between some users a *mark*. In general, a *fingerprint* in each legal copy is a collection of ℓ marks, each mark has q possible values. A collection of fingerprints is referred to as a *code*. In other words, a collection of all fingerprints in the system is a q-ary code of length ℓ and a fingerprint is a *codeword* that belongs to the code. Once a copy of fingerprinted data is sold, the corresponding fingerprint is securely mapped to the purchaser's identity, so that legal responsibility on the use of a copy of data is tied to the customer, while the effectiveness of this mechanism depends on both the design of the code itself as well as the security of the mapping. The scope of this thesis covers only the former concern.

If a naive purchaser redistributed his copy illegally, the fingerprint embedded in that copy will allow the publisher to identify the malicious user and proceed with an appropriate legal action. However, if a group of users pool their copies together, they can detect a part of fingerprint where their copies are different and then modify the fingerprints to create a new illegal copy that differs from their copies to avoid the legal responsibility. They are able to change all the marks they are able to detect to some other value or make them unreadable, and if they are lucky enough, a new copy will be identical to an unfortunate innocent user. We commonly refer to a guilty user as a *traitor* and refer to a group of malicious users as a *coalition*.

The ability of a coalition to create a pirate copy with a new fingerprint is referred to as a *marking assumption*. In this thesis, we assume that a coalition can substitute the marks that they are able to detect by any of the values in the corresponding positions

in their copies. Our marking assumption is known as the *narrow-sense model*. There are some other models that allows a coalition to change the marks to any arbitrary values or make them unreadable. We describe these models in the next section, related work.

Here is an example for a better understanding of the fingerprinting notion:

Broadcast encryption schemes: This example is borrowed from a survey paper by Blackburn [9], which describes an application first introduced by Fiat and Naor [20]. A broadcast company transmits encrypted broadcast content over its network. A session key $\mathcal{K}$, that was used to encrypt the content, is changed regularly. $\mathcal{K}$ is split into ℓ shares, encrypted separately, then transmitted along with the encrypted content. There are ℓ types of key $1, 2, ..., \ell$, each key of type i is able to decrypt the ith share of $\mathcal{K}$. Each type contains q different keys. Each user who purchases a subscription is issued a 'decoder box' with a 'smart card' containing a set of ℓ keys, one of each type. Each key allows a user to decrypt a certain share of $\mathcal{K}$, hence a user is able to obtain all ℓ shares of the $\mathcal{K}$. After obtaining all shares of the session key, a decoder can reconstruct $\mathcal{K}$, then decrypts the content and allows the user to view the content.

In this setting, the illegal redistribution of data itself is costly as it requires continuous decryption and re-broadcasting of the stream. Instead, traitors may redistribute their own decryption keys, or collude and create a new set of pirate keys, that allows other people to decrypt and view the content. The coalition can create a new pirate key set only by simply picking ℓ keys, one per type, from what they have originally.

For additional examples and more detail, we suggest reading a well-written survey paper by Blackburn [9]. The precise combinatorial definitions of fingerprinting codes are given in the next chapter.

There are many types of fingerprinting codes each corresponding to the security properties it guarantees. This thesis focuses on four well-known types of fingerprinting codes: Frameproof Codes (FP codes), Secure Frameproof Codes (SFP codes), Identifiable Parent Property Codes (IPP codes) and Traceability Codes (TA codes). The

precise combinatorial definitions of these codes are given in the next chapter. Properties of each type of code can be summarised as follows, given that the coalition size is at most k: a coalition under a frameproof code cannot frame any innocent user outside the coalition by producing a pirate copy that is identical to that user's copy; two disjoint coalitions under a secure frameproof code can not frame each other by producing the same pirate copy; an intercepted pirate copy produced under an identifiable parent property code can be traced back to at least one member of the coalition that produced it; lastly, in traceability codes, the user that owns a copy of data that is the most similar to a pirate copy is a member of the coalition that produced the pirate copy.

One of the general interests can be summarised as in the following question:

With fixed length ℓ and fixed alphabet size q, what is the largest possible size of the code with a certain fingerprinting? How can a code achieving or approaching such bound be constructed?

The size of the code allows us to estimate the largest possible number of users in our system. The larger the code is, the more users we can have. Generally, this is very important since most content distributors want to sell as many copies of their product as possible. The higher threshold and the stronger property of the codes imply higher security. One of the aims of this thesis is to find the best bounds we can on the size of codes for the four well-known types of fingerprinting codes, assuming that q is large.

All the codes above operate under the assumption that the coalition size is at most k. Once a coalition size is greater than k, there is no guarantee that the code still possesses such properties. In response to that, Anthapadmanabhan and Barg introduced the notion of two-level fingerprinting codes [4] in the context of traceability codes. They suggest the users are divided into several groups of the same size. As in the classical traceability codes, for a given fingerprint from a pirate copy, we are able to identify one of the traitors as long as the coalition size does not exceed a certain threshold k. However, in two-level codes we have an extra property: when the coalition is of a larger size K, greater than k, we are still able to trace one of the groups

containing at least one traitor.

In this thesis, we extend the concept of two-level fingerprinting codes to other types of fingerprinting codes and study bounds on the size of these codes.

In the next section, Section 1.2, we mention some related work that is beyond the scope of this thesis. It is followed by a brief summary of the contents of each of the following chapters of the thesis in Section 1.3.

1.2 Related Work

Here are the four most well-known models of marking assumptions, which refer to the ability of a coalition to create a pirate copy with a new fingerprint. Note that we only work on the narrow-sense model in this thesis.

1. Narrow-sense model: as in an example from broadcast encryption scheme, a coalition can substitute the marks that they are able to detect by any mark from the corresponding positions in their copies.

2. Expanded narrow-sense model: a coalition can substitute the marks that they are able to detect by any mark from the corresponding positions in their copies, or make it unreadable.

3. Wide-sense model: a coalition can substitute the marks that they are able to detect by any arbitrary mark from the alphabet set.

4. Expanded wide-sense model: a coalition can substitute the marks that they are able to detect by any arbitrary mark from the alphabet set, or make it unreadable.

With 4 models and 4 types of codes, there are 16 cases to consider. Panoui [26] shows that expanded narrow-sense model and narrow-sense model are equivalent for frameproof and secure frameproof codes; expanded wide-sense model and wide-sense model are equivalent for identifiable parent property codes. Moreover she shows that

all models are equivalent for traceability codes. Hence, by just considering the narrow-sense model, we cover more than half of the 16 possible cases.

One year after introducing two-level traceability codes under the notion of two-level fingerprinting codes, Anthapadmanabhan and Barg introduce a (claimed to be) stronger definition for their two-level fingerprinting codes [3]. The new codes still possess the property of classical traceability codes for a coalition of size at most k, but instead of an ability to trace a group when a coalition size is at most K, the codes have an ability to prevent an innocent group from being framed which is more similar to frameproof property. One might be interested in how the bounds on the size of the codes vary, under a different combinations of type or marking assumption. However, that is out of the scope of this thesis.

1.3 Structure of the Thesis

This section provides a brief summary of contents of each of the following chapters of the thesis. The high level structure of this thesis is as follows: Chapters 2-3 give the basic knowledge and well-known results on fingerprinting codes; Chapters 4-7 are concerned with new, more powerful, models of fingerprinting codes; and Chapters 8-10 are concerned with improving upper bounds for codes in the classical (and well studied) fingerprinting settings.

Our original contributions appear in various chapters: Chapter 4, 5, 6, 7, 9, and 10. The main contributions of this thesis are in Chapter 5, 9 and 10. We give more details of these contributions at the start of each chapter.

In Chapter 2, we introduce important notation that will be using throughout the thesis, and restate the combinatorial definitions of the four types of fingerprinting codes. We also borrow some basic notation and concepts from coding theory and assume the reader has some familiarity with those concepts. Then, we give the relationships between the different types of fingerprinting codes. Theorems regarding the relationships

between the different types of these codes are borrowed from other papers, but we provide our own proofs. Chapter 3 contains a survey of the important previously known bounds and constructions for each of the four types of fingerprinting codes. Each type of code contributes one section.

We extend the concept of two-level traceability codes [4], to other types of fingerprinting codes in Chapter 4. We state our own definitions. Then, we present relationships between different types of fingerprinting codes we have defined.

In the next chapter, Chapter 5, we propose the first explicit non-trivial constructions for two-level codes in the fingerprint context. Our proposals for two-level fingerprinting codes are suitable for the situation when the number of groups is small, i.e. less than or equal to the size of the alphabet. In Chapters 6 and 7 we propose a different method for constructing two-level frameproof codes and two-level identifiable parent property codes, respectively, with a more general number of groups. Both constructions give codes with at least the same size as a construction that is based on high minimum distance codes; furthermore the resulting codes have a significantly larger size under certain conditions.

In Chapter 8, we introduce another interesting combinatorial object, separating hash families. Separating hash families have various applications in, for instance, Secret sharing schemes [12, 40], Visual cryptography systems [22], Broadcast encryption schemes [20, 32], Key distribution [32], Re-Keying schemes [27] and Traceability schemes [28, 29, 31, 33, 34, 30]. Certain classes of separating hash families are equivalent to frameproof and secure frameproof codes. We discuss how the concept of separating hash families relates to fingerprinting codes, then provide the best previously known bounds for framepoof and secure frameproof codes through the best previously known bounds for separating hash families.

Chapter 9 contains one of the main original contributions of the thesis and is dedicated to improving the previously known bounds for frameproof codes. The chapter is written in the language of separating hash families. We achieve a new tight up-

per bound for the size of frameproof codes when the length ℓ of the code satisfies $\ell = 1 \bmod k$. Our bound is optimal for many set of parameters. This is followed by new tight upper bounds for the size of frameproof codes when the length ℓ of the code satisfies $k < \ell \leq 2k$. Our new bounds are much cleaner and better than the previously known bounds.

Chapter 10 aims to improve previously known bounds for secure frameproof codes. This chapter is written in the language of separating hash families. The first section is devoted to secure frameproof codes when the coalition size is at most 2, when the code has short length. (By short length we mean length 4 or less.) This is followed by the main original contribution of the chapter, the special case of length 5 for which we reduce the size of upper bound by a factor of $\frac{2}{3}$ compared with the best previously known bound. We prove new tight upper bounds for the size of secure frameproof codes when the length ℓ of the code is $2k$. We then explore and improve bounds for secure frameproof codes, when the coalition is of size at most k, of short length ($\ell \leq k$), in the last section using colored graphs.

Lastly, Chapter 11 collects a series of open problems arising from this thesis, including potential future work.

Chapter 2

One-level Fingerprinting Codes

Many different digital fingerprinting schemes have been proposed for the purpose of digital data copyright protection. Out of all these schemes, in this thesis, we focus on studying four well-known types of fingerprinting codes which we will refer to as *one-level* fingerprinting codes. As a result, this early chapter is dedicated to introducing the relevant concepts needed to understand our work. To make is easier and more enjoyable for the reader to follow the definition of each type of codes, we motivate our work with a real life scenario. Imagine that we are a movie seller distributing fingerprinted copies of a movie to customers who purchased it legally. The fingerprint on each copy then serves as an evidence to prosecute the customer in case his or her copy is observed to have been distributed illegally. A coalition of (at most k) smart but adversarial customers might collude to produce a pirate copy with a new fingerprint, thus cheating the copyright protection mechanism. If they upload the pirate copy to a free download site on the internet, this will be a great loss to us. In order to sue for damages, one might be interested in tracing back from the obtained pirate copy to at least one member of the coalition. Then we can put that member in court and expect to learn more about the other members in the coalition from him or her. Or we might at least, given a pirate copy, want to exclude some innocent customers to narrow down the investigation area. Such abilities are considered as key criteria in designing

fingerprinting codes.

In this chapter, we introduce important notation which we use throughout this thesis, and restate the definitions of the classical one-level codes. Then, we give the relationships between the different types of these codes. The results and definitions in this chapter can all be found in the literature, though we do provide our own proofs to some of the results.

2.1 Notation

Let C be a code of length ℓ on an alphabet Q of (finite) size q, i.e. $C \subseteq Q^\ell$. The elements of Q^ℓ are called *words*, and elements of C are called *codewords*. We sometimes refer to both as *fingerprints* to ease our explanation. For copyright protection purposes, fingerprints from a code C are used as follows. First, each copy of the digital data is embedded with a unique fingerprint from C. Then we issue data with a different fingerprint from C to each user/customer. Once a copy is sold, the corresponding fingerprint is securely mapped to the purchaser's identity, so that legal responsibility on the use of a copy of data is tied to the customer.

The *hamming distance* between words x, y will be written as $d_H(x, y)$. Further, for any $X \subseteq Q^\ell$ and $y \in Q^\ell$, let $d_H(X) = \min_{\substack{x,y \in X \\ x \neq y}} d_H(x, y)$ denote the *minimum distance* of X and let $d_H(X, y) = \min_{x \in X} d_H(x, y)$.

For example, let $a = 1111, b = 1100, c = 0001$ and $y = 0110$, and let $X = \{a, b, c\}$. Then $d_H(X, y) = 2$, since $d_H(a, y) = 2$, $d_H(b, y) = 2$ and $d_H(c, y) = 3$.

For each word $x \in Q^\ell$, we write x_i for the ith component of x. For instance, $b_2 = c_4 = 1$ in the example above.

For any positive integer n, denote by $[n]$ the set of integers from 1 to n, in other words, $[n] = \{1, 2, 3, ..., n\}$.

Let ℓ', ℓ'' and ℓ be positive integers such that $\ell = \ell' + \ell''$, let Q' and Q'' be non-empty finite sets. Then for any words $x \in {Q'}^{\ell'}$ and $y \in {Q''}^{\ell''}$, *the concatenation of x and y*,

denoted by $x||y$, is the word $z \in (Q' \cup Q'')^{\ell}$ such that

$$z_i = \begin{cases} x_i & \text{if } i \leq \ell'; \\ y_{i-\ell'} & \text{if } i > \ell', \end{cases}$$

for all $i \in [\ell]$. Similarly, for any subsets $X \subseteq {Q'}^{\ell'}$ and $Y \subseteq {Q''}^{\ell''}$, the *concatenation set of X and Y*, denoted by $X||Y$, is the set

$$Z = \{x||y : x \in X \text{ and } y \in Y\} \subseteq \left(Q' \cup Q''\right)^{\ell}.$$

For example, let $X = \{a, b\}$ where $a = 1111$ and $b = 1100$, and $Y = \{a', b'\}$ where $a' = 101$ and $b' = 010$. Then

$$a||a' = 1111101 \qquad a||b' = 1111010$$
$$b||a' = 1100101 \qquad b||b' = 1100010$$

and so

$$X||Y = \{1111101, 1111010, 1100101, 1100010\}.$$

2.2 Defining One-Level Fingerprinting Codes

Before defining one-level fingerprinting codes, it is necessary to mention the concept of *descendants*. For $X \subseteq Q^{\ell}$, *the set of descendants* of X, denoted $\text{desc}(X)$ is a subset of Q^{ℓ} such that:

$$\text{desc}(X) = \{d \in Q^{\ell} : \forall i \in [\ell], \exists x \in X \text{ such that } d_i = x_i\}.$$

As an example (taken from [9]), let $P = \{1100, 2102, 1122\}$, then

$$\text{desc}(P) = \{1100, 1102, 1120, 1122, 2100, 2102, 2120, 2122\}$$

because the first coordinate can be either 1 or 2, the second coordinate can only be 1 and the third and the last coordinate is either 0 or 2. If we relate this example back to our movie seller scenario, we consider P as a coalition of 3 users. Then, $\text{desc}(P)$ is the set of all possible fingerprints this coalition can produce for its pirate copies under a certain marking assumption.

Let k be a positive integer. For a code C define the *k-descendant code* of C, denoted $\text{desc}_k(C)$, as follows:

$$\text{desc}_k(C) = \bigcup_{\substack{X \subseteq C \\ |X| \leq k}} \text{desc}(X).$$

In other words, $\text{desc}_k(C)$ is the set of all words that are descendants of a coalition of size at most k of C; or $\text{desc}(C)$ is the set of all possible codewords used in pirate copies, created by a coalition of at most k of our customers.

Using the same coalition P as in the previous example and returning to our scenario again, we may think of P as the set of all codewords embedded in copies of the movie we have sold. Only 3 copies of the movie have been sold so far, and codewords hidden inside those copies are the 3 different elements of P. (It looks like we are at the earliest stage of our business.) We now assume that our situation is not too bad, in that at least one of our customers is honest, so $\text{desc}_2(P)$ is the set of all possible fingerprints used in pirate copies, created by a coalition of at most 2 of our customers. In this example,

$$\text{desc}_2(P) = \{1100, 1102, 1120, 1122, 2100, 2102, 2122\}.$$

This is because all possible nonempty subsets of size at most 2 of P, in other words,

all coalitions of size at most 2, are $\{1100\}$, $\{2102\}$, $\{1122\}$, $\{1100, 2102\}$, $\{1100, 1122\}$ and $\{2102, 1122\}$.

Considering all sets of descendants of these subsets, we get $\text{desc}(\{1100\}) = \{1100\}$, $\text{desc}(\{2102\}) = \{2102\}$, $\text{desc}(\{1122\}) = \{1122\}$, $\text{desc}(\{1100, 2102\}) = \{1100, 1102, 2100, 2102\}$, $\text{desc}(\{1100, 1122\}) = \{1100, 1102, 1120, 1122\}$ and $\text{desc}(\{2102, 1122\}) = \{1102, 1122, 2102, 2122\}$. Hence

$$\begin{aligned}\text{desc}_2(P) = &\bigcup_{\substack{X \subseteq C \\ |X| \leq 2}} \text{desc}(X) \\ =&\{1100\} \cup \{2102\} \cup \{1122\} \cup \{1100, 1102, 2100, 2102\} \\ &\cup \{1100, 1102, 1120, 1122\} \cup \{1102, 1122, 2102, 2122\} \\ =&\{1100, 1102, 1120, 1122, 2100, 2102, 2122\}.\end{aligned}$$

We are now ready to define some well known one-level fingerprinting codes, namely Frameproof (FP) codes, Secure frameproof (SFP) codes, Identifiable parent property (IPP) codes and Traceability (TA) codes. The concepts of frameproof, secure frameproof, IPP and traceability codes were first introduced by Boneh and Shaw [14], Stinson, van Trung and Wei [33], Hollman et al. [21] and Chor et al. [15], respectively.

Definition 2.1 (One-Level fingerprinting codes)**.** Let C be a q-ary code of length ℓ and let k be a positive integer.

(i) C has the *k-frameproof* property (or is k-FP) if for all $X \subseteq C$ such that $|X| \leq k$,

$$\text{desc}(X) \cap C \subseteq X.$$

(ii) C has the *k-secure frameproof* property (or is k-SFP) if for all $X_1, X_2 \subseteq C$ of size at most k, $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$ implies $X_1 \cap X_2 \neq \emptyset$.

(iii) C has the *k-identifiable parent property* (or is k-IPP) if for all $x \in \text{desc}_k(C)$, it

holds that

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \mathrm{desc}(X)}} X \neq \emptyset.$$

(iv) C has the *k-traceability* property (or is k-TA) if for all $X \subseteq C$ such that $|X| \leq k$ and for all $x \in \mathrm{desc}(X)$, then $z \in X$ for all $z \in C$ with $d_H(x, z)$ minimal.

From this point onwards we refer to the codes in Definition 2.1 as *one-level fingerprinting codes*, as opposed to two-level fingerprinting codes, which we define in Chapter 4.

Each type of code can be motivated as follows, assuming that the coalition size is at most k. For frameproof codes, no coalition possessing a collection of data, can produce a codeword that does not belong to the coalition. As its name thus suggests, frameproof codes guarantee that no coalition can frame a customer outside the coalition by producing a copy of movie that is identical to that customer's copy. In secure frameproof codes, the descendant sets of two disjoint coalitions are always disjoint, that is, two completely different coalitions of users cannot produce the same new word. Hence they can never frame each other and/or produce copies with the same new fingerprint. If two or more coalitions in an IPP code can produce a common descendant, then they must all have at least one member in common. This means that just from considering the fingerprint in a copy of the movie produced by a coalition, we can always trace back to at least one member of that coalition. Lastly, in traceability codes, a codeword that is most similar to the given descendant, is always a member of the coalition. We are certain that the customer that owns a copy of movie with a fingerprint most similar to the pirate copy is guilty.

Compared to IPP and traceability codes, frameproof and secure frameproof codes do not provide any traceability, and hence they are regarded as codes with weaker properties. This is made clearer when we state the relationships between different

types of one-level codes below.

We now give some examples of one-level fingerprinting codes.

Example 1. $C = \{1100, 1001, 1010\} \subseteq \{0,1\}^4$ is a 2-FP code.

This is easy to see since $\text{desc}(\{1100, 1001\}) \cap C = \{1100, 1001\}$, $\text{desc}(\{1100, 1010\}) \cap C = \{1100, 1010\}$ and $\text{desc}(\{1001, 1010\}) \cap C = \{1001, 1010\}$.

Example 2 ([9])**.** Let Q be an alphabet containing 0, and let ℓ and let k be positive integers. Define $C = \{x \in Q^\ell : \text{there exists a unique } i \in [\ell] \text{ such that } x_i \neq 0\}$. Then C is a k-FP code for any positive integer k.

This is not too difficult to see since each non-zero symbol is used by exactly 1 codeword in each coordinate. Hence, no matter how big the coalition is, it cannot create a codeword outside the coalition.

The next example is an example of a 2-SFP code.

Example 3. Let $C = \{1001, 1200, 0010, 2211\} \subseteq \{0,1,2\}^4$. Then C is a 2-SFP code.

This is also easy to check as $\text{desc}(\{1001, 1200\}) \cap \text{desc}(\{0010, 2211\}) = \emptyset$, $\text{desc}(\{1001, 0010\}) \cap \text{desc}(\{1200, 2211\}) = \emptyset$ and $\text{desc}(\{1001, 2211\}) \cap \text{desc}(\{1200, 0010\}) = \emptyset$.

For a 3-SFP code, we borrow an example by Staddon, Stinson and Wei [29].

Example 4. ([29]) Let $C = \{1111111111, 1111000000, 1000111000, 0100100110, 0010010101, 0001001011\} \subseteq \{0,1\}^{10}$, then C is a 3-SFP code.

The following example from [9] is a 2-IPP code.

Example 5. ([9]) $C = \{0000, 0111, 0222, 1012, 1120, 1201, 2021, 2102, 2210\} \subseteq \{0,1,2\}^4$ is a 2-IPP code.

Our last example is an example of a 2-TA code.

Example 6. $C = \{000, 111, 222\} \subseteq \{0,1,2\}^3$ is a 2-TA code.

This is clear since any word from any two codewords' set of descendants agrees with one of its codewords in at least 2 coordinates.

2.3 Relationships between One-Level Codes

Without much effort one may check from the definitions that the relationships among different types of codes are as follows: k-TA codes are k-IPP codes, k-IPP codes are k-SFP codes and k-SFP codes are k-FP codes. We restate the results from [29] in the next three lemma and theorems. They are not too difficult to verify. However, we provide our own proofs for the sake of completeness.

Lemma 2.3.1 ([29])**.** *A k-TA code is a k-IPP code.*

We include our own proof here.

Proof. Let C be a k-TA code. Let x be a word in $\text{desc}_k(C)$. Let z be a codeword in C such that $d_H(x, z)$ is minimal. By the k-TA property, $z \in X$ for any $X \subseteq C$ such that $\text{desc}(X)$ contains x.
Hence

$$z \in \bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \text{desc}(X)}} X.$$

That is to say

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \text{desc}(X)}} X \neq \emptyset.$$

Therefore, C is a k-IPP code. □

Here we give an example to illustrate Lemma 2.3.1.

Example 7. The 2-TA code C from Example 6 is a 2-IPP code.

We calculate all descendant sets of coalitions of C of size at most 2:

$$\begin{aligned}
\text{desc}(\{000\}) &= \{000\} \\
\text{desc}(\{111\}) &= \{111\} \\
\text{desc}(\{222\}) &= \{222\} \\
\text{desc}(\{000, 111\}) &= \{000, 001, 010, 011, 100, 101, 110, 111\} \\
\text{desc}(\{000, 222\}) &= \{000, 002, 020, 022, 200, 202, 220, 222\} \\
\text{desc}(\{111, 222\}) &= \{111, 112, 121, 122, 211, 212, 221, 222\}
\end{aligned}$$

Hence $\text{desc}_2(C) = \{000, 001, 002, 010, 011, 020, 022, 100, 101, 110, 111, 112, 121, 122, 200, 202, 211, 212, 220, 221, 222\}$.

Considering all words in $\mathrm{desc}_2(C)$, we get

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 000\in\mathrm{desc}(X)}} X = \{000\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 001\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 002\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 010\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 011\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 020\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 022\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 100\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 101\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 110\in\mathrm{desc}(X)}} X = \{000, 111\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 111\in\mathrm{desc}(X)}} X = \{111\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 112\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 121\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 122\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 200\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 202\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 211\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 212\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 220\in\mathrm{desc}(X)}} X = \{000, 222\} \neq \emptyset, \qquad \bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 221\in\mathrm{desc}(X)}} X = \{111, 222\} \neq \emptyset,$$

$$\bigcap_{\substack{X\subseteq C:|X|\leq 2\\ 222\in\mathrm{desc}(X)}} X = \{222\} \neq \emptyset,$$

Hence, C is also 2-IPP.

However, the converse of this lemma is not always true. A counterexample is the code C in Example 5. C is 2-IPP but not 2-TA since $2110 \in \mathrm{desc}(\{2102, 2210\})$, but

$d_H(C, 2110) = d_H(0111, 2110) = 2$ and $0111 \notin \{2102, 2210\}$.

The next two theorems are stated without proof in [29].

Theorem 2.3.2. *A k-IPP code is a k-SFP code.*

Proof. Let C be a k-IPP code. Let X_1, X_2 be subsets of C of size at most k. Assume that $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$. Let $x \in \text{desc}(X_1) \cap \text{desc}(X_2)$. Hence $x \in \text{desc}_k(C)$. Now, by the k-IPP property,

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \text{desc}(X)}} X \neq \emptyset$$

and

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \text{desc}(X)}} X \subseteq X_1 \cap X_2.$$

Hence,

$$X_1 \cap X_2 \neq \emptyset.$$

Therefore, C is a k-SFP code. □

The following example shows that the converse of Theorem 2.3.2 does not always hold.

Example 8. $C = \{1001, 1200, 0010, 2211\} \subseteq \{0, 1, 2\}^4$ is a 2-SFP code, but not a 2-IPP code.

We know from Example 3 that C is 2-SFP. However, C is not 2-IPP, since $1201 \in \text{desc}\{1001, 2211\}$, $1201 \in \text{desc}\{1001, 1200\}$ and $1201 \in \text{desc}\{1200, 2211\}$, but $\{1001, 2211\} \cap \{1001, 1200\} \cap \{1200, 2211\} = \emptyset$.

Theorem 2.3.3. *A k-SFP code is a k-FP code.*

Proof. Let C be a k-SFP code. Let X_0 be a subset of C of size at most k, and let $x \in \text{desc}(X_0) \cap C$. Hence, $\text{desc}(X_0) \cap \{x\} = \text{desc}(X_0) \cap \text{desc}(\{x\}) \neq \emptyset$. By the k-SFP property, $X_0 \cap \{x\} \neq \emptyset$. Then, $x \in X_0$. Therefore, C is a k-FP code. □

The following example shows that a k-FP code is not necessarily a k-SFP code.

Example 9. $C = \{1100, 1001, 1010, 0100\} \subseteq \{0,1\}^4$ is a 2-FP code, but not a 2-SFP code.

It is easy to see that C is a 2-FP code since $\text{desc}(\{1100, 1001\}) \cap C = \{1100, 1001\}$, $\text{desc}(\{1100, 1010\}) \cap C = \{1100, 1010\}$, $\text{desc}(\{1100, 0100\}) \cap C = \{1100, 0100\}$, $\text{desc}(\{1001, 1010\}) \cap C = \{1001, 1010\}$, $\text{desc}(\{1001, 0100\}) \cap C = \{1001, 0100\}$ and $\text{desc}(\{1010, 0100\}) \cap C = \{1010, 0100\}$. But, it is not 2-SFP as $\text{desc}(\{1100, 1001\}) \cap \text{desc}(\{1010, 0100\}) = \{1000, 1100\} \neq \emptyset$, but $\{1100, 1001\} \cap \{1010, 0100\} = \emptyset$.

We can represent the relationships between different types of one-level fingerprinting codes in the following diagram.

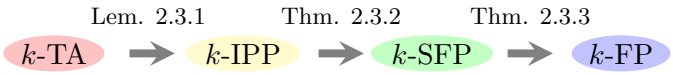


Figure 2.1: Relationships among different types of one-level fingerprinting codes

Chapter 3

Bounds on the Size of One-Level Fingerprinting Codes

This chapter aims to survey the important previously known bounds for each of four types of (one-level) fingerprinting codes, both for completeness of the thesis and convenience of being able to refer to the theorems, under the narrow sense marking assumption. An extensive survey of the other models can be found in [26]. Each section is devoted to one type of code.

3.1 Frameproof codes

Frameproof codes were first introduced by Boneh and Shaw [14] in 1995, under a different marking assumption, the expanded wide-sense model. In 1998, Stinson and Wei [34] provided some constructions for frameproof codes using various combinatorial objects, including separating hash families; these are presented in later chapters of the thesis. Further combinatorial properties of frameproof codes and the other types of fingerprinting codes were studied by Staddon, Stinson and Wei in [29] using the narrow-sense model. Through studying cover-free families and separating hash families they obtained an upper bound on the size of a q-ary k-FP codes of length ℓ as presented

below. Note that Staddon *et al.* state that this bound is valid for all four types of codes they consider.

Theorem 3.1.1 ([29], Theorem 3.7)**.** *Let C be a q-ary k-FP code of length ℓ. Then*

$$|C| \leq k(q^{\lceil \frac{\ell}{k} \rceil} - 1).$$

Later, an open problem in [29] inspired Blackburn to find a new upper bound on the size of k-FP codes as a function of q when ℓ and k are fixed [10]. The bound in that paper, presented below, was obtained from applying extremal set theory related to the Erdos-Ko-Rado Theorem. In the paper, an example of a 3-FP code of size meeting the leading term of the bound was also provided.

Theorem 3.1.2 ([10], Corollary 12)**.** *Let C be a q-ary k-FP code of length ℓ. Then*

$$|C| \leq \left(\frac{\ell}{\ell - (r-1)\lceil \frac{\ell}{k} \rceil} \right) q^{\lceil \frac{\ell}{k} \rceil} + O\left(q^{\lceil \frac{\ell}{k} \rceil - 1} \right),$$

where r is a unique positive integer in $\{1, 2, ..., k\}$ such that $r = \ell \bmod k$.

Note that when the length ℓ of the code does not exceed the size k of the coalition the bounds from Theorem 3.1.1 are better than the bounds from Theorem 3.1.2. However, if $\ell > k$, assuming that q tends to infinity, Theorem 3.1.1 gives a better bound only when $r = k$; Theorem 3.1.2 gives a bound with the best leading term for any other value of r.

Apart from the result above, in the same paper, Blackburn also gave a bound on the size of 2-FP codes, whose leading term is tight, namely $2q^{\lceil \frac{\ell}{2} \rceil}$ when ℓ is even and $q^{\lceil \frac{\ell}{2} \rceil}$ when ℓ is odd; he includes two explicit constructions for 2-FP codes that meet the leading term of the bound.

We give a better bound for the case $r = 1$ in Chapter 10, where we eliminate the $O\left(q^{\lceil \frac{\ell}{k} \rceil - 1} \right)$ term from Theorem 3.1.2 resulting in a new clean and neat bound.

Cohen and Encheva [17] provided an explicit construction for a k-FP code, in the case that the length ℓ of the code does not exceed the size k of the coalition, as follows.

Construction 3.1.1 ([17], Proposition 1)**.** *Let $Q = \{0, 1, ..., q-1\}$. The set C containing all elements of Q^ℓ with exactly one nonzero component forms a k-FP code of cardinality $\ell(q-1)$.*

Note that the resulting codes are of the size that meets the bound in Theorem 3.1.1 in the case $k = \ell$.

Sarkar and Stinson [28] observed that the union of frameproof codes can give a larger frameproof code, then used a recursive construction of separating hash families to show that, given a q-ary k-FP code length ℓ of size n it is possible to construct a $2^i q$-ary k-FP code length ℓ of size $2^i n$, for any positive integer i. Finally, they concluded that there exists an infinite class of q-ary k-FP length ℓ codes of size $O\left(k^{\log \ell} \log \ell\right)$.

The next theorem gives a sufficient condition on the minimum distance of a code to make it a k-FP code.

Theorem 3.1.3 ([9], Theorem 3.1)**.** *Let C be a length ℓ error correcting code with minimum distance d. If $d > (1 - 1/k)\ell$ for some positive integer k, then C is a k-FP code.*

The following frameproof codes construction uses high minimum distance codes that satisfy Theorem 3.1.3. The construction was introduced in [17], also see [10].

Construction 3.1.2 ([17], Theorem 1.1)**.** *Let ℓ and k be positive integers such that $\ell \geq 2$ and $k \geq 2$. Let q be a prime power greater than ℓ. Let $\mathbb{F}_q$ be a finite field of cardinality q, and let $\alpha_1, \alpha_2, ..., \alpha_\ell \in \mathbb{F}_q$ be distinct. Define a code C over $\mathbb{F}_q$ by*

$$C = \{(f(\alpha_1), f(\alpha_2), ..., f(\alpha_\ell)) : f \in \mathbb{F}_q[X] \text{ and } \deg f < \lceil \ell/k \rceil\}.$$

Then, C is a k-FP code of cardinality $q^{\lceil \ell/k \rceil}$.

In this construction, using interpolation, one can retrieve a codeword just by knowing only $\lceil \ell/k \rceil$ of its positions. This is because we use polynomials of degree at most $\lfloor \ell/k \rfloor$ over a finite field.

3.2 Secure frameproof codes

Stinson, van Trung and Wei [33] added some traceability to frameproof codes and introduced them as secure frameproof codes in 1997. Since the frameproof codes only prevent a coalition from framing a user outside the coalition, the distributer cannot trace an illegal copy back to any of the coalition members, so Boneh and Shaw [14] suggested a notion of codes that can trace back to at least one of the coalition members to prevent such a problem. However, they also gave a discouraging result that no such a code exists under the expanded wide-sense model. Stinson, van Trung and Wei slightly weakened the property, resulting in secure frameproof codes. They also gave the first explicit construction and give non-constructive existence results based on probabilistic arguments in [14], again, under the expanded wide-sense model.

Staddon, Stinson and Wei [29] studied secure frameproof codes under the narrow-sense model and obtained the following result using separating hash families.

Theorem 3.2.1 ([29], Theorem 3.10)**.** *Let C be a q-ary k-SFP code of length ℓ. Then*

$$|C| \leq q^{\lceil \frac{\ell}{k} \rceil} + 2k - 2.$$

In 2008, Stinson and Zaverucha presented a bound based on their improved bounds on the size of separating hash families [31].

Theorem 3.2.2 ([31], Corollary 2.8)**.** *Let C be a q-ary k-SFP code of length ℓ. Then*

$$|C| \leq (2k^2 - 3k + 2)q^{\lceil \frac{\ell}{2k-1} \rceil} + 2k^2 + 3k - 1.$$

Recently, Bazrafshan and van Trung [7] provide an upper bound on the maximum

size of k-SFP codes, through the existence of separating hash families.

Theorem 3.2.3 ([7]). *Let C be a q-ary k-SFP code of length ℓ. Then*

$$|C| \leq (2k-1)q^{\lceil \frac{\ell}{2k-1} \rceil}.$$

We later give much better bounds on the size of secure frameproof codes in some special cases through improving the bounds on the size of separating hash family in Chapters 9 and 10.

We mention a few papers that give the explicit constructions for secure frameproof codes using various high minimum distance codes. Cohen, Encheva, Litsyn and Schaathun refered to secure frameproof codes as separating codes [19], then gave a construction involving the concatenation of BCH codes. Encheva and Cohen [18] constructed 2-SFP codes based on Hadamard matrices which are also 3-FP codes. Tonien and Safavi-Naini [37] also presented an explicit construction for 2-SFP codes using Hadamard matrices.

3.3 IPP codes

Identifiable parent property codes were first defined by Hollman, van Lint, Linnart and Tolhuizen [21] in 1998 in the case of coalitions of the size two, i.e., the 2-IPP case. They gave a sufficient condition on the minimum distance of a code so that it is a 2-IPP code, and presented some upper and lower bounds on the size of codes with short length, and codes of arbitrary length. In 2000, Alon, Fischer and Szegedy [1] studied the bounds on the size of 2-IPP codes of length 4. They showed that for any $\epsilon > 0$, there exists $q_0 = q_0(\epsilon)$ such that $|C| < \epsilon q^2$ for every code with alphabet size q, where $q > q_0$.

The k-IPP case was first studied by Staddon, Stinson and Wei in [29] where combinatorial properties of all four types of codes we are interested in were studied. They showed that:

Theorem 3.3.1 ([29], Corollary 2.8)**.** *A q-ary k-IPP code of length ℓ does not exist when $q < k$.*

Blackburn [11] examined the case of short length $\ell < \lfloor (k/2+1)^2 \rfloor$, then used the techniques in the paper [21] to derive a bound for k-IPP codes of arbitrary length as presented below.

Theorem 3.3.2 ([11], Theorem 3)**.** *Let C be a q-ary k-IPP code of length ℓ. Let $u = \lfloor (k/2+1)^2 \rfloor$. Then, we have that*

$$C \leq \frac{1}{2} u(u-1) q^{\lceil n/(u-1) \rceil}.$$

At the same time Alon and Stav [2] used a similar method and derived the following better upper bound independently.

Theorem 3.3.3 ([2], Lemma 2.2, Lemma 2.3)**.** *Let C be a q-ary k-IPP code of length ℓ. Let $u = \lfloor (k/2+1)^2 \rfloor$. Then, we have that*

$$C \leq (u-1) q^{\lceil n/(u-1) \rceil}.$$

Alon and Stav also gave bounds on the size of IPP codes in the case of short length, i.e., when $\ell < \lfloor (k/2+1)^2 \rfloor - 1$, where the size of the code is linear to q. These results can be summarised as follows.

Theorem 3.3.4 ([2], Lemma 4.1, Theorem 4.2, Lemma 4.4)**.** *Let C be a q-ary k-IPP code of length ℓ. Let $u = \lfloor (k/2+1)^2 \rfloor$ and let b and m be positive integers such that $b + m - 1 \leq k$. Then, we have that*

$$C \leq \begin{cases} q & ;\ \textit{when } \ell \leq k \\ \left(1 + \frac{1}{k - \frac{3}{2}} - o(1)\right) q & ;\ \textit{when } \ell \leq k+1 \end{cases}$$

$$C < b(q-1) + m \quad ;\ \textit{when } k+1 < \ell < u-1$$

An example of a code that meets the bound in the first case is a repetition code of any length, so the bound is tight. The bound for the second case was also shown to be tight in [2]. However, the bound on the last case is not always tight.

In [5], Barg, Cohen and Encheva used the probabilistic method to establish the following probabilistic existence results for k-IPP codes; in their case, they fixed k and q, and let ℓ grow.

Theorem 3.3.5 ([5], Lemma 3.5)**.** *Let $u = \lfloor (k/2+1)^2 \rfloor$. There exists a q-ary k-IPP length ℓ code C of size $|C| = q^{R\ell}$ where*

$$R \geq \frac{1}{u-1} \log_q \frac{(q-k)!q^u}{(q-k)!q^u - q!(q-k)^{u-k}}.$$

Yemane [39] fixed k and ℓ, then let q tend to infinity, to establish the following result.

Theorem 3.3.6 ([39])**.** *Let $\epsilon > 0$. There exists a q-ary k-IPP length ℓ code C where*

$$|C| \geq q^{\ell(\frac{1}{u-1}-\epsilon)}.$$

3.4 Traceability codes

Traceability codes were defined by Chor, Fiat and Naor [15] as a scheme to prevent illegal redistribution of digital data (also see their joint work with Pinkas [16]). It was the first type of fingerprinting codes to be introduced. Some constructions and a sufficient condition for a code to be a k-TA code were also introduced at the same time. The next theorem shows a sufficient condition for a code to possess the k-TA property.

Theorem 3.4.1 ([16])**.** *Let C be a length ℓ error correcting code with minimum distance d. If $d > (1-1/k^2)\ell$ for some positive integer k, then C is a k-TA code.*

The theorem above provides a useful and simple construction for certain k-TA codes. Since k-TA code possesses all the properties of the other types of code, the

theorem is also valid for k-IPP, k-SFP and k-FP codes.

Using Theorem 3.4.1, Staddon, Stinson and Wei used Reed-Solomon codes to construct k-TA codes and established the following bounds.

Theorem 3.4.2 ([29], Theorem 4.5)**.** *Suppose q is prime, $\ell < q$ and $t \geq 2$ are integers. Then there exists a q-ary length ℓ k-TA code C such that*

$$|C| \geq q^{\frac{\ell}{k^2}}.$$

Most constructions/examples of k-traceability codes known to the author use an error correcting code that satisfies the condition of Theorem 3.4.1, Theorem 3.4.2 for example. Other constructions that do not depend on error correcting codes often have size tending to zero when q grows. Moreover, for some certain parameters, due to the Plotkin bounds there are no high minimum distance codes that satisfies Theorem 3.4.1. Blackburn, Etzion and Ng [13] explored whether there exist an infinite family of q-ary k-TA codes of rate bounded away from zero in this situation. They used probabilistic techniques to show the following result.

Theorem 3.4.3 ([13])**.** *Let k and q be integers such that $k \geq 2$. When*

$$k^2 - \lceil k/2 \rceil + 1 \leq q$$

or when $k = 2$ and $q = 3$, the following statement holds. There exists a positive constant R (depending on q and k) and a sequence of q-ary k-TA codes $C_1, C_2, \ldots$ with the property that C_ℓ has length ℓ and $|C_\ell| \to q^{R\ell}$ as $\ell \to \infty$.

Recall from before the statement of Theorem 3.1.1 that Staddon, Stinson and Wei [29] provided an upper bound on the size of a q-ary length ℓ k-TA code.

Theorem 3.4.4 ([29], Theorem 3.7)**.** *Let C be a q-ary k-TA code of length ℓ. Then*

$$|C| \leq k(q^{\lceil \frac{\ell}{k} \rceil} - 1).$$

Blackburn, Etzion and Ng also improved the upper bound on the size of 2-TA codes in [13] by showing that there exists a constant c, depending only on ℓ, such that a q-ary 2-TA code of length ℓ contains at most $cq^{\lceil \ell/4 \rceil}$ codewords. This shows that when ℓ is fixed and q is a sufficiently large compared with ℓ, the high minimum distance codes construction produces good q-ary 2-TA codes of length ℓ. In fact, when q is a prime power sufficiently large compared with ℓ, a Reed-Solomon code can be used to construct q-ary 2-TA codes of cardinality $q^{\lceil \ell/4 \rceil}$.

Chapter 4

Two-Level Codes: Definitions and Relationships

In one-level fingerprinting codes, a coalition is restricted to have size k or less. When the size of a coalition exceeds this threshold, one-level fingerprinting codes may no longer retain their properties (e.g. traceability) for such a coalition. This gives rise to the idea of two-level fingerprinting codes, i.e., codes that also give weaker information for large coalitions. This may be best understood with a scenario. A digital document is distributed to several companies, each with equal number of distinct copies. Then those companies assign each copy in their hands to an individual employee. Here each company is acting as a *group* in our two-level model described below. Once a piracy has occurred, apart from tracing back to an individual traitor or protecting an innocent user from being framed, one might be interested in just tracing back to a company that employs one of the coalition members and sue the whole company, or to make sure that none of those companies can cooperate and frame the other innocent companies without getting one of their employees involved in their crime.

Two-level fingerprinting codes were first introduced by Anthapadmanabhan and Barg (2009) in context of traceability (TA) codes [4]. In this chapter, we extend this concept to other types of fingerprinting codes, namely identifiable parent property

(IPP) codes, secure frameproof (SFP) codes, and frameproof (FP) codes. We state the codes' definitions and provide corresponding examples. Then, we present relationships between different types of two-level fingerprinting codes.

4.1 Definitions of Two-Level Fingerprinting Codes

In traditional one-level fingerprinting codes, we assign a different fingerprint from C to each user. Again, C is a code of length ℓ over an alphabet Q of (finite) size q. In two-level fingerprinting codes, we use a code C with a certain specified partition and call it a *two-level code*.

Definition 4.1. Let C be a q-ary length ℓ code containing gp codewords, for some positive integers g and p. Divide C into g disjoint subsets (groups) of p elements each, denoted by $C_1, C_2, ...C_g$. The code $C = C_1 \cup C_2 \cup ... \cup C_g$ together with such a partition is a *two-level code*.

We call C a q-ary length ℓ two-level code, containing g groups of size p. We refer to each C_i as a *group*.

For all $i, j \in \{1, 2, ..., g\}$, we have $|C_i| = p$ and $C_i \cap C_j = \emptyset$ when $i \neq j$. Define $\mathcal{G} : C \to [g]$ by $\mathcal{G}(c) = i$ for all $c \in C_i$, then for any $c \in C$, $\mathcal{G}(c)$ represents its *group index* i.

For any subset X of C, we define $\mathcal{G}(X)$ to be the set of all group indices of elements in X, i.e., $\mathcal{G}(X) = \{\mathcal{G}(x) : x \in X\}$.

Define the *group distance* $d_1(C)$ and the *code distance* $d_2(C)$ as follows:

$$d_1(C) = \min_{\substack{x,y \in C \\ \mathcal{G}(x) \neq \mathcal{G}(y)}} d_H(x, y);$$

$$d_2(C) = d_H(C).$$

Now we are ready to define the four types of two-level fingerprinting codes, we consider in this thesis.

Definition 4.2. Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level q-ary length ℓ code and let K, k be positive integers where $K \geq k$. The code C has the (K, k)-*frameproof* property (or is (K, k)-FP) if

1. C is k-FP when viewed as a q-ary length ℓ code, and
2. C has the *second level frameproof* property, denoted by $(K, *)$-FP: for all $X \subseteq C$ such that $|X| \leq K$ and for all $x \in \text{desc}(X) \cap C$, $\mathcal{G}(x) \in \mathcal{G}(X)$.

The intuition behind two-level FP codes is as follows. A two-level FP code is also a one level FP code, which means that framing an innocent employee is not possible for coalition of size k or less. Further, the second property assures that framing an innocent group (e.g. company in our scenario), i.e., a company that does not employ any coalition member, is also not possible provided that the coalition has size K or less.

Definition 4.3. Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level q-ary length ℓ code and let K, k be positive integers where $K \geq k$. The code C has the (K, k)-*secure frameproof* property (or is (K, k)-SFP) if

1. C is k-SFP when viewed as a q-ary length ℓ code, and
2. C has the *second level secure frameproof* property, denoted by $(K, *)$-SFP: for all $X_1, X_2 \subseteq C$ of size at most K, if $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$, then $\mathcal{G}(X_1) \cap \mathcal{G}(X_2) \neq \emptyset$.

A (K, k)-SFP code is a k-SFP code. The second property ensures that a coalition of K or less companies cannot frame a coalition from other disjoint group of K or less companies. Here by coalition of companies we mean a coalition of employees, with at least one employee from each of those companies.

Definition 4.4. Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level q-ary length ℓ code and let K, k be positive integers where $K \geq k$. The code C has the (K, k)-*identifiable parent property* (or is (K, k)-IPP) if

1. C is k-IPP when viewed as a q-ary length ℓ code, and

2. C has the *second level identifiable parent* property, denoted by $(K, *)$-IPP: for all $x \in \text{desc}_K(C)$,

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \neq \emptyset.$$

Two-level IPP codes allow the identification of a group containing a parent, provided that the parent coalition has size K or less. Thus, by just considering an illegal fingerprint in a copy of the data, we can always trace back to at least one company that employs a member of the coalition.

Definition 4.5 ([4])**.** Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level q-ary length ℓ code and let K, k be positive integers where $K \geq k$. The code C has the (K, k)-*traceability* property (or is (K, k)-TA) if

1. C is k-TA when viewed as a q-ary length ℓ code, and

2. C has the *second level traceability* property, denoted by $(K, *)$-TA: for all $X \subseteq C$ that $|X| \leq K$ and for all $x \in \text{desc}(X)$,
   for all $z \in C$ with $d_H(x, z)$ minimal (i.e. $d_H(x, z) = d_H(C, x)$), then $\mathcal{G}(z) \in \mathcal{G}(X)$.

Two-level TA codes also possess one-level TA property, so that the employee that owns the copy of data most similar to the pirate copy is guilty when the coalition size is at most k. Additionally, the second level property shows that the company that owns the copy of data most similar to the pirate copy is responsible for the crime, provided the coalition size is at most K.

We refer to (K, k)-FP codes, (K, k)-SFP codes, (K, k)-IPP codes and (K, k)-TA codes as *two-level fingerprinting codes*, and refer to all codes in Definition 2.1 as *one-level fingerprinting codes*.

Here are examples of two-level fingerprinting codes in the same order as their definitions.

Example 10. Let $C = C_1 \cup C_2 \cup C_3 \subseteq \{0,1,2,3\}^3$, where

$$C_1 = \{100, 010, 001\}$$
$$C_2 = \{200, 020, 002\}$$
$$C_3 = \{300, 030, 003\}.$$

Then, C is a (K,k)-FP code for any positive integers K and k such that $K \geq k$.

Example 2 shows that C is k-FP for any positive integer k. Also, the $(K,*)$-FP property follows from the fact that, for any $i \in [3]$, only codewords in group i contain the symbol i.

Example 11. Let $C = C_1 \cup C_2 \subseteq \{0,1\}^{10}$, where

$$C_1 = \{0100100110, 0010010101, 0001001011\}$$
$$C_2 = \{1111111111, 1111000000, 1000111000\}.$$

Then, C is a $(K,3)$-SFP code for any integer $K \geq 3$.

We know from Example 4 that C is 3-SFP. Since for any $i \in [2]$, only codewords in group i contain the symbol $i-1$ in the first coordinate, C is $(K,*)$-SFP.

Example 12. Let $C = C_1 \cup C_2 \cup C_3 \subseteq \{0,1,2\}^4$, where

$$C_1 = \{0000, 0111, 0222\}$$
$$C_2 = \{1012, 1120, 1201\}$$
$$C_3 = \{2021, 2102, 2210\}.$$

Then, C is a $(K,2)$-IPP code for any positive integer $K \geq 2$.

From Example 5, C is a 2-IPP code. The $(K, *)$-IPP property arises from observing that, for any $i \in [3]$, only codewords in group i contain the symbol $i-1$ in the first coordinate.

The next example is generalised from a family of 3-TA codes by Blackburn, Etzion and Ng [13].

Example 13. Let $q = kr + 1$, where k is an integer and r is a positive integer. Let $Q = \{0, 1, ..., kr\}$.
Define $C = C_1 \cup C_2 \cup ... \cup C_{k+1}$, where

$$\begin{aligned}
C_1 &= \{(0, i, i, ..., i) : i \in [r]\} \\
C_2 &= \{(i, 0, r+i, ..., r+i) : i \in [r]\} \\
C_3 &= \{(r+i, r+i, 0, 2r+i, ..., 2r+i) : i \in [r]\} \\
&\vdots \\
C_k &= \{((k-2)r+i, ..., (k-2)r+i, 0, (k-1)r+i) : i \in [r]\} \\
C_{k+1} &= \{((k-1)r+i, ..., (k-1)r+i, 0) : i \in [r]\}
\end{aligned}$$

Then C is a two-level code containing $k+1$ groups of size r that has the $((k+1)r, k)$-TA property.

Proof. We will show the k-TA property first, and then the $((k+1)r, *)$-TA property.

Consider a coalition $X_0 \subseteq C$ of size at most k. Let $y \in \text{desc}(X_0)$ and let $z \in C$ such that $d_H(y, z)$ minimal. Define $I = \{i \in [k+1] : z_i = y_i\}$. Then $|I|$ must be at least 2. Let $j \in I$ be such that $y_j \neq 0$. So, there exists $x \in X_0$ that $x_j = y_j$. Since there exists only one codeword in C that has y_j in the jth position, $z = x$. Therefore $z \in X_0$, which shows C is k-TA.

Let $X_1 \subseteq C$ be a coalition of size at most $|C| = (k+1)r$ and let $y' \in \text{desc}(X_1)$. Let z' be a codeword in C that minimises $d_H(y', z')$ and define $I = \{i \in [k+1] : z'_i = y'_i\}$. It is obvious that I has at least 1 element. Let $j \in I$. When $y'_j \neq 0$, the situation

is similar to the previous paragraph. Consider the case when $y'_j = 0$. There exists $x' \in X_1$ such that $x'_j = 0$. Hence, both x' and z' are from C_j. Therefore $z' \in \mathcal{G}(U)$.

Here we conclude that C is a $((k+1)r, k)$-TA code. □

4.2 Relationships between Two-Level Codes

The relationships among different types of two-level fingerprinting codes and one-level ingerprinting codes are illustrated in Diagram 4.1. An arrow from type A to type B signifies that a code of type A is a code of type B.

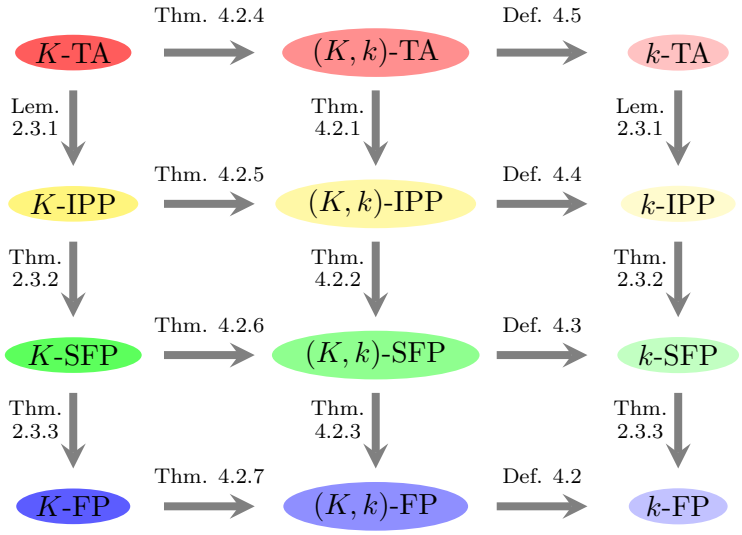


Figure 4.1: Relationships among different types of fingerprinting codes

The proofs of all the relationships in Diagram 4.1 are straightforward. We start by establishing the relationships in the middle column, then continue to the relationships given by the left horizontal arrows. The implications corresponding to the vertical arrows at the sides of the diagram have been proved in Chapter 2.

Theorem 4.2.1. *A (K, k)-TA code is a (K, k)-IPP code.*

Proof. Let C be a (K, k)-TA code. Hence, C is k-TA when viewed as a one-level code. By Lemma 2.3.1, C is a k-IPP code.

Let $x \in \text{desc}_K(C)$ and let D_x be a set of all elements z in C that give the minimum value of $d_H(z, x)$. Let $X_0 \subseteq C$ be of size at most K with $x \in \text{desc}(X_0)$. By the definition of two-level traceability code, we know that $\mathcal{G}(z) \in \mathcal{G}(X_0)$ for all $z \in \mathcal{G}(D_x)$. Hence $\mathcal{G}(D_x) \subseteq \mathcal{G}(X_0)$. Then, we have

$$\mathcal{G}(D_x) \subseteq \bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X).$$

Consequently,

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \neq \emptyset.$$

Therefore C is a (K, k)-IPP code. □

Theorem 4.2.2. *A (K, k)-IPP code is a (K, k)-SFP code.*

Proof. Let C be a (K, k)-IPP code. Hence, C is k-IPP when viewed as a one-level code. By Theorem 2.3.2, C is a k-SFP code.

Let X_1, X_2 be subsets of C of size at most K. Assume that $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$. Let $x \in \text{desc}(X_1) \cap \text{desc}(X_2)$. Hence $x \in \text{desc}_K(C)$.

Since

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \neq \emptyset$$

and

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \subseteq \mathcal{G}(X_1) \cap \mathcal{G}(X_2),$$

we find

$$\mathcal{G}(X_1) \cap \mathcal{G}(X_2) \neq \emptyset.$$

Therefore C is a (K, k)-SFP code. □

Theorem 4.2.3. *A (K, k)-SFP code is a (K, k)-FP code.*

Proof. Let C be a (K, k)-SFP code. Hence C is k-SFP when viewed as a one-level code. By Theorem 2.3.3, C is a k-FP code.

Let X_0 be a subset of C of size at most K, and let $x \in \text{desc}(X_0) \cap C$. Hence $\mathcal{G}(x) \in \mathcal{G}(\text{desc}(X_0)) \cap \mathcal{G}(\{x\}) = \mathcal{G}(\text{desc}(X_0)) \cap \mathcal{G}(\text{desc}(\{x\}))$, which implies $\mathcal{G}(\text{desc}(X_0)) \cap \mathcal{G}(\text{desc}(\{x\})) \neq \emptyset$. By the $(K, *)$-SFP property, $\mathcal{G}(X_0) \cap \mathcal{G}(\{x\}) \neq \emptyset$. Then $\mathcal{G}(x) \in \mathcal{G}(X_0)$.

Therefore C is a (K, k)-FP code. □

The converse of Theorems 4.2.1, 4.2.2 and 4.2.3 are not true. The counterexamples used for the one-level cases can be used again here, since two-level codes possess the one-level code properties.

Theorem 4.2.4. *A K-TA code is a (K, k)-TA code.*

Proof. Let C be a K-TA code. It is easy to see that C is also a k-TA code for any $k \leq K$. Let X_0 be a subset of C of size at most K, let $x \in \text{desc}(X_0)$ and let D_x be a set of all elements z in C that give the minimum value of $d_H(z, x)$. By the definition of K-TA code, we know that $z \in X_0$ for all $z \in D_x$. Hence $\mathcal{G}(z) \in \mathcal{G}(X_0)$ for all $z \in D_x$. So, C is a (K, k)-TA code. □

Theorem 4.2.5. *A K-IPP code is a (K, k)-IPP code.*

Proof. Let C be a K-IPP code. Then C is also a k-IPP code for any $k \leq K$. Let

$x \in \text{desc}_K(C)$. By the definition of K-IPP code,

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} X \neq \emptyset.$$

Hence,

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \supseteq \mathcal{G}\Big(\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} X \Big) \neq \emptyset.$$

Therefore C is a (K, k)-IPP code. □

Theorem 4.2.6. *A K-SFP code is a (K, k)-SFP code.*

Proof. Let C be a K-SFP code. Thus C is also a k-SFP code for any $k \leq K$. Let X_1, X_2 be subsets of C of size at most K such that $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$. By the definition of K-SFP code, $X_1 \cap X_2 \neq \emptyset$. Hence, $\mathcal{G}(X_1) \cap \mathcal{G}(X_2) \neq \emptyset$. Which implies C is a (K, k)-SFP code. □

Theorem 4.2.7. *A K-FP code is a (K, k)-FP code.*

Proof. Let C be a K-FP code. It is easy to see that C is also a k-FP code for any $k \leq K$. Let X_0 be a subset of C of size at most K, and let $x \in \text{desc}(X_0) \cap C$. By the definition of K-FP code,

$$\text{desc}(X_0) \cap C \subseteq X_0.$$

Therefore,

$$\mathcal{G}(\text{desc}(X_0) \cap C) \subseteq \mathcal{G}(X_0).$$

This makes C a (K, k)-FP code as required. □

Chapter 5

Two-Level Code Constructions

All two-level fingerprinting codes, by definition, possess the properties of their corresponding one-level fingerprinting codes. It is thus natural to consider constructing two-level fingerprinting codes from existing one-level codes.

In this chapter, we define and propose the first explicit non-trivial constructions for two-level IPP, SFP and FP codes. Our proposals for two-level fingerprinting codes are suitable for the situation when the number of groups is small, i.e. less than or equal to the size of the alphabet.

We explain our two constructions in the following sections.

5.1 A Simple Construction

Straight from the definitions, it is easy to see that the upper bounds on the size of one-level codes are also relevant to two-level codes. The construction we present here is simple and provides two-level codes that meet the best existence bounds for one-level codes. In some special cases, the simple construction is better than the more general construction in the next section. The idea is to take a one-level code, and choose a suitable partition to produce a two-level code with the properties we want.

Example 14. Let C be a k-FP code on an alphabet Q, constructed as in Construction

3.1.2. Then, for any integer $K > k$, there exists a (K, k)-FP code C' of the same cardinality as C with q groups, where $q = |Q|$.

Proof. Partition C into q groups, $C_1, C_2, ..., C_q$, by letting $C_i = \{x \in C : x_1 = i\}$. Let C' be $C_1 \cup C_2 \cup ... \cup C_q$. Since $|C| = q^{\lceil \ell/k \rceil}$ we have a code C' containing q groups of cardinality $q^{\lceil \ell/k \rceil - 1}$. It is easy to see that $\mathcal{G}(\text{desc}(X)) = \mathcal{G}(X)$ for any subset X of C'. Hence none of the coalitions of C' can produce the codeword that belong to a group outside their own group. Hence C' has the $(K, *)$-FP property. The k-FP property holds for C' since it holds for C. So we see that C' is a (K, k)-FP code of the same cardinality as C. □

Example 14 has shown that there exist (K, k)-FP codes as large as k-FP codes in some cases. In fact, for any k-IPP, k-SFP or k-FP code C, if there exists a coordinate i such that the symbols appearing in that coordinate of codewords in C occur equally often, we can construct a two-level code by partitioning the codewords of C into groups, so that codewords in each group have the same symbol in the ith coordinate. As the symbols in the ith coordinate are uniformly distributed, these groups have the same size and we thus obtain a two-level code C' with the same cardinality as C. The construction can be rewritten as follows.

Construction 5.1.1. *Let q, g and ℓ be integers greater than 1, where $g \leq q$. Let C be an q-ary code of length ℓ. Suppose there exists $i \in [\ell]$ such that only $g \leq q$ symbols $q_1, q_2, ..., q_g$ from Q occur as the ith coordinate of codewords, and the symbols occur equally often.*

For each $j \in [g]$, let $C_i = \{x \in C : x_i = q_j\}$. Then, $C = C_1 \cup C_2 \cup ... \cup C_g$ and $|C_j| = \frac{|C|}{g}$ for all $j \in [g]$, and so we have a two-level code.

We now show that this simple construction is valid for any k-IPP, k-SFP and k-FP code.

Theorem 5.1.1. *Let q, g and ℓ be integers greater than 1, where $g \leq q$. Let C be an q-ary length ℓ k-FP code. If there exists $i \in [\ell]$ such that only $g \leq q$ symbols from Q occur*

as the ith coordinate of codewords, and the symbols occur equally often, Construction 5.1.1 gives a q-ary length ℓ two-level (K, k)-FP code containing g groups of the same size.

Proof. Since C is k-FP, the only thing we need to prove is the $(K, *)$-FP property.

Let X be any subset of C containing at most K codewords. Let $x \in \text{desc}(X) \cap C'$. We show that $\mathcal{G}(x) \in \mathcal{G}(X)$. Since $x \in \text{desc}(X) \cap C$, we have $\mathcal{G}(x) \in \mathcal{G}(\text{desc}(X))$. However, the group index can be identified by the ith coordinate only. Therefore $\mathcal{G}(\text{desc}(X)) = \mathcal{G}(X)$, which implies $\mathcal{G}(x) \in \mathcal{G}(X)$.

Thus, we have shown that $C = C_1 \cup C_2 \cup ... \cup C_g$ is a q-ary length ℓ two-level (K, k)-FP code containing g groups of the same size. □

Theorem 5.1.2. *Let q, g and ℓ be integers greater than 1, where $g \leq q$. Let C be an q-ary length ℓ k-SFP code. If there exists $i \in [\ell]$ such that only $g \leq q$ symbols from Q occur as the ith coordinate of codewords, and the symbols occur equally often, Construction 5.1.1 gives a q-ary length ℓ two-level (K, k)-SFP code containing g groups of the same size.*

Proof. Since C is k-SFP, we only need to prove the $(K, *)$-SFP property.

Let X_1, X_2 be any two subsets of C containing at most K codewords such that $\mathcal{G}(\text{desc}(X_1)) \cap \mathcal{G}(\text{desc}(X_2)) \neq \emptyset$. We will show that $\mathcal{G}(X_1) \cap \mathcal{G}(X_2) \neq \emptyset$. Since the group index is identified by the ith coordinate only, $\mathcal{G}(\text{desc}(X_1)) = \mathcal{G}(X_1)$ and $\mathcal{G}(\text{desc}(X_2)) = \mathcal{G}(X_2)$. Hence $\mathcal{G}(X_1) \cap \mathcal{G}(X_2) = \mathcal{G}(\text{desc}(X_1)) \cap \mathcal{G}(\text{desc}(X_2)) \neq \emptyset$. Thus, $C = C_1 \cup C_2 \cup ... \cup C_g$ is a q-ary length ℓ two-level (K, k)-SFP code containing g groups of the same size. □

Theorem 5.1.3. *Let q, g and ℓ be integers greater than 1, where $g \leq q$. Let C be an q-ary length ℓ k-IPP code. If there exists $i \in [\ell]$ such that only $g \leq q$ symbols from Q occur as the ith coordinate of codewords, and the symbols occur equally often, Construction 5.1.1 gives a q-ary length ℓ two-level (K, k)-IPP code containing g groups of the same size.*

Proof. Since C is k-IPP, now we only need to show the $(K, *)$-IPP property.

Let $x \in \text{desc}_K(C)$. There exists a subset X_0 of C containing at most K codewords such that $x \in \text{desc}(X_0)$. Since the group index can only be identified by the ith coordinate, for all $X \subseteq C$ such that $|X| \leq K$ and $x \in \text{desc}(X)$, $\mathcal{G}(X)$ must contain x_i. Hence

$$\bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) \neq \emptyset.$$

Therefore $C = C_1 \cup C_2 \cup ... \cup C_g$ is a q-ary length ℓ two-level (K, k)-SFP code containing g groups of the same size. □

Construction 5.1.1 is only useful for two-level IPP, SFP and FP codes, but not for two-level TA codes (see example 15 below). This is because there is no guarantee that the nearest codeword will carry the correct symbol on the coordinate we use to construct our partition C_i. Construction 5.1.1 can be used to produce a two-level code without causing any change to the code, when a one-level code has a coordinate such that the distribution of alphabet symbols is uniform. Even when the distribution is almost uniform, Construction 5.1.1 can still be used once we remove some codewords to make the code uniform in a coordinate. However, there are one-level codes such that the distribution of alphabet symbols in any coordinate is non-uniform (and there are examples that are the largest known for some parameters). In this case, the above simple two-level construction cannot be used. The construction we propose in the next section is general enough to work in these cases, though at a cost of reducing the size of the code by a factor of up to 2.

Example 15. Let C be a 2-TA code as follows,

$$\{0000, 0111, 1102, 1210, 2012, 2120\}.$$

Partition C by the first coordinate, then C is not a (3,2)-TA code.

Proof. Consider the word $0012 \in \text{desc}(\{0000, 0111, 1102\})$, we have $d_H(C, 0012) = 1 = d_H(2012, 0012)$. However, $\mathcal{G}(2012) = 2$ is not a member of $\mathcal{G}(\{0000, 0111, 1102\}) = \{0, 1\}$. Thus, C is not a (3,2)-TA code. □

5.2 A General Construction

In this section, we aim to construct two-level fingerprinting codes from existing one-level codes. Our construction produces codes whose number of groups is at most the alphabet size. It begins with a one-level code, and involves removal of some codewords, as well as grouping and modifying the remaining codewords. The results are two-level codes which are guaranteed to be at least half the size of the original one-level codes.

The next theorem is the core of our construction.

Theorem 5.2.1. *Let q, g and ℓ be integers greater than 1, where $g \leq q$. Let C be a q-ary length ℓ code. Then there exists a q-ary length ℓ code C' of cardinality at least $\frac{|C|}{2}$, where C' possesses the following properties;*

1. *there exists an injection from C' to C with changes occuring only in the first coordinate of the codewords,*
2. *C' can be partitioned into g groups of the same size, each with at least $\left\lceil \frac{|C|}{2g} \right\rceil$ codewords,*
3. *the first coordinate of codewords in each group of C' are distinct from those of any other group.*

The explicit construction of two-level codes is embedded in the proof of Thorem 5.2.1. Before giving the detailed proof, we provide the following example to give a rough idea about how to construct C' from a given code C satisfying Theorem 5.2.1.

Example 16. Let $n = 91$, $q = 11$, $g = 9$ and $p = \left\lceil \frac{n}{2g} \right\rceil = 6$. Let g_i be the number of

codewords of C beginning with the symbol i, for $i \in [q]$. Suppose we have:

$$g_1 = 4 \qquad g_2 = 5 \qquad g_3 = 10 \qquad g_4 = 11$$
$$g_5 = 17 \qquad g_6 = 5 \qquad g_7 = 2 \qquad g_8 = 4$$
$$g_9 = 18 \qquad g_A = 10 \qquad g_B = 5.$$

Our aim is to form new 9 groups of size 6, where the first coordinate of codewords in each new group are distinct from any other group. Our method can be divided into 3 main steps: splitting, merging and replacing. We illustrate the construction of C' in Table 5.1.

Step 1: Splitting

The purpose of this step is to split all big groups, which contain p codewords or more, into one or more smaller groups of size at least p. Observe that g_3, g_4 and g_A provide 1 group each, while g_5 and g_9 give 2 and 3 groups, respectively. We now have 8 groups of size at least $p = 6$, call them $C_1, C_2, ..., C_8$, with only 5 different first coordinates. Also, remove any 3 other groups, for instance, g_2, g_6 and g_B, so that we have room available to add 3 more first coordinates. Hence, 2, 6 and B have become unused first coordinates.

Step 2: Merging

In this step, we aim to create more groups, of size at least p, by merging at least 2 smaller groups together. So, we merge the remaining groups together to obtain the last group C_9 of size $g_1 + g_7 + g_8 = 10$.

Step 3: Replacing

Since some first coordinates of the groups we have constructed are repeated, we replace them with those unused first coordinates in this step. To do so, we first reduce the number of codewords in each group to p. Then, let $\{a_i : i \in [8]\}$ be a permutation of $\{2, 3, 4, 5, 6, 9, A, B\}$, we replace the first coordinate of codewords in each group C_i (excluding C_9) by a_i.

The result after completing these 3 steps is 9 disjoint groups of size 6, where the first coordinate of codewords in each group are distinct from any other groups.

Table 5.1 illustrates what we did earlier. We use c^i and c'^i to denote the ith codeword of C and C', respectively. For a codeword c beginning with a symbol k, we represented c by $c : k * *...*$. We use g'_k to denote the new number of codewords beginning with the symbol k.

We now prove Theorem 5.2.1 using a similar approach to Example 16.

Proof of Theorem 5.2.1. Let q and ℓ be integers greater than 1. Let C be a one-level code length ℓ over an alphabet Q, where $|Q| = q$. Let g be a positive integer less than or equal to q, and let $p = \left\lceil \frac{|C|}{2g} \right\rceil$. For each symbol $a \in Q$, let $G_a = \{x \in C : x_1 = a\}$ and denote the size of G_a by g_a. Let $g_a = \alpha_a p + \beta_a$, where α_a, β_a are integers such that $0 \leq \beta_a < p$. Let Q_1 be $\{a \in Q : \alpha_a > 0\}$, $q_1 = |Q_1|$ and $v = \sum_{a \in Q} \alpha_a = \sum_{a \in Q_1} \alpha_a$. As in Example 16, we construct g groups of size p using three main steps: splitting, merging and replacing.

Step 1: Splitting

For each $a \in Q_1$, we pick $\alpha_a p$ codewords from each G_a, then divide these codewords into α_a sets of p codewords. At this stage, we obtain v disjoint sets of p codewords, call them $C_1, C_2, ..., C_v$, with the property that all the codewords within the same set have the same symbol in the first coordinate. However, some of the symbols are still being used by more than one group.

If $v \geq g$ we are done: for $i \in [g]$, we replace the first coordinate of the codewords in C_i by the symbol i, to form C'_i, and define $C' = \bigcup_{i=1}^{g} C'_i$. So without loss of generality, we may assume $v < g$. To construct the first v groups from $C_1, C_2, ..., C_v$, we need v different symbols to replace the first coordinate of each group. Besides the q_1 symbols in Q_1, we need $v - q_1$ extra symbols. Let Q_2 be any subset of Q of cardinality v containing Q_1. We discard all codewords in G_a where $a \in Q_2 \backslash Q_1$.

Step 2: Merging

Original group size	Original codewords	New codewords	New group size
g_2=5	$c^{11} : 2 * *...*$	Discarded	Removed
	$\vdots$	$\vdots$	
	$c^{15} : 2 * *...*$	Discarded	
g_3=10	$c^{16} : 3 * *...*$	$c'^{7} : 2 * *...*$	$g'_2 = 6$
	$c^{17} : 3 * *...*$	$c'^{8} : 2 * *...*$	
	$c^{18} : 3 * *...*$	$c'^{9} : 2 * *...*$	
	$c^{19} : 3 * *...*$	$c'^{10} : 2 * *...*$	
	$c^{20} : 3 * *...*$	$c'^{11} : 2 * *...*$	
	$c^{21} : 3 * *...*$	$c'^{12} : 2 * *...*$	
	$c^{22} : 3 * *...*$	Discarded	
	$\vdots$	$\vdots$	
	$c^{25} : 3 * *...*$	Discarded	
g_4=11	$c^{26} : 4 * *...*$	$c'^{13} : 3 * *...*$	$g'_3 = 6$
	$c^{27} : 4 * *...*$	$c'^{14} : 3 * *...*$	
	$c^{28} : 4 * *...*$	$c'^{15} : 3 * *...*$	
	$c^{29} : 4 * *...*$	$c'^{16} : 3 * *...*$	
	$c^{30} : 4 * *...*$	$c'^{17} : 3 * *...*$	
	$c^{31} : 4 * *...*$	$c'^{18} : 3 * *...*$	
	$c^{32} : 4 * *...*$	Discarded	
	$\vdots$	$\vdots$	
	$c^{36} : 4 * *...*$	Discarded	
g_5=17	$c^{37} : 5 * *...*$	$c'^{19} : 4 * *...*$	$g'_4 = 6$
	$c^{38} : 5 * *...*$	$c'^{20} : 4 * *...*$	
	$c^{39} : 5 * *...*$	$c'^{21} : 4 * *...*$	
	$c^{40} : 5 * *...*$	$c'^{22} : 4 * *...*$	
	$c^{41} : 5 * *...*$	$c'^{23} : 4 * *...*$	
	$c^{42} : 5 * *...*$	$c'^{24} : 4 * *...*$	
	$c^{43} : 5 * *...*$	$c'^{25} : 5 * *...*$	$g'_5 = 6$
	$c^{44} : 5 * *...*$	$c'^{26} : 5 * *...*$	
	$c^{45} : 5 * *...*$	$c'^{27} : 5 * *...*$	
	$c^{46} : 5 * *...*$	$c'^{28} : 5 * *...*$	
	$c^{47} : 5 * *...*$	$c'^{29} : 5 * *...*$	
	$c^{48} : 5 * *...*$	$c'^{30} : 5 * *...*$	
	$c^{49} : 5 * *...*$	Discarded	
	$\vdots$	$\vdots$	
	$c^{53} : 5 * *...*$	Discarded	
g_6=5	$c^{54} : 6 * *...*$	Discarded	Removed
	$\vdots$	$\vdots$	
	$c^{58} : 6 * *...*$	Discarded	

Original group size	Original codewords	New codewords	New group size
g_1=4	$c^{1} : 1 * *...*$	$c'^{1} : 1 * *...*$	$g'_{1,7,8} = 6$
	$c^{2} : 1 * *...*$	$c'^{2} : 1 * *...*$	
	$c^{3} : 1 * *...*$	$c'^{3} : 1 * *...*$	
	$c^{4} : 1 * *...*$	Discarded	
g_7=2	$c^{5} : 7 * *...*$	$c'^{4} : 7 * *...*$	
	$c^{6} : 7 * *...*$	Discarded	
g_8=4	$c^{7} : 8 * *...*$	$c'^{5} : 8 * *...*$	
	$c^{8} : 8 * *...*$	$c'^{6} : 8 * *...*$	
	$c^{9} : 8 * *...*$	Discarded	
	$c^{10} : 8 * *...*$	Discarded	
g_9=18	$c^{59} : 9 * *...*$	$c'^{31} : 6 * *...*$	$g'_6 = 6$
	$c^{60} : 9 * *...*$	$c'^{32} : 6 * *...*$	
	$c^{61} : 9 * *...*$	$c'^{33} : 6 * *...*$	
	$c^{62} : 9 * *...*$	$c'^{34} : 6 * *...*$	
	$c^{63} : 9 * *...*$	$c'^{35} : 6 * *...*$	
	$c^{64} : 9 * *...*$	$c'^{36} : 6 * *...*$	
	$c^{65} : 9 * *...*$	$c'^{37} : 9 * *...*$	$g'_9 = 6$
	$c^{66} : 9 * *...*$	$c'^{38} : 9 * *...*$	
	$c^{67} : 9 * *...*$	$c'^{39} : 9 * *...*$	
	$c^{68} : 9 * *...*$	$c'^{40} : 9 * *...*$	
	$c^{69} : 9 * *...*$	$c'^{41} : 9 * *...*$	
	$c^{70} : 9 * *...*$	$c'^{42} : 9 * *...*$	
	$c^{71} : 9 * *...*$	$c'^{43} : A * *...*$	$g'_A = 6$
	$c^{72} : 9 * *...*$	$c'^{44} : A * *...*$	
	$c^{73} : 9 * *...*$	$c'^{45} : A * *...*$	
	$c^{74} : 9 * *...*$	$c'^{46} : A * *...*$	
	$c^{75} : 9 * *...*$	$c'^{47} : A * *...*$	
	$c^{76} : 9 * *...*$	$c'^{48} : A * *...*$	
g_A=10	$c^{77} : A * *...*$	$c'^{49} : B * *...*$	$g'_B = 6$
	$c^{78} : A * *...*$	$c'^{50} : B * *...*$	
	$c^{79} : A * *...*$	$c'^{51} : B * *...*$	
	$c^{80} : A * *...*$	$c'^{52} : B * *...*$	
	$c^{81} : A * *...*$	$c'^{53} : B * *...*$	
	$c^{82} : A * *...*$	$c'^{54} : B * *...*$	
	$c^{83} : A * *...*$	Discarded	
	$\vdots$	$\vdots$	
	$c^{86} : A * *...*$	Discarded	
g_B=5	$c^{87} : B * *...*$	Discarded	Removed
	$\vdots$	$\vdots$	
	$c^{91} : B * *...*$	Discarded	

Table 5.1: Dividing into groups in Example 16

We merge some of the remaining G_a, where $a \in Q \backslash Q_2$, into $g-v$ groups of size between p and $2p-2$, where the first coordinates of each group are different from the other groups. This can be done as follows.

Observe that

$$\begin{aligned}
\sum_{a \in Q \backslash Q_2} g_a &= |C| - \sum_{a \in Q_2} g_a \\
&= |C| - \sum_{a \in Q_2} (\alpha_a p + \beta_a) \\
&= |C| - \left(\sum_{a \in Q_2} \alpha_a p + \sum_{a \in Q_2} \beta_a \right) \\
&\geq |C| - (vp + v(p-1)) \\
&\geq 2gp - vp - v(p-1) \\
&= 2(g-v)p + v.
\end{aligned}$$

Hence, apart from $\cup_{a \in Q_2} G_a$, we have at least $2(g-v)p + v > 2(g-v)p$ codewords left in the code. And, since $g_a = \alpha_a + \beta_a = 0 + \beta_a = \beta_a \leq p-1$ for all $a \in Q \backslash Q_2$, we can automatically group $G_a, a \in Q \backslash Q_2$ in a greedy fashion into $g-v$ sets of size between p and $2p-2$, so that each G_a is not split into two or more sets. Let these sets be $C_{v+1}, C_{v+2} ..., C_g$. Note that the first coordinate of codewords in each new group may vary, but differs from any other group.

Step 3: Replacing

Here we construct the groups of C' as follows: Let $Q_2 = \{a_1, a_2, ..., a_v\}$.

1. for $i = 1$ to v, let C'_i be a set of codewords obtained from C_i by replacing the first coordinate by the symbol $a_i \in Q_2$,
2. for $i = v+1$ to g, let C'_i be a set of any p codewords from C_i,
3. let $C' = \bigcup_{i=1}^{g} C'_i$.

Now, we need to show that our constructed code C' satisfies Theorem 5.2.1.

Let the mapping $\varphi : C' \longrightarrow C$ map each codeword of C' to the codeword it was modified from in C. It is not difficult to see that φ is an injection that makes changes in only the first coordinate of any codeword.

The second condition is also satisfied since each group C'_i is of size $p = \left\lceil \frac{|C|}{2g} \right\rceil$.

The last condition follows from the first part of step 3, as well as the construction of $C_{v+1}, ..., C_g$.

Note that to construct C', we have eliminated $\left(\sum_{a \in Q} \beta_a \right) - (g-v)p$ codewords from C. Here the first term represents all the remainders, and the second term is derived from $(g - v)$ merged groups. Since $\left(\sum_{a \in Q} \beta_a \right) - (g - v)p = (|C| - vp) - (g - v)p = |C| - gp \leq |C| - \frac{|C|}{2} = \frac{|C|}{2}$, we are guaranteed to eliminate at most $\frac{|C|}{2}$ codewords from $|C|$. □

We will show in the next subsection that if C is any one-level FP, SFP and IPP code, the two-level code C' satisfying Theorem 5.2.1 has the corresponding two-level fingerprinting property. To make it more convenient for us to show this, we define some mappings and prove a lemma to be used in the next subsection.

Let the mapping $\pi : Q \longrightarrow Q$ be defined as follows. Let $\pi(a) = a$ when a does not appear as the first coordinate of any codeword of C', otherwise let $\pi(a) = b \in Q$ when there exists a codeword $c' \in C'$ with $c'_1 = a$ that was derived from $c \in C$ with $c_1 = b$.

Let $\psi : Q^\ell \longrightarrow Q^\ell$ be defined by mapping $x \in Q^\ell$ to $\psi(x) \in Q^\ell$ where

$$\psi(x)_i = \begin{cases} \pi(x_i) & \text{if } i = 1; \\ x_i & \text{otherwise.} \end{cases}$$

It is not difficult to see that ψ is a well-defined function and φ from Theorem 5.2.1 is actually ψ when restricted to C', i.e. $\varphi = \psi|_{C'}$.

Observe that for any $i \in [\ell]$ and any codewords $y, z \in C'$, if $y_i = z_i$, then $\varphi(y)_i = \varphi(z)_i$. Moreover, $\varphi(y)_i = y_i = z_i = \varphi(z)_i$ when $i \neq 1$.

Lemma 5.2.2. *Let C and C' be codes length ℓ over Q satisfying Theorem 5.2.1. Let X be a subset of C'. Then*

$$\psi(\text{desc}(X)) \subseteq \text{desc}(\psi(X)).$$

Proof. Let X be a subset of C' and let y be a codeword in $\psi(\text{desc}(X))$. Then, there exists a codeword x in $\text{desc}(X)$ such that $\psi(x) = y$. For any coordinate i in $[\ell]$, there exists a codeword x^i in X, where $x_i = x_i^i$. Hence $\psi(x)_i = \psi(x^i)_i$ for all $i \in [\ell]$. Which implies $\psi(x) \in \text{desc}(\{\psi(x^1), \psi(x^2), ..., \psi(x^\ell)\}) \subseteq \text{desc}(\psi(X))$. Therefore $y \in \text{desc}(\psi(X))$, which implies $\psi(\text{desc}(X)) \subseteq \text{desc}(\psi(X))$. □

5.2.1 The Existence of Codes

In this last part of the chapter, we demonstrate that the codes C' satisfying Theorem 5.2.1 are two-level FP, SFP or IPP codes if the original codes C are FP, SFP or IPP codes, respectively. Also, we provide an example showing that the two-level code constructed from a TA code using Theorem 5.2.1 does not always possess two-level TA property.

Theorem 5.2.3. *Let k, q, g and ℓ be integers greater than 1, where $g \leq q$. Let K be a positive integer such that $K \geq k$. Suppose that there exists a q-ary length ℓ one-level k-FP code C. Then there exists a q-ary length ℓ two-level (K, k)-FP code C' of cardinality at least $\frac{|C|}{2}$, where C' contains g groups of the same size.*

Proof. Let C' be a code obtained from the k-FP code C as in Theorem 5.2.1. It is easy to see that no coalition X of the codewords in C' can frame a codeword in a group disjoint from X, since any pair of codewords from different groups have different symbols in the first coordinate. So, only k-FP property of C' needs to be proved.

Let U be any subset of C' containing at most k codewords. Let $x \in \text{desc}(U) \cap C'$. We will show that $x \in U$. We have $\varphi(U) \subseteq C$ and $|\varphi(U)| \leq |U| \leq k$.

Since $x \in \text{desc}(U) \cap C'$, then $x \in \text{desc}(U)$ and $x \in C'$. By Lemma 5.2.2, $\varphi(x) \in$

$\text{desc}(\varphi(X))$. Also, it is easy to see that $\varphi(C') \subseteq C$. Hence, $\varphi(x) \in \text{desc}(\varphi(X)) \cap C$. So $\varphi(x) \in \varphi(U)$ by the k-FP property of C. Hence $x \in U$, which implies C' has the k-FP property, i.e. C' is a (K, k)-FP code.

Thus we can conclude that there exists a q-ary length ℓ two-level (K, k)-FP code C' of size at least $\frac{|C|}{2}$, containing g groups (each of size at least $\left\lceil \frac{|C|}{2g} \right\rceil$). □

Theorem 5.2.4. *Let k, q, g and ℓ be integers greater than 1, where $g \leq q$. Let K be a positive integer such that $K \geq k$. Suppose that there exists a q-ary length ℓ one-level k-SFP code C. Then there exists a q-ary length ℓ two-level (K, k)-SFP code C' of cardinality at least $\frac{|C|}{2}$, where C' contains g groups of the same size.*

Proof. Let C' be a code obtained from the k-SFP code C as in Theorem 5.2.1.

1. Let X_1, X_2 be subsets of C of size at most k, where $\text{desc}(X_1) \cap \text{desc}(X_2) \neq \emptyset$. We will show that $X_1 \cap X_2 \neq \emptyset$.

   Let $x \in \text{desc}(X_0) \cap \text{desc}(X_1)$. Then $\varphi(x) = \psi(x) \in \psi(\text{desc}(X_0) \cap \text{desc}(X_1))$. Now

   $$\begin{aligned} \varphi(x) &\in \psi(\text{desc}(X_1) \cap \text{desc}(X_2)) \\ &\subseteq \psi(\text{desc}(X_1)) \cap \psi(\text{desc}(X_2)) \\ &\subseteq \text{desc}(\psi(X_1)) \cap \text{desc}(\psi(X_2)) \text{ by Lemma 5.2.2} \\ &= \text{desc}(\varphi(X_1)) \cap \text{desc}(\varphi(X_2)). \end{aligned}$$

   Therefore $\text{desc}(\varphi(X_1)) \cap \text{desc}(\varphi(X_2)) \neq \emptyset$. By the k-SFP property of C, we deduce that $\varphi(X_1) \cap \varphi(X_2) \neq \emptyset$. Since φ is an injection, $\varphi(X_1) \cap \varphi(X_2) = \varphi(X_1 \cap X_2)$. Therefore $X_1 \cap X_2 \neq \emptyset$, which implies C' has the k-SFP property.

2. Let Y_1, Y_2 be subsets of C of size at most K, where $\text{desc}(Y_1) \cap \text{desc}(Y_2) \neq \emptyset$. We will show that $\mathcal{G}(Y_0) \cap \mathcal{G}(Y_1) \neq \emptyset$.

   Let $x \in \text{desc}(Y_1) \cap \text{desc}(Y_2)$. Then there exist codewords a in Y_1 and b in Y_2, where $a_1 = x_1 = b_1$. Since the first coordinate of each group is different from the

others, we can conclude that $\mathcal{G}(a) = \mathcal{G}(b)$. Therefore $\mathcal{G}(a) \in \mathcal{G}(Y_1) \cap \mathcal{G}(Y_2) \neq \emptyset$, so C' is a (K, k)-SFP code.

Therefore, there exists a q-ary length ℓ two-level (K, k)-SFP code C' of size at least $\frac{|C|}{2}$, containing g groups (each of size at least $\left\lceil \frac{|C|}{2g} \right\rceil$). □

Theorem 5.2.5. *Let k, q, g and ℓ be integers greater than 1, where $g \leq q$. Let K be a positive integer such that $K \geq k$. Suppose that there exists a q-ary length ℓ one-level k-IPP code C. Then there exists a q-ary length ℓ two-level (K, k)-IPP code C' of cardinality at least $\frac{|C|}{2}$, where C' contains g groups of the same size.*

Proof. Let C' be a code obtained from the k-IPP code C as in Theorem 5.2.1.

1. Let $x \in \mathrm{desc}_k(C')$. Then, there exists $U \subseteq C'$ such that $|U| \leq k$ and $x \in \mathrm{desc}(U)$. By Lemma 5.2.2, we know that $\psi(x) \in \mathrm{desc}(\varphi(U))$ and $\varphi(U) \subseteq C$. Observe that $|\varphi(U)| \leq |U| \leq k$. Hence $\psi(x) \in \mathrm{desc}_k(C)$. Since C is an IPP code,

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ \psi(x) \in \mathrm{desc}(X)}} X \neq \emptyset.$$

Also, for any $X \subseteq C'$, $x \in \mathrm{desc}(X)$ implies $\psi(x) \in \mathrm{desc}(\varphi(X))$ and $|X| = |\varphi(X)|$. Hence

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ \psi(x) \in \mathrm{desc}(X)}} X \subseteq \bigcap_{\substack{X \subseteq C' : |X| \leq k \\ x \in \mathrm{desc}(X)}} \varphi(X).$$

Since φ is injective, we have

$$\bigcap_{\substack{X \subseteq C' : |X| \leq k \\ x \in \mathrm{desc}(X)}} \varphi(X) = \varphi\Big(\bigcap_{\substack{X \subseteq C' : |X| \leq k \\ x \in \mathrm{desc}(X)}} X \Big).$$

Hence

$$\varphi(\bigcap_{\substack{X\subseteq C':|X|\leq k\\ x\in\mathrm{desc}(X)}} X)\neq\emptyset.$$

Therefore

$$\bigcap_{\substack{X\subseteq C':|X|\leq k\\ x\in\mathrm{desc}(X)}} X\neq\emptyset,$$

which shows that C' is a k-IPP code.

2. Let $y\in\mathrm{desc}_K(C')$. Then, there exists $V\subseteq C'$ such that $|V|\leq K$ and $y\in\mathrm{desc}(V)$. Let v be a codeword in V such that $v_1=y_1$, and let $i\in[g]$ such that $v\in C'_i$. Hence $\mathcal{G}(v)=i$. For any $X\subseteq C'$ of cardinality at most K with $\mathrm{desc}(X)$ containing y, there exists a codeword y^X such that $y_1=y_1^X$. Since the group index of a codeword can be determined from its first coordinate, we have $\mathcal{G}(y^X)=i$. That implies

$$i\in\bigcap_{\substack{X\subseteq C':|X|\leq K\\ y\in\mathrm{desc}(X)}}\mathcal{G}(X).$$

Consequently,

$$\bigcap_{\substack{X\subseteq C':|X|\leq K\\ y\in\mathrm{desc}(X)}}\mathcal{G}(X)\neq\emptyset.$$

Therefore, C' has the $(K,*)$-IPP property and is a (K,k)-IPP code.

Thus, there exists a q-ary length ℓ two-level (K,k)-IPP code C' of size at least $\frac{|C|}{2}$, containing g groups (each of size at least $\left\lceil\frac{|C|}{2g}\right\rceil$). □

The two-level codes satisfying Theorem 5.2.1 preserve the fingerprinting property

from their corresponding one-level codes for IPP, SFP and FP codes. However, this is not always true in the case of TA codes as can be seen in the following example.

Example 17. Let $C = \{122, 133, 144, 155, 216, 317, 418, 519, 661, 771, 881, 991\} \subseteq \{1, 2, 3, ..., 9\}^3$. It is not difficult to check that C is a 2-TA code. Let $g = 4$, then $p = \lceil \frac{12}{8} \rceil = 2$. Then Theorem 5.2.1 does not guarantee two-level traceability code from C.

Proof. Here we have $g_1 = 4$, $g_2 = g_3 = ... = g_9 = 1$. Consider $C_1 = \{122, 133\}$, $C_2 = \{144, 155\}$, $C_3 = \{216, 661\}$ and $C_4 = \{317, 771\}$, which leads to $C'_1 = \{122, 133\}$, $C'_2 = \{944, 955\}$, $C'_3 = \{216, 661\}$ and $C'_4 = \{317, 771\}$ by Theorem 5.2.1. Let $U = \{122, 216, 661\}$, then $111 \in \text{desc}(U)$ and $\mathcal{G}(U) = \{1, 3\}$. Observe that 317 is a codeword of C' with $d_H(111, 317)$ minimal, but $\mathcal{G}(317) = 4 \notin \mathcal{G}(U)$. Therefore C' is not a $(K, 2)$-TA code for any integer K greater than 2. □

Theorem 5.2.1 ensures that we can always construct two-level IPP, SFP and FP codes, with $g \leq q$, of size at least half of the size of the existing one-level codes. When the one-level code is of exponential size, throwing away half of its codewords would not effect the codes' size significantly. However, we do not have the same result for TA codes.

The example above shows the construction does not work, not that the analogue of Theorems 5.2.3-5.2.5 is false for TA codes.

Chapter 6

Constructing Two-Level Frameproof Codes

Since the only known general method for constructing frameproof codes involves using error-correcting codes with high minimum distance (see Chapter 3), one might naturally try to construct two-level codes based on high minimum distance codes. In this chapter, we state a sufficient condition on the minimum distance of a code that makes it two-level frameproof. Then we propose a different method for constructing frameproof codes which produces code of significantly larger size than that based on high minimum distance codes.

To clearly illustrate our improvement, we first state the sufficient conditions for a high minimum distance code to be a (K, k)-FP code. Recall the definition of $d_1(C)$ and $d_2(C)$ from Section 4.1.

Theorem 6.0.6. *Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level length ℓ code containing g groups of p codewords. If*

$$d_1(C) > (1 - 1/K)\ell \text{ and}$$
$$d_2(C) > (1 - 1/k)\ell,$$

then C is a (K, k)-FP code.

Proof. Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level length ℓ code containing g groups of p codewords such that $d_1(C) > (1 - 1/K)\ell$ and $d_2(C) > (1 - 1/k)\ell$. Theorem 3.1.3 implies that C is a k-FP code. Now we only need to show the validity of the $(K, *)$-FP property.

Let X be a subset of C, where $|X| \leq K$. Let y be any codeword in $C \backslash X$ such that $\mathcal{G}(y) \notin \mathcal{G}(X)$. We need to show that $y \notin \text{desc}(X) \cap C$. Since $\mathcal{G}(y) \notin \mathcal{G}(X)$, each codeword in X can agree with y in at most $\ell - d_1(C)$ coordinates. By using all the codewords from X, we can construct a descendant of X that agrees with y in up to $|X|\,(\ell - d_1(C))$ coordinates. But $|X|\,(\ell - d_1(C)) < |X|(\ell - (1- 1/K)\ell) \leq K(1/K)\ell = \ell$. Thus $d_H(y, \text{desc}(X)) \geq \ell - |X|\,(\ell - d_1(C)) > \ell - \ell = 0$, and so $y \notin \text{desc}(X)$. Therefore $y \notin \text{desc}(X) \cap C$. Hence, C is (K, k)-FP. □

6.1 Constructing Two-Level FP Codes

Although Theorem 6.0.6 suggests sufficient conditions on $d_1(C)$ and $d_2(C)$ for C to be a (K, k)-FP code, it is not obvious how codes satisfying those conditions can be constructed. We could construct (K, k)-FP codes using Theorem 3.1.3. However, the size of the two-level codes obtained from this construction is rather small. In this section, we propose a construction for two-level FP codes which is not based on high minimum distance codes. The construction we describe give bigger codes than those of high minimum distance.

The general idea of our two-level FP codes construction is to obtain a (K, k)-FP code by combining a K-FP and a k-FP code with certain properties together in a particular way.

Let Q_1 and Q_2 be finite sets, and let ℓ be a positive integer. For any x in Q_1^ℓ and y in Q_2^ℓ, let $c_{xy} = ((x_1, y_1), (x_2, y_2), ..., (x_\ell, y_\ell))$.

For $i \in \{1, 2\}$, let P_i be a projection from $(Q_1 \times Q_2)^\ell$ to Q_i^ℓ, defined by

$$P_1(c_{xy}) = x,$$
$$P_2(c_{xy}) = y,$$

for any $c_{xy} \in (Q_1 \times Q_2)^\ell$.

Construction 6.1.1. *Let ℓ, K and k be positive integers such that $\ell \geq 2$ and $K \geq k \geq 2$. Let q_1 and q_2 be prime powers greater than ℓ. Let $\mathbb{F}_{q_1}$ and $\mathbb{F}_{q_2}$ be finite fields of cardinality q_1 and q_2 respectively. Let D_1 be a K-FP code of length ℓ over $\mathbb{F}_{q_1}$ and let D_2 be a k-FP code of length ℓ over $\mathbb{F}_{q_2}$, constructed as in Construction 3.1.2.*

Let Q denote $\mathbb{F}_{q_1} \times \mathbb{F}_{q_2}$. Define a length ℓ code C over Q by

$$C = \bigcup_{x \in D_1} C_x,$$

where $C_x = \{c_{xy} : y \in D_2\}$ for each $x \in D_1$. Then, C is a (K, k)-FP code containing $|D_1|$ groups of $|D_2|$ codewords each.

Proof. We know that D_1 is K-FP and D_2 is k-FP from Construction 3.1.2, and it is clear that $P_1(C) = D_1$ and $P_2(C) = D_2$. Let $\mathcal{G} = P_1$.

(i) Let U be a subset of C of size at most K. Let $x \in \text{desc}(U) \cap C$. Then, $\mathcal{G}(x) = P_1(x) \in P_1(\text{desc}(U) \cap C)$. Since $P_1(\text{desc}(U) \cap C) = \text{desc}(P_1(U)) \cap D_1$, it implies $\mathcal{G}(x) \in \text{desc}(P_1(U)) \cap D_1$. And since $P_1(U)$ is a subset of D_1 of cardinality less than K, and D_1 is K-FP, then, $\mathcal{G}(x) \in P_1(U)$. Therefore, $\mathcal{G}(x) \in \mathcal{G}(U)$.

(ii) Let V be subset of C of size at most k. Let $c_{xy} \in \text{desc}(V) \cap C$. So, c_{xy} must agree with some codeword $c_{x'y'} \in V$ in at least $\lceil \ell/k \rceil$ positions. Hence, x agrees with x' in at least $\lceil \ell/k \rceil \geq \lceil \ell/K \rceil$ positions and y agrees with y' in at least $\lceil \ell/k \rceil$ positions. Therefore, by the minimum distance of D_1 and D_2, $x = x'$ and $y = y'$. That means $c_{xy} = c_{x'y'} \in V$.

Therefore, C is a (K, k)-FP code containing $|D_1|$ groups of $|D_2|$ codewords each. □

6.2 More on the Constructions

A natural question regarding the previous construction is whether $C = \{c_{xy} : x \in D_1$ and $y \in D_2\}$ is a (K, k)-FP code for any choice of K-FP code D_1 and k-FP code D_2. If not, what are the necessary and sufficient conditions for D_1 and D_2 that make C a (K, k)-FP code? This problem is considered in this section.

Actually, given that D_1 is a K-FP code and D_2 is a k-FP code, $C = \{c_{xy} : x \in D_1$ and $y \in D_2\}$ is not always a (K, k)-FP code. Here is an example.

Example 18. Let D_1 be a 4-FP code constructed from Construction 3.1.1 and let D_2 be a 2-FP code from Example 1, i.e., $D_1 = \{1000, 0100, 0010, 0001, 2000, 0200, 0020, 0002\}$ and $D_2 = \{1100, 1001, 1010\}$. A two-level code $C\{c_{xy} : x \in D_1$ and $y \in D_2\}$ is not a $(4, 2)$-FP code.

Proof. Consider $X = \{((1, 1), (0, 0), (0, 1), (0, 0)), ((0, 1), (2, 0), (0, 0), (0, 1))\}$. We have $((1, 1), (0, 0), (0, 0), (0, 1)) \in \text{desc}(X) \cap C$, which contradicts the $(4, 2)$-FP property. □

Before discussing the main problem, we define some additional notation.

For any codewords x, y in Q^ℓ and any subset X of Q^ℓ, let

$$\mathcal{I}(x, y) := \{i \in [\ell] \ : x_i = y_i\}$$

be the set of all components of x that agree with y, and let

$$\mathcal{I}(x, X) := \bigcup_{y \in X \setminus \{x\}} \mathcal{I}(x, y)$$

be the set of all components of x that agree with at least one member of X.

For any code $C \subseteq Q^\ell$ and any positive integer m, define

$$\mathcal{I}_m(C) := \{\mathcal{I}(x, X) : X \subseteq C, |X| \leq m \text{ and } x \in C \backslash X\}.$$

The following theorem provides necessary and sufficient conditions on D_1 and D_2 so that they can be used to construct a two-level FP code in a similar fashion to Construction 6.1.1.

Theorem 6.2.1. *Let ℓ, K and k be positive integers such that $\ell \geq 2$ and $K \geq k \geq 2$. Let q_1 and q_2 be prime powers greater than ℓ. Let D_1 be a K-FP code of length ℓ over $\mathbb{F}_{q_1}$, and let D_2 be a k-FP code of length ℓ over $\mathbb{F}_{q_2}$. Then $C = \{c_{xy} : x \in D_1 \text{ and } y \in D_2\}$ is a (K, k)-FP code off and only if $I \cup J \neq [\ell]$ for any $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$, where s and t are positive integers such that $s + t \leq k$.*

Proof. Suppose C has the (K, k)-FP property. Let s and t be positive integers such that $s + t \leq k$. We assume for a contradiction that there exist $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$ such that $I \cup J = [\ell]$. Let $X_1 \subseteq D_1$ and $x' \in D_1 \backslash X_1$ be such that $|X_1| \leq s$ and $\mathcal{I}(x', X_1) = I$. And let $X_2 \subseteq D_2$ and $x'' \in D_2 \backslash X_2$ be such that $|X_2| \leq t$ $\mathcal{I}(x'', X_2) = J$. Let U be $\{c_{xx''} : x \in X_1\} \cup \{c_{x'y} : y \in X_2\}$. Then $|U| \leq s + t \leq k$.

For any $i \in I$, there exists a codeword $x \in X_1$ such that $x_i = x'_i$. Hence $c_{xx''} \in U$ and $c_{x'x''i} = c_{xx''i}$. Similarly, for any $i \in J$, there exists a codeword $y \in X_2$ such that $y_i = x''_i$. Hence $c_{x'y} \in U$ and $c_{x'x''i} = c_{x'y_i}$. Therefore $c_{x'x''}$ is in $\text{desc}(U) \cap C$. However, $c_{x'x''}$ is not in U, since $x' \notin X_1$ and $x'' \notin X_2$, contradicting the (K, k)-FP property of C. Hence $I \cup J \neq [\ell]$ for any $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$.

Conversely, with similar arguments to the first part of the proof of Construction 6.1.1, we can show that C has the $(K, *)$-FP property for any choice of K-FP code D_1 and k-FP code D_2. Hence, only k-FP property is yet to be established.

Suppose that $I \cup J \neq [\ell]$ for any $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$, where s and t are positive integers such that $s + t \leq k$.

Let U be a subset of C of size at most k. Let $c_{xy} \in \text{desc}(U) \cap C$. Replace any

$c_{x'y'} \in U$ such that both $x' \neq x$ and $y' \neq y$ by either $c_{xy'}$ or $c_{x'y}$; name the new set U'. It is clear that $c_{xy} \in \text{desc}(U') \cap C$. Moreover, U' can be written as $X_1 \cup X_2$ where

$$X_1 := \{c_{ab} \in U' : a = x\},$$
$$X_2 := \{c_{ab} \in U' : b = y\}.$$

Note that $X_1 \neq \emptyset$ since D_1 is K-FP and $x = P_1(c_{xy}) \in P_1(U') \cap D_1$. Also $X_2 \neq \emptyset$ since D_2 is k-FP and $y = P_2(c_{xy}) \in P_2(U') \cap D_2$. Since $c_{xy} \in \text{desc}(U') \cap C$, we have $\mathcal{I}(c_{xy}, X_1) \cup \mathcal{I}(c_{xy}, X_2) = [\ell]$. Observe that,

$$\mathcal{I}(c_{xy}, X_1) \cup \mathcal{I}(c_{xy}, X_2) = \mathcal{I}(y, P_2(X_1)) \cup \mathcal{I}(x, P_1(X_2)),$$

so $\mathcal{I}(y, P_2(X_1)) \cup \mathcal{I}(x, P_1(X_2)) = [\ell]$. By assumption, when $s + t \leq k$ there exist no $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$ such that $I \cup J = [\ell]$. Therefore, $|P_2(X_1)| + |P_1(X_2)|$ must be bigger than k. Hence, $X_1 \cap X_2 \neq \emptyset$. This implies $c_{xy} \in U'$. But, by the construction of U', this could happen only if $c_{xy} \in U$. Therefore, $c_{xy} \in U$. Hence, C is k-FP.

□

Although the theorem give both necessary and sufficient conditions on D_1 and D_2, it is not an easy task to check if those properties hold. The the sufficient conditions in the next corollary are easier to verify, but the size of the resulting code will also be reduced.

Corollary 6.2.2. *Let K and k be integers such that $2 \leq k < K$. Let D_1 and D_2 be length ℓ error correcting codes of minimum distance $d_1 > (1-1/K)\ell$ and $d_2 > (1-1/k)\ell$, respectively. Then, $C = \{c_{xy} : x \in D_1 \text{ and } y \in D_2\}$ is a (K, k)-FP code containing $|D_1|$ groups of $|D_2|$ codewords each.*

Proof. By Theorem 3.1.3, D_1 and D_2 are K-FP and k-FP codes, respectively. Hence, C has the $(K, *)$-FP property. For any positive integers s and t, where $s + t \leq k$, and

any $I \in \mathcal{I}_s(D_1)$ and $J \in \mathcal{I}_t(D_2)$, we have

$$\begin{aligned} |I| + |J| &\leq s(\ell/K) + t(\ell/k) \\ &< s(\ell/k) + t(\ell/k) \\ &= (s+t)(\ell/k) \\ &\leq \ell. \end{aligned}$$

Therefore, $|I| + |J| < \ell$. Hence, $I \cup J \neq [\ell]$. Applying Theorem 6.2.1, we find that C is a (K, k)-FP code as required. □

An alternative proof, which is much shorter, for this corollary is as follows.

Proof. Consider $d_1(C) \geq d_H(D_1) > (1 - 1/K)\ell$ and $d_2(C) \geq d_H(D_2) > (1 - 1/k)\ell$. By Theorem 6.0.6, C is a (K, k)-FP code. □

To visualise the idea, we give an example of a two-level code that can be constructed using Theorem 6.2.1.

Example 19. Consider a 4-FP code D_1 of length 6 inspired by a construction for a 3-FP code of length 5 in [10] (see below) and a 2-FP code D_2 of length 6 constructed as in Construction 3 of [10] (again, see below). Neither D_1 or D_2 is of high minimum distance. We show that $C = \{c_{xy} : x \in D_1 \text{ and } y \in D_2\}$ is a (4,2)-FP code of length 6, containing $\frac{6}{4}(q_1^2 - 2q_1 + 1)$ groups of $2(q_2 - 1)^3(1 - \frac{1}{2\sqrt{q_2 - 1}})$ codewords each.

The construction of D_1

We define six sets S_1, S_2, S_3, S_4, S_5, and S_6 of words of length 6 over the alphabet

$\mathbb{Z}_4 \cup \{\infty\}$ as follows:

$$S_1 = \{(\infty, a, a, a, a, a) : a \in \mathbb{Z}_4\}$$
$$S_2 = \{(a, \infty, a+2, a+3, a, a+1) : a \in \mathbb{Z}_4\}$$
$$S_3 = \{(a, a, \infty, a+1, a+3, a+2) : a \in \mathbb{Z}_4\}$$
$$S_4 = \{(a, a+1, a+3, \infty, a+2, a) : a \in \mathbb{Z}_4\}$$
$$S_5 = \{(a, a+2, a+1, a, \infty, a+3) : a \in \mathbb{Z}_4\}$$
$$S_6 = \{(a, a+3, a, a+2, a+1, \infty) : a \in \mathbb{Z}_4\}$$

The sets S_i are pairwise disjoint and contain 4 words. Moreover, each codeword in $S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6$ is uniquely determined by any two components.

Let m_1 be a prime power such that $m_1 \geq 5$. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and α_5 be distinct elements of $\mathbb{F}_{m_1}$. We define six sets T_1, T_2, T_3, T_4, T_5, and T_6 of words of length 6 over the alphabet $\mathbb{F}_{m_1} \cup \{\infty\}$ as follows:

$$T_1 = \{(\infty, f(\alpha_1), f(\alpha_2), f(\alpha_3), f(\alpha_4), f(\alpha_5)) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$
$$T_2 = \{(f(\alpha_1), \infty, f(\alpha_2), f(\alpha_3), f(\alpha_4), f(\alpha_5)) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$
$$T_3 = \{(f(\alpha_1), f(\alpha_2), \infty, f(\alpha_3), f(\alpha_4), f(\alpha_5)) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$
$$T_4 = \{(f(\alpha_1), f(\alpha_2), f(\alpha_3), \infty, f(\alpha_4), f(\alpha_5)) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$
$$T_5 = \{(f(\alpha_1), f(\alpha_2), f(\alpha_3), f(\alpha_4), \infty, f(\alpha_5)) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$
$$T_6 = \{(f(\alpha_1), f(\alpha_2), f(\alpha_3), f(\alpha_4), f(\alpha_5), \infty) : f \in \mathbb{F}_{m_1}[X], \deg f \leq 1\}$$

The sets T_i are pairwise disjoint and have cardinality m_1^2. Moreover, two distinct codewords x, y in T_i can agree at no more than one component other than ith, since $\deg f \leq 1$.

Define sets A_1, A_2, A_3, A_4, A_5, and A_6 of words of length 6 over the alphabet $Q_1 = (\mathbb{Z}_4 \times \mathbb{F}_{m_1}) \cup \{(\infty, \infty)\}$ by

$$A_i = \{c_{xy} : x \in S_i \text{ and } y \in T_i\}$$

for all $i \in [6]$. Then $|A_i| = |S_i| \times |T_i| = 4m_1^2$.

Let $D_1 = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6$. Then D_1 is a 4-FP code of size $\frac{6}{4}(q_1^2 - 2q_1 + 1)$ over the alphabet set Q_1, where $q_1 = |Q_1| = 4m_1 + 1$.

The construction of D_2

Let m_2 be a prime power such that $m_2 \geq 7$ and fix $q_2 = m_2^2 + 1$. Let $\mathbb{F}_{m_2}$ be the finite field of order m_2, and define Q_2 to be $(\mathbb{F}_{m_2})^2 \cup \{\infty\}$. Let $\beta_1, \beta_2, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ be distinct elements of $\mathbb{F}_{m_2}$. Define B_1 and B_2 by,

$$\begin{aligned} B_1 =& \{(\infty, (f(\alpha_1), g(\alpha_1)), (f(\alpha_2), g(\alpha_2)), (f(\alpha_3), g(\alpha_3)), (f(\alpha_4), g(\alpha_4)), \\ & (f(\alpha_5), g(\alpha_5))) : f, g \in \mathbb{F}_{m_2}[X], \deg f = 2, \deg g \leq 2\} \\ B_2 =& \{((t(\beta_1), t(\beta_2)), (s(\alpha_1), t(\alpha_1)), (s(\alpha_2), t(\alpha_2)), (s(\alpha_3), t(\alpha_3)), (s(\alpha_4), t(\alpha_4)), \\ & (s(\alpha_5), t(\alpha_5))) : s, t \in \mathbb{F}_{m_2}[X], \deg s \leq 1, \deg t \leq 3\} \end{aligned}$$

Define $D_2 = B_1 \cup B_2$. Then D_2 is a 2-FP code of cardinality $2(q_2 - 1)^3(1 - \frac{1}{2\sqrt{q_2 - 1}})$ over the alphabet set Q_2.

The construction of C

Let $C = \{c_{xy} : x \in D_1 \text{ and } y \in D_2\}$ and define $\mathcal{G} = P_1$. We show that C is a (4,2)-FP code of length 6, using Theorem 6.2.1.

The only possible pair of positive integers s and t such that $s + t \leq k = 2$ is $s = 1, t = 1$. Hence, we need to consider only $\mathcal{I}_1(D_1)$ and $\mathcal{I}_1(D_2)$, where

$$\mathcal{I}_1(D_1) = \{I \subseteq [\ell] : |I| \leq 2\} \text{ and}$$
$$\mathcal{I}_1(D_2) = \{I \subseteq [\ell] : |I| \leq 3\}.$$

We can easily see that for any $I \in \mathcal{I}_1(D_1), J \in \mathcal{I}_1(D_2)$, $|I \cup J| \leq 5$. Hence, $I \cup J \neq [\ell]$. Therefore, C is a $(4, 2)$-FP code of length 6, containing $\frac{6}{4}(q_1^2 - 2q_1 + 1)$ groups of $2(q_2 - 1)^3(1 - \frac{1}{2\sqrt{q_2 - 1}})$ codewords.

The size of a two-level code constructed from high minimum distance codes as in Corollary 6.2.2 is at most $q_1^2 q_2^3$, since the singleton bound shows that $|D_1| \leq q_1^2$ and $|D_2| \leq q_2^3$, respectively. The two-level FP code in Example 19 has a much larger number of groups and a significantly larger size than any code constructed from high minimum distance codes.

We remark that our construction is for FP codes only: even if D_1 and D_2 are SFP codes that satisfied the condition in Theorem 6.2.1, C is not always a SFP code. We give an example here.

Example 20. Let D_1 be a 4-SFP code from Example 24 and let D_2 be a 2-SFP code from Example 3, i.e., $D_1 = \{0000, 0111, 1122, 2220\}$ and $D_2 = \{1001, 1200, 0010, 2211\}$. A two-level code $C\,\{c_{xy} : x \in D_1 \text{ and } y \in D_2\}$ is not a $(4, 2)$-SFP code.

Proof. The only possible pair of positive numbers s and t such that $s + t \leq k = 2$ is $s = 1, t = 1$. Hence, we need to consider only $\mathcal{I}_1(D_1)$ and $\mathcal{I}_1(D_2)$, where

$$\mathcal{I}_1(D_1) = \{I \subseteq [\ell] : |I| \leq 1\} \text{ and}$$
$$\mathcal{I}_1(D_2) = \{I \subseteq [\ell] : |I| \leq 2\}.$$

We can easily see that for any $I \in \mathcal{I}_1(D_1), J \in \mathcal{I}_1(D_2)$, $|I \cup J| \leq 3$. Hence, $I \cup J \neq [\ell]$.

Consider

$$X = \{((0,2),(0,2),(0,1),(0,1)),((0,1),(0,0),(0,0),(0,1))\}$$

and

$$Y = \{((0,1),(0,2),(0,0),(0,0)),((2,1),(2,0),(2,0),(0,1))\},$$

two distinct coalitions of size at most 4. We have $((0,1),(0,2),(0,0),(0,1)) \in \text{desc}(X) \cap \text{desc}(Y)$, which contradicts the $(4,2)$-SFP property. □

Chapter 7

Constructing Two-Level IPP Codes

Recall that a one-level IPP code can be constructed from an error-correcting code with high minimum distance (see Theorem 3.4.1). In this chapter, we give sufficient conditions for a high minimum distance code to be a two-level IPP code. Then we propose a construction for two-level IPP codes which is not based on high minimum distance error-correcting codes. Our construction gives codes with at least the same size as codes based on high minimum distance codes. Under a certain condition, our construction gives codes with larger size.

Now we state the sufficient conditions for a high minimum distance code to be a (K, k)-TA code given by Anthapadmanabhan and Barg [4]. Recall the definition of $d_1(C)$ and $d_2(C)$ from Section 4.1.

Theorem 7.0.3. *Let $C = C_1 \cup C_2 \cup ... \cup C_g$ be a two-level code of length ℓ containing g groups of p codewords. If*

$$d_1(C) > (1 - 1/K^2)\ell \text{ and}$$
$$d_2(C) > (1 - 1/k^2)\ell,$$

for some positive integers k and K, where $K > k \geq 2$, then C is a (K, k)-TA code.

Since it is not obvious how to construct a two-level IPP code using Theorem 7.0.3, we can use Theorem 3.4.1 to construct a K-IPP code which is also a (K, k)-IPP code. However, such a code can be small.

Example 13 shows that there exists a two-level TA (or IPP) code that is larger than any TA (or IPP) code constructed using Theorem 3.4.1. Consider a q-ary error correcting code C of length ℓ with minimum distance $d_H(C) = \left\lceil (1 - \frac{1}{(\ell-1)^2})\ell \right\rceil = \ell$. By Singleton bound, C has at most $q^{\ell - d_H(C)+1} = q^{\ell-\ell+1} = q$ codewords. While TA codes in Example 13 contain $(k+1)r = \frac{k+1}{k}(q-1) = q + \frac{q-k-1}{k} > q$ codewords, when $k < q + 1$.

7.1 Construction of Two-Level IPP Codes

In this section, we propose a new construction for two-level IPP codes. The general idea of our two-level IPP code construction is to obtain a (K, k)-IPP code by concatenating a K-IPP and a k-IPP code together in a particular way.

Construction 7.1.1. *Let Q be an alphabet of size q, and let g and p be positive integers such that $g \leq q$. Let D_1 be a code of length ℓ_1 over Q and D_2 be a code of length ℓ_2 over Q such that $|D_2| = gp$ and $|D_1| \geq g$.*

Construct a two-level code C containing g groups of size p as follows,

1. *Partition D_2 into g disjoint sets of the same size p, say $U_1, U_2, ..., U_g$.*
2. *Let $c_1, c_2, ..., c_g$ be g distinct codewords from D_1.*
3. *For any $i \in [g]$, let $C_i = \{x||c_i : x \in U_i\}$.*
4. *Let $C = C_1 \cup C_2 \cup ... \cup C_g$.*

For $i \in [2]$ and $\ell = \ell_1 + \ell_2$, let P_i be the projection from Q^ℓ to Q^{ℓ_i} defined by

$$P_1(x) = (x_{\ell_2+1}, x_{\ell_2+2}, ..., x_\ell), \text{ and}$$
$$P_2(x) = (x_1, x_2, ..., x_{\ell_2}), \tag{7.1}$$

for all $x = (x_1, x_2, ..., x_\ell) \in Q^\ell$.

It is easy to see that when C is defined as above, for any $c = x||c_i \in C$, where $x \in D_2$ and $i \in [g]$, we have

1. $P_2(c) = x$ and P_2 is bijective,
2. $P_1(c) = c_i$,
3. $\mathcal{G}(c) = i$.

Theorem 7.1.1. *Let C be a two-level code constructed as in Construction 7.1.1. Assume that D_1 is a K-IPP code and D_2 is a k-IPP code, then C is (K, k)-IPP.*

Proof. Assume that D_1 is K-IPP and D_2 is k-IPP.

(i) Let $x \in \text{desc}_k(C)$. Then there exists a subset $X_0 \subseteq C$ of size at most k such that $x \in \text{desc}(X_0)$. Hence, $P_2(x) \in P_2(\text{desc}(X_0))$. Note that P_2 is bijective, thus $P_2(\text{desc}(X_0)) = \text{desc}(P_2(X_0))$. Since $|P_2(X_0)| \leq |X_0| \leq k$, $P_2(X) \subseteq D_2$ and D_2 is k-IPP,

$$\bigcap_{\substack{X \subseteq D_2 : |X| \leq k \\ P_2(x) \in \text{desc}(X)}} X \neq \emptyset.$$

Recall that P_2 is bijective. Therefore

$$\bigcap_{\substack{X \subseteq C : |X| \leq k \\ x \in \text{desc}(X)}} X \neq \emptyset.$$

Which implies C is k-IPP.

(ii) Let $y \in \mathrm{desc}_K(C)$. Then there exists a subset $Y_0 \subseteq C$ of size at most K such that $y \in \mathrm{desc}(Y_0)$. Then $P_1(y) \in P_1(\mathrm{desc}(Y_0)) = \mathrm{desc}(P_1(Y_0))$.

Since D_1 is K-IPP and $|P_1(Y_0)| \leq |Y_0| \leq K$. Then

$$\bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ P_1(y) \in \mathrm{desc}(X)}} X \neq \emptyset.$$

Also

$$\begin{aligned}
\bigcap_{\substack{X \subseteq C : |X| \leq K \\ y \in \mathrm{desc}(X)}} \mathcal{G}(X) &= \bigcap_{\substack{X \subseteq C : |X| \leq K \\ y \in \mathrm{desc}(X)}} \{i \in [g] : c_i = P_1(x) \text{ for some } x \in X\} \\
&= \bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ P_1(y) \in \mathrm{desc}(X)}} \{i \in [g] : c_i \in X\} \\
&= \bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ P_1(y) \in \mathrm{desc}(X)}} \mathcal{G}(\{c_i \in X\}) \\
&= \mathcal{G}\left(\bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ P_1(y) \in \mathrm{desc}(X)}} \{c_i \in X\} \right) \\
&= \mathcal{G}\left(\bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ P_1(y) \in \mathrm{desc}(X)}} X \right) \\
&\neq \emptyset.
\end{aligned}$$

Therefore, C is (K, k)-IPP. □

It is quite obvious from our construction that D_2 has to be a k-IPP code. The next theorem will show that D_1 also has to be a K-IPP code.

Theorem 7.1.2. *Let C be a two-level code constructed as in Construction 7.1.1. Assume that D_2 is a k-IPP code. Then C has the (K, k)-IPP property if and only if D_1 is a K-IPP code.*

Proof. Let D_2 be a k-IPP code. If D_1 is a K-IPP code, then C is (K, k)-IPP, by Theorem 7.1.1.

Conversely, suppose that D_1 is not a K-IPP code. We need to show that C is not a (K, k)-IPP code. Since D_1 is not a K-IPP code, there exist $I \subseteq [g]$ of size at most K and $x_0 \in \text{desc}(\{c_i : i \in I\})$ such that

$$\bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ x_0 \in \text{desc}(X)}} X = \emptyset.$$

Let X_0 be a subset of C of size at most K which $\mathcal{G}(X_0) = I$. Since $x_0 \in \text{desc}(\{c_i : i \in I\}) = \text{desc}(P_1(X_0)) = P_1(\text{desc}(X_0))$, there exists $x \in \text{desc}(X_0)$ such that $P_1(x) = x_0$. Similar to the proof of Theorem 7.1.1, it follows that

$$\begin{aligned} \bigcap_{\substack{X \subseteq C : |X| \leq K \\ x \in \text{desc}(X)}} \mathcal{G}(X) &= \mathcal{G}\left(\bigcap_{\substack{X \subseteq D_1 : |X| \leq K \\ x_0 \in \text{desc}(X)}} X \right) \\ &= \emptyset. \end{aligned}$$

Hence C does not have the $(K, *)$-IPP property.

Therefore, C is (K, k)-IPP if and only if D_1 is a K-IPP code. □

Given two one-level IPP codes D_1 and D_2 that have size larger than high minimum distance codes, we can always construct a two-level IPP code of size larger than codes which could be obtained from Theorem 3.4.1. However our construction is valid only for IPP codes: even if D_1 and D_2 are TA codes, C is not always a TA code. We give an example here.

Example 21. Let D_1 be a 3-TA code represented by $\{000, 111, 222\}$, and let $D_2 =$

$D'_1 \cup D'_2 \cup D'_3$ where for all $i \in [r]$,

$$D'_1 = \{(0, i, i) : i \in [2]\}$$
$$D'_2 = \{(i, 0, r + i) : i \in [2]\}$$
$$D'_3 = \{(r + i, r + i, 0) : i \in [2]\}$$

By Example 13, D_2 is a $(6, 2)$-TA code and it is also a 2-TA code (as proposed by Blackburn, Etzion and Ng [13]). Then a code C constructed from D_1 and D_2 using Construction 7.1.1 is not a (3,2)-TA code.

Proof. Observe that one possible result from the construction is $C_i = D'_i||\{iii\}$ where $i \in [3]$. Let $X = \{103222, 330333\}$. Then $X \subseteq C$ and $|X| \leq 2$. Consider the word 103333 from desc(X). We have $d_H(103333, 103222) = d_H(103333, 330333) = d_H(103333, 440333)$=3, but $440333 \notin X$. Hence, C is not a 2-TA code, and so cannot be a $(3, 2)$-TA code. □

Chapter 8

Separating Hash Families

The notion of a separating hash family was first introduced by Stinson, van Trung, and Wei [33] in 1997 as a tool to create an explicit construction for frameproof codes. In this chapter, we first state the definition of separating hash families, and then discuss how this concept relates to fingerprinting codes. We present some previously known bounds on the size of separating hash families in the last section. All the material in this chapter is well-known.

8.1 Introduction to Separating Hash Families

Special cases of separating hash families have been studied in various literatures under many different names. Our definition is an adaptation of the definition in [33]. We first define a hash family.

Definition 8.1. Let X and Y be two finite sets such that $|X| = n$ and $|Y| = m$. A *hash family* $\mathcal{F}$ is a family of functions $\{f_i : X \to Y, i \in [N]\}$, for some positive integer N.

The term 'separating' in 'separating hash families' comes from the following definition.

Definition 8.2. Let X and Y be two finite sets. Let f be a function mapping from X to Y. Let A and $B \subseteq X$. We say f *separates* A and B when $f(A) \cap f(B) = \emptyset$.

Let m, n and t be positive integers, and let $w_1, w_2, ..., w_t$ be positive integers in non-decreasing order. To avoid trivial cases, assume that $m \geq 2$ and $t \geq 2$.

Definition 8.3. Let X and Y be two finite sets such that $|X| = n$ and $|Y| = m$. Let $\mathcal{F}$ be a family of functions $\{f_i : X \to Y, i \in [N]\}$, for some positive integer N. Then $\mathcal{F}$ is an $(N; n, m, \{w_1, w_2, ..., w_t\})$-*separating hash family*, or an SHF$(N; n, m, \{w_1, w_2, ..., w_t\})$, if for any pairwise disjoint $C_1, C_2, ..., C_t \subseteq X$ such that $|C_j| \leq w_j, j \in [t]$, there exists $i \in [N]$ such that f_i separates $C_1, C_2, ..., C_t$ (i.e. $f_i(C_1), f_i(C_2), ..., f_i(C_t)$ are pairwise disjoint).

For any SHF$(N; n, m, \{w_1, w_2, ..., w_t\})$, we define its *size*, *length*, and *type* to be n, N, and $\{w_1, w_2, ..., w_t\}$, respectively.

The original definition of separating hash family had the stronger constraint that $|C_i| = w_i$. In our definition $|C_i| \leq w_i$. Such a change makes no difference when n is large enough, but allows separating hash families to exist even when n is less than $\sum_{i=1}^{t} w_i$. This is a similar situation to fingerprinting codes in the sense that the coalition size can be anything up to k.

Example 22. Let $X = \{1, 2, 3, 4\}$ and $Y = \{a, b, c, d\}$. Let $\mathcal{F} = \{f_1, f_2, f_3\}$, where

$$\begin{aligned} f_1(1) = a, \quad f_1(2) = a, \quad f_1(3) = b, \quad f_1(4) = d, \\ f_2(1) = a, \quad f_2(2) = b, \quad f_2(3) = c, \quad f_2(4) = c, \\ f_3(1) = b, \quad f_3(2) = c, \quad f_3(3) = c, \quad f_3(4) = d. \end{aligned}$$

Then $\mathcal{F}$ is an SHF$(3; 4, 4, \{1, 3\})$, but not an SHF$(3; 4, 4, \{2, 2\})$.

It is easy to see that $\mathcal{F}$ is an SHF$(3; 4, 4, \{1, 3\})$ since f_2 can separate $\{2\}$ from $\{1, 3, 4\}$, and $\{1\}$ from $\{2, 3, 4\}$, f_1 can separate $\{3\}$ from $\{1, 2, 4\}$, and $\{4\}$ from $\{1, 2, 3\}$. However, $\{1, 3\}$ and $\{2, 4\}$ can not be separated by any member of $\mathcal{F}$. Thus, $\mathcal{F}$ is not an SHF$(3; 4, 4, \{2, 2\})$.

Once the parameters $N, m, \{w_1, w_2, ..., w_t\}$ have been given, it is not difficult to construct a separating hash family provided that n is small enough. The problem is therefore to maximise n. If n is the largest number possible for the given $N, m, \{w_1, w_2, ..., w_t\}$, then we say that the separating hash family is *optimal*.

A separating hash family $\mathcal{F}$ can be portrayed as an $N \times n$ matrix where each column is represented by a member of X and each row i represents the function f_i. Here, the element in row i and column x, for some $x \in X$, is the value of $f_i(x)$. The family $\mathcal{F}$ in Example 22 can be illustrated by the following matrix:

$$\begin{pmatrix} a & a & b & d \\ a & b & c & c \\ b & c & c & d \end{pmatrix}.$$

Let C be q-ary length ℓ code over an alphabet Q. We can construct a hash family $\mathcal{H}(C)$ from C by defining $\mathcal{H}(C)$ to be a hash family $\{f_i : C \to Q, i \in [\ell]\}$ where $f_i(x) = x_i$ for any x in C.

On the other hand, we can construct a code C over an alphabet Q from an existing hash family. Let $|X| = n$ and $|Y| = m$, and let $\mathcal{F} = \{f_i : X \to Y, i \in [N]\}$, for some positive integer N, be a hash family. Let $C(\mathcal{F}) = \{(f_1(x), f_2(x), ..., f_\ell(x)) : x \in X\}$. Then $C(\mathcal{F})$ is a m-ary length N code over an alphabet Y containing n codewords.

The next two theorems show how separating hash families are related to secure frameproof and frameproof codes. (See [30] for a more extensive literature review.)

Theorem 8.1.1 ([33]). *There is a q-ary length ℓ k-FP code C if and only if there is an* $\mathrm{SHF}(\ell; |C|, q, \{1, k\})$.

Proof. Let C be a q-ary length ℓ k-FP code. Let $C_1, C_2 \subseteq C$ be disjoint, where $|C_1| \leq 1$ and $|C_2| \leq k$. By the k-FP property, $\mathrm{desc}(C_2) \cap C = C_2$. Hence, $\mathrm{desc}(C_2) \cap C_1 = \emptyset$. Therefore, there exists a coordinate $i \in [\ell]$ where none of codewords in C_2 agree with the codeword in C_1. This implies $f_i(C_1) \cap f_i(C_2) = \emptyset$. We find that $\mathcal{H}(C)$ is an $\mathrm{SHF}(\ell; |C|, q, \{1, k\})$ as required.

Let $|X| = n, |Y| = q$, and let $\mathcal{F} = \{f_i : X \to Y, i \in [N]\}$ be an SHF$(\ell; |C|, q, \{1, k\})$. Let $C = C(\mathcal{F})$.

Let C_2 be a subset of C of size at most k. For any codeword $x \in C \backslash C_2$, there exists an $i \in [\ell]$ such that $f_i(\{x\}) \cap f_i(C_2) = \emptyset$. Hence, $\text{desc}(C_2) \cap \{x\} = \emptyset$. Consider

$$\begin{aligned}
\text{desc}(C_2) \cap C &= \text{desc}(C_2) \bigcap (C_2 \cup C \backslash C_2) \\
&= (\text{desc}(C_2) \cap C_2) \bigcup (\text{desc}(C_2) \cap C \backslash C_2) \\
&= (\text{desc}(C_2) \cap C_2) \bigcup \left(\text{desc}(C_2) \cap \bigcup_{x \in C \backslash C_2} \{x\} \right) \\
&= (\text{desc}(C_2) \cap C_2) \bigcup \left(\bigcup_{x \in C \backslash C_2} (\text{desc}(C_2) \cap \{x\}) \right) \\
&= (\text{desc}(C_2) \cap C_2) \bigcup \left(\bigcup_{x \in C \backslash C_2} \emptyset \right) \\
&= C_2 \cup \emptyset \\
&= C_2.
\end{aligned}$$

Hence, C is a q-ary length ℓ k-FP code. □

Theorem 8.1.2 ([33]). *There is a q-ary length ℓ k-SFP code if and only if there is an* SHF$(\ell; |C|, q, \{k, k\})$.

Proof. Let C_1, C_2 be disjoint subsets of C of size at most k. By the k-SFP property, $\text{desc}(C_1) \cap \text{desc}(C_2) = \emptyset$. Therefore, there exists a coordinate $i \in [\ell]$ where none of the codewords in C_1 agree with the codewords in C_2. This implies $f_i(C_1) \cap f_i(C_2) = \emptyset$. Again, we get $\mathcal{H}(C)$ is an SHF$(\ell; |C|, q, \{k, k\})$ as required.

Let $|X| = n, |Y| = q$, and let $\mathcal{F} = \{f_i : X \to Y, i \in [N]\}$ be an SHF$(\ell; |C|, q, \{k, k\})$. Let $C = C(\mathcal{F})$.

Let C_1, C_2 be disjoint subsets of C of size at most k. Then there exists an $i \in [\ell]$ such that $f_i(C_1) \cap f_i(C_2) = \emptyset$. Hence, none of the codewords of C_1 agree with the

codewords of C_2 in coordinate i. Thus $\text{desc}(C_1) \cap \text{desc}(C_2) = \emptyset$. Therefore, C is a q-ary k-SFP code of length ℓ. □

We can abuse terminology and say that k-FP codes are separating hash families of type $\{1, w\}$ and k-SFP codes are separating hash families of type $\{w, w\}$. IPP and traceability codes are not in general equivalent to a class of separating hash families, although 2-IPP codes are separating hash families of type {1, 1, 1} and {2, 2} simultaneously, see [21].

Since the special cases of separating hash families that are closely related to fingerprinting codes are of type $\{w_1, w_2\}$ only, from this point onward, we reduce our field of interest down to separating hash families of type $\{w_1, w_2\}$. Furthermore the following two theorems give us a stronger reason to narrow our field of interest down. Both theorem are easy consequences of the definition of separating hash family.

Theorem 8.1.3 ([30, 36]). *Suppose that $\mathcal{F}$ is an* $\text{SHF}(N; n, m, \{w_1, w_2, ..., w_t\})$*, and let $w_1' \leq w_1$. Then $\mathcal{F}$ is also an* $\text{SHF}(N; n, m, \{w_1', w_2, ..., w_t\})$.

This theorem holds for the following reason. If there exists a hash function $f \in \mathcal{F}$ that pairwise separates a collections of sets C_i of size w_i, f can also separate the sets C_i when C_1 is replaced by a smaller set.

Theorem 8.1.4 ([30, 36]). *Suppose that $\mathcal{F}$ is an* $\text{SHF}(N; n, m, \{w_1, w_2, ..., w_t\})$*, and define $w_1' = w_1 + w_2$. Then $\mathcal{F}$ is also an* $\text{SHF}(N; n, m, \{w_1', w_3, ..., w_t\})$.

This theorem holds for the following reason. If there exists a hash function $f \in \mathcal{F}$ that pairwise separates a collections of sets C_i of size w_i, f can also separate the sets C_i when C_1 and C_2 are replaced by a union of C_1 and C_2.

Hence, the size of a separating hash family of type $\{w_1, w_2, ..., w_t\}$ is always bounded above by the size of the largest separating hash family of type $\{w_1', w_2'\}$ for some w_1', w_2'.

8.2 Bounds on the Size of Separating Hash Families

In this section, we present the known bounds on the size of separating hash families of type $\{w_1, w_2\}$. The best general result on the upper bounds on the size of separating hash families of type $\{w_1, w_2\}$ is stated by Bazrafshan and van Trung [7]. Their result is as follows.

Theorem 8.2.1. *Let m, n be positive integers, and let w_1, w_2 be positive integers in non-decreasing order. If there exists an* $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$*, then* $n \leq (w_1 + w_2 - 1)m^{\left\lceil \frac{N}{w_1+w_2-1} \right\rceil}$.

The case of length 1 is trivial as can be seen in the two following theorems.

Theorem 8.2.2. *Let m, n be positive integers greater than 1, and let w_1, w_2 be positive integers in non-decreasing order. If there exists an* $\mathrm{SHF}(1; n, m, \{w_1, w_2\})$*, then* $n \leq m$.

Proof. Let $\mathcal{F} = \{f_1 : X \to Y\}$ be an $\mathrm{SHF}(1; n, m, \{w_1, w_2\})$. Assume that $n \geq m + 1$. Then there exist x and y in X such that $x \neq y$ and $f(x) = f(y)$. Let C_1 and $C_2 \subseteq C$ be disjoint, such that $x \in C_1$, $y \in C_2$. Then f cannot separate C_1 and C_2, contradicting the $\mathrm{SHF}(1; n, m, \{w_1, w_2\})$ property. Therefore, $n \leq m$. □

The following matrix gives an $\mathrm{SHF}(1; m, m, \{w_1, w_2\})$.

$$\begin{pmatrix} 1 & 2 & \dots & m \end{pmatrix}$$

Hence, one can see that the following theorem holds.

Theorem 8.2.3. *Let m be a positive integer greater than 1, and let w_1, w_2 be positive integers in non-decreasing order. An* $\mathrm{SHF}(1; m, m, \{w_1, w_2\})$ *exists and is optimal.*

The case of type $\{1, 1\}$ is also trivial as can be seen in the two following theorems.

Theorem 8.2.4. *Let m, n, N be positive integers greater than 1. If there exists an* $\mathrm{SHF}(N; n, m, \{1, 1\})$*, then* $n \leq m^N$.

Proof. Let $\mathcal{F} = \{f_1, f_2, ..., f_N : X \to Y\}$ be an SHF$(N; n, m, \{1, 1\})$. Assume that $n \geq m^N + 1$. Then there exist x and y in X such that $x \neq y$ and $(f_1(x), f_2(x), ..., f_N(x)) = (f_1(y), f_2(y), ..., f_N(y))$. Let $C_1 = \{x\}$ and $C_2 = \{y\}$. Then C_1, C_2 are disjoint and none of the functions $f \in \mathcal{F}$ can separate C_1 and C_2, contradicting the SHF$(N; n, m, \{1, 1\})$ property. Therefore, $n \leq m^N$. □

The following matrix gives an SHF$(N; m^N, m, \{1, 1\})$. Note that every possible column appears exactly once.

$$\left(\begin{array}{ccccc|ccccc|c|ccccc}
1 & 1 & ... & 1 & 1 & 2 & 2 & ... & 2 & 2 & ... & m & m & ... & m & m \\
1 & 1 & ... & m & m & 1 & 1 & ... & m & m & ... & 1 & 1 & ... & m & m \\
\vdots & \vdots & ... & \vdots & \vdots & \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots & \vdots \\
1 & 1 & ... & m & m & 1 & 1 & ... & m & m & ... & 1 & 1 & ... & m & m \\
1 & 2 & ... & m-1 & m & 1 & 2 & ... & m-1 & m & ... & 1 & 2 & ... & m-1 & m
\end{array}\right)$$

Hence, one can see that the following theorem holds.

Theorem 8.2.5. *Let m, N be positive integers greater than 1. An* SHF$(N; m^N, m, \{1, 1\})$ *exists and is optimal.*

The Construction 3.1.2 for k-FP codes gives the following lower bound:

Theorem 8.2.6. *Let q be a prime power. There exists an* SHF$(N; q^{\lceil \frac{N}{k} \rceil}, q, \{1, k\})$.

The probabilistic results on the existence of k-IPP codes in Theorem 3.3.5 also gives a lower bound for separating hash families of type $\{2, 2\}$.

Corollary 8.2.7. *Let m be a positive integer greater than 1.*
There exists an SHF$(N; m^{RN}, m, \{2, 2\})$ *where*

$$R \geq \frac{1}{3} \log_m \frac{m^3}{5m^2 - 8m + 4}.$$

Liu and Shen proposed explicit constructions of separating hash families from algebraic curves over finite fields in [24] and achieved the following results.

Theorem 8.2.8 ([24], Theorem 3.2)**.** *Let q be a prime power and let $c_1 \geq c_2 \geq 2$ be real numbers. There exists an* $\mathrm{SHF}((q-1)q^i; q^{(c_1c_2-2)q^i}, q^2, \{\left\lfloor \frac{\sqrt{2}}{c_1}(q^{\frac{1}{2}}-1)\right\rfloor, \left\lfloor \frac{\sqrt{2}}{c_2}(q^{\frac{1}{2}}-1)\right\rfloor\})$.

Stinson, Wei and Zhu [35] provide constructions for separating hash families using error-correcting codes and orthogonal arrays, and establish the following existence results for separating hash families of type $\{w_1, w_2\}$; they fixed n, m, w_1 and w_2, and let N grow.

Theorem 8.2.9 ([35], Theorem 4.4)**.** *For any positive integers m, w_1 and w_2, there exists an infinite class of* $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ *for which N is $O((w_1w_2)^{\log^*(n)}(\log n))$.*

Where $\log^*(n)$ is defined recursively as follows:

$$\log^*(n) = \begin{cases} 1 & \text{when n=1} \\ \log^*(\lceil \log n \rceil) + 1 & \text{otherwise} \end{cases}.$$

Chapter 9

Frameproof Codes: Separating Hash Families Type $\{1, k\}$

In this chapter we focus on improving the previously known bounds for separating hash families in some special cases that are related to frameproof codes. We achieve new tight upper bounds for the size of 2-FP codes and k-FP codes when the length $\ell = 1 \bmod k$. Our bounds are the best possible for many parameter values. We also achieve new tight upper bounds for the size of k-FP codes when the length ℓ satisfies $k < \ell \leq 2k$. Our new bounds resemble Theorem 3.1.2, which gives the best previously known bounds for frameproof codes, but without the term $O\left(q^{\lceil \frac{\ell}{k} \rceil - 1}\right)$.

9.1 2-FP codes

In this section we aim to improve the previously known upper bound on the size of hash families of type $\{1, 2\}$. This is equivalent to improving the known upper bounds on 2-FP codes.

We aim to prove the following result.

Theorem 9.1.1. *Let m, n be positive integers greater than 1, and let d be a non-negative integer. If there exists an* $\mathrm{SHF}(2d + r; n, m, \{1, 2\})$*, where $r \in [2]$, then $n \leq$*

$r(m-(r-1))^{d+1}$.

Which is equivalent to proving:

Theorem 9.1.2. *Let C be a q-ary 2-FP code of length $\ell = 2d + r$, where d is a non-negative integer and $r \in [2]$. Then*

$$|C| \leq r(m-(r-1))^{d+1}.$$

When $r = 2$, i.e., when N is even in Theorem 9.1.1, the theorem follows from Theorem 3.1.1. For if we substitute ℓ by $2d + 2$ in Theorem 8.2.1, we obtain the following corollary.

Corollary 9.1.3. *Let m, n be positive integers greater than 1, and let d be a positive integer. If there exists an* $\mathrm{SHF}(2d; n, m, \{1, 2\})$*, then $n \leq 2(m-1)^d$.*

However, note that Theorem 3.1.1. does not give the result we want when N is odd. The best previously known bound in this case can be obtained by substituting ℓ by $2d + 1$ in Theorem 3.1.2, to obtain the following corollary.

Corollary 9.1.4. *Let m, n be positive integers greater than 1, and let d be a non-negative integer. If there exists an* $\mathrm{SHF}(2d+1; n, m, \{1, 2\})$*, then $n \leq m^{d+1} + O\left(m^d\right)$.*

Thus, only the case $r = 1$, i.e., the case when N is odd, is yet to be shown.

To make it easier for us to generate proofs for better bounds on the size of $\mathrm{SHF}(N; n, m, \{1, 2\})$, it is necessary that we introduce some additional terms and notation.

Definition 9.1. Let $\mathcal{F} = \{f_i : X \to Y, i \in [N]\}$ be an $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$.

For any $x \in X$, any $i \in [N]$, and any $I \subseteq [N]$, let $x_i = f_i(x)$, and let $x_I = (f_j(x))_{j \in I}$. We say x is *unique* under I if $|\{z \in X : z_I = x_I\}| = 1$, and we say x is *non-unique* under I when $|\{z \in X : z_I = x_I\}| > 1$. For any $I \subseteq \{1, 2, ..., N\}$, let $U_I = \{x \in X : x \text{ is unique under } I\}$, and let $V_I = \{x \in X : x \text{ is non-unique under } I\}$.

The following theorem gives new bounds on the size of $\mathrm{SHF}(N; n, m, \{1, 2\})$ in the case when N is odd, thus establishing Theorem 9.1.1.

Theorem 9.1.5. *Let m, n be positive integers greater than 1, and let d be a non-negative integer. If there exists an $\mathrm{SHF}(2d+1; n, m, \{1, 2\})$, then $n \leq m^{d+1}$.*

The bounds in Theorem 9.1.5 are the best possible for many parameter values. Li, van Rees and Wei [23] show that the bounds in Theorem 9.1.5 are optimal when $N = 3$ and provide an explicit construction of an optimal $\mathrm{SHF}(3; m^2, m, \{1, 2\})$ using Steiner triple systems of order m, when $m = 1, 3 \bmod 6$. The bounds in Theorem 9.1.5 are also optimal when m is a prime power. The easiest example would be constructing a corresponding frameproof code by Reed-Solomon code of minimum distance $d = \frac{1}{2}(\ell + 1)$. (See Construction 3.1.2 for 2-FP codes.)

Proof of Theorem 9.1.5. Let $\mathcal{F} = \{f_1, f_2, ..., f_{2d+1} : X \to Y\}$ be an $\mathrm{SHF}(2d+1; n, m, \{1, 2\})$. Assume that $n \geq m^{d+1} + 1$.

Let $\mathcal{I}_{d+1}$ be the set of all $(d+1)$-subsets of $[2d+1]$. For any $I \in \mathcal{I}_{d+1}$, since $|X| \geq m^{d+1} + 1$ there are at least $\left\lceil \frac{m^{d+1}+1}{m^{d+1}} \right\rceil = 2$ elements x and y from X such that $x_I = y_I$, by the pigeonhole principle.

Let $I \in \mathcal{I}_{d+1}$ maximise the number of $x \in X$ where x is non-unique under I. Let $X' = V_I$. Denote $|X'|$ by s. Then $s \geq 2$.

Observe that for any $x \in X'$, it is necessary that x is unique under $J = [2d+1] \backslash I$. To show this, assume that there exists $z \in X \backslash \{x\}$ such that $z_J = x_J$. So, x and z can not be separated by any $f_i \in \mathcal{F}$ where $i \in J$. Since $x \in X'$ there exists $x' \in X \backslash \{x\}$ such that $f_I(x') = f_I(x)$. We have that $x' \neq z$ for otherwise we get $f_i(z) = f_i(x)$ for all $i \in [2d+1]$, which implies $x = z$. Therefore $f(\{x\}) \cap f(\{x', z\}) \neq \emptyset$ for all $f \in \mathcal{F}$. Hence the contradiction occurs.

Now we consider $X \backslash X'$. All $x \in X'$ are unique under J, hence the image of $f_J|_{X \backslash X'}$ has size at most $m^d - s$. Denote the distinct images of $X \backslash X'$ by $y_1, y_2, ..., y_{m^d - s}$. Let $X_i = \{x \in X \backslash X' : f_J(x) = y_i\}$.

Let $j \in I$ be fixed. Define $I_0 \in \mathcal{I}_{d+1}$ by $I_0 = J \cup \{j\}$. Observe that each X_i contributes at least $|X_i| - m$ non-unique images under f_{I_0}, since f_j maps to at most m alphabet symbols. Therefore, the number of $x \in X$ that are non-unique under I_0 is at least

$$\begin{aligned}\sum_{k=1}^{m^d-s} (|X_i| - m) &= \sum_{k=1}^{m^d-s} |X_i| - \sum_{k=1}^{m^d-s} m \\ &= |X \backslash X'| - m(m^d - s) \\ &\geq (m^{d+1} + 1 - s) - m(m^d - s) \\ &= m^{d+1} + 1 - s - m^{d+1} + sm \\ &= s(m-1) + 1.\end{aligned}$$

Thus, the number of $x \in X$ such that $f_{I_0}(x)$ is non-unique is at least $s(m-1)+1 \geq s+1 > s$, which contradicts our choice of I. Therefore $n \leq m^{d+1}$. □

We apply the same technique as in the proof of Theorem 9.1.5 to prove a new upper bound for the more general case in the next section.

The results by Bazrafshan and van Trung [8] also assert that our bound is the best possible for 2-FP code of odd length. They derive a contradiction by finding an existence of a forbidden recursive pattern when the size of the code is greater than m^{d+1}.

9.2 k-FP codes of length ℓ where $\ell = 1 \bmod k$

In this section, we aim to prove the following result which is actually a generalised version of Theorem 9.1.5.

Theorem 9.2.1. *Let m, n and k be positive integers greater than 1 where $m \geq 2(k-1)$, and let h be a non-negative integer. If there exists an* $\mathrm{SHF}(kh+1; n, m, \{1, k\})$, *then* $n \leq m^{h+1}$.

Which is equivalent to proving the following theorem.

Theorem 9.2.2. *Let C be a q-ary k-FP code of length $\ell = kh+1$, where $q \geq 2(k-1)$. Then*

$$|C| \leq q^{h+1}.$$

At a glance, one can see that the bound is much tighter than the bound from Theorem 8.2.1. Moreover, this bound gives the same leading term as in Theorem 3.1.2 without the term $O\left(q^{\lceil \frac{\ell}{k} \rceil - 1}\right)$.

Again, the Construction 3.1.2 ensures that the bounds in Theorem 9.2.1 are optimal when m is a prime power.

We use the same techniques as in the proof of Theorem 9.1.5 to prove Theorem 9.2.1. However, proving the existence of the set J is not as simple. Hence, it requires some extra work.

Here we give the definition and a relevant theorem of a combinatorial object that will be useful in the proof of Theorem 9.2.1.

Definition 9.2. A family $\mathcal{S}$ of subsets of a set is *t-colliding* if $\mathcal{S}$ does not contain t pairwise disjoint subsets.

Theorem 9.2.3 ([10], Theorem 11)**.** *Let t, k and ℓ be positive integers such that $\ell \geq tk$. Let $\mathcal{S}$ be a t-colliding family of subsets of $[\ell]$, where $|S| = k$ for all $S \in \mathcal{S}$. Then*

$$|\mathcal{S}| \leq \binom{\ell}{k} \frac{(t-1)k}{\ell}.$$

Proof of Theorem 9.2.1. Let $\mathcal{F} = \{f_1, f_2, ..., f_{kh+1} : X \to Y\}$ be an $\mathrm{SHF}(kh+1; n, m, \{1, k\})$. Assume that $n \geq m^{h+1} + 1$.

For any $i \in [kh+1]$, let $\mathcal{I}_i$ be the set of all i-subsets of $[\ell]$. For any $I \in \mathcal{I}_{h+1}$, there are at least 2 elements $x, x' \in X$ with $x_I = x'_I$, by the pigeonhole principle.

Let $I_{\max} \in \mathcal{I}_{h+1}$ maximise the number of $x \in X$ where x is non-unique under $I_{\max}$. Let $s = |V_{I_{\max}}|$. The previous paragraph shows that $s \geq 2$.

Claim. There exists $J \in \mathcal{I}_h$ such that $J \cap I_{\max} = \emptyset$ and at least $\frac{1}{k-1}|V_{I_{\max}}|$ elements $x \in V_{I_{\max}}$ are unique under J.

Once we are confident that the claim is true, the rest of the proof follows as in the later part of the proof of Theorem 9.1.5. However, unlike in Theorem 9.1.5, the validity of the claim is not easy to see. Hence we need extra work to justify the claim.

For each $x \in V_{I_{\max}}$ define $\mathcal{J}_x$ to be the set of all h-subsets I' of $[\ell]\backslash I_{\max}$ such that x is non-unique under I'. Then, $\mathcal{J}_x$ must be a $(k-1)$-colliding family; we can see this as follows.

Assume that $\mathcal{J}_x$ is not a $(k-1)$-colliding family. Then there exist pairwise disjoint sets $J_1, J_2, ..., J_{k-1}$ in $\mathcal{J}_x$ such that

$$\bigcup_{i=1}^{k-1} J_i = [\ell]\backslash I_{\max}.$$

Let z be an element of $V_{I_{\max}}\backslash\{x\}$ such that $z_{I_{\max}} = x_{I_{\max}}$, and, for each $i \in [k-1]$, let z^i be an element of $X\backslash\{x\}$ such that $z^i_{J_i} = x_{J_i}$. This makes $f(\{x\}) \cap f(\{z^1, z^2, ..., z^{k-1}, z\}) \neq \emptyset$ for all $f \in \mathcal{F}$, contradicting to the SHF$(kh+1; n, m, \{1, k\})$ property of $\mathcal{F}$. Hence, $\mathcal{J}_x$ is a $(k-1)$-colliding family.

Therefore by Theorem 9.2.3,

$$\begin{aligned} |\mathcal{J}_x| &\leq \binom{\ell-(h+1)}{h} \frac{((k-1)-1)h}{\ell-(h+1)} \\ &= \binom{(k-1)h}{h} \frac{(k-2)h}{(k-1)h} \\ &= \binom{(k-1)h}{h} \frac{(k-2)}{(k-1)}. \end{aligned}$$

Since there are $\binom{\ell-(h+1)}{h} = \binom{(k-1)h}{h}$ different h-subsets of $[\ell]\backslash I_{\max}$, the number of

h-subsets I of $[\ell]\backslash I_{\max}$ such that x is unique under I is

$$\begin{aligned}\binom{(k-1)h}{h} - |\mathcal{J}_x| &\geq \binom{(k-1)h}{h} - \binom{(k-1)h}{h}\frac{(k-2)}{(k-1)} \\ &= \binom{(k-1)h}{h}\frac{1}{(k-1)}.\end{aligned}$$

This implies each $x \in V_{I_{\max}}$ is unique in at least $\binom{(k-1)h}{h}\frac{1}{(k-1)}$ h-subsets of $[\ell]\backslash I_{\max}$. Hence, there exists $J \in \mathcal{I}_h$ such that $J \cap I_{\max} = \emptyset$ and at least $\frac{1}{k-1}|V_{I_{\max}}|$ elements $x \in V_{I_{\max}}$ are unique under J. This establishes our claim and from this point we can deploy the technique we use in Theorem 9.1.5.

Now we consider $X\backslash V_{I_{\max}}$. The number of h-tuples of the form x_J when $x \in X\backslash V_{I_{\max}}$ is at most $m^h - \frac{1}{k-1}s$, by our choice of J. Denote the distinct h-tuples x_J when $x \in X\backslash V_{I_{\max}}$ by $y_1, y_2, ..., y_{m^h - \frac{1}{k-1}s}$. Let $X_i = \{x \in X\backslash V_{I_{\max}} : x_J = y_i\}$.

Let $j \in [kh+1]\backslash J$ be fixed, and define $I_0 \in \mathcal{I}_{h+1}$ by $I_0 = J \cup \{j\}$. Observe that each X_i contributes at least $|X_i| - m$ non-unique $(h+1)$-tuples under I_0, since there are at most m symbols occur in the j^{th} coordinate.

Therefore, the number of $x \in X$ that x_{I_0} is non-unique is at least

$$\begin{aligned}\sum_{h=1}^{m^h - \frac{1}{k-1}s}(|X_i| - m) &= \sum_{h=1}^{m^h - \frac{1}{k-1}s}|X_i| - \sum_{h=1}^{m^h - \frac{1}{k-1}s} m \\ &= |X\backslash V_{I_{\max}}| - (m^h - \frac{1}{k-1}s)m \\ &\geq (m^{h+1} + 1 - s) - (m^h - \frac{1}{k-1}s)m \\ &= m^{h+1} + 1 - s - m^{h+1} + \frac{1}{k-1}sm \\ &= 1 + s(\frac{1}{k-1}m - 1).\end{aligned}$$

Thus, the number of $x \in X$ that non-unique under I_0, is at least

$$1 + s(\frac{1}{k-1}m - 1).$$

Recall that $m \geq 2(k-1)$. Hence

$$\begin{aligned} 1 + s(\frac{1}{k-1}m - 1) &\geq 1 + s \\ &> s, \end{aligned}$$

which contradicts our choice of I_{max}.

Therefore $n \leq m^{h+1}$. Now, we obtain Theorem 9.2.1 as required. □

9.3 k-FP codes of length ℓ where $k < \ell \leq 2k$

In this Section, we aim to prove the following theorem.

Theorem 9.3.1. *Let m, n, k be positive integers greater than 1, where $m > k$, and let r be an integer such that $0 < r \leq k$. If $\mathcal{F}$ is an* SHF$(k + r; n, m, \{1, k\})$. *Then*

$$n \leq \gamma m^2,$$

where $\gamma = \frac{k+r}{k-r+2}$.

Which is equivalent to proving the following theorem in frameproof codes language.

Theorem 9.3.2. *Let C be a q-ary k-FP code of length ℓ where $k < \ell \leq 2k$ and $m > k$. Let r be positive integer such that $r = \ell - k$. Then $0 < r \leq k$ and*

$$n \leq \gamma m^2,$$

where $\gamma = \frac{k+r}{k-r+2}$.

The recent recursive construction by Meng Chee and Zhang [25] ensures that when $r = 2$ there always exists a k-FP code of size $\frac{k+2}{k}(m-1)^2 + 1$. Hence our bound is tight for some parameters.

We first compare our bound with two previously known results, one from separating hash families, one from frameproof codes. Here is a bound derived from the general

result on upper bounds of separating hash families by Bazrafshan and van Trung [7], see Theorem 8.2.1.

Corollary 9.3.3. *Let m, n, k be positive integers greater than 1 and let r be an integer such that $0 < r \leq k$. If $\mathcal{F}$ is an* SHF$(k + r; n, m, \{1, k\})$. *Then*

$$n \leq km^2.$$

We now state the best previously known result on the upper bound of frameproof codes by Blackburn [10]. Note that the original result is written in frameproof codes language (see Theorem 3.1.2).

Corollary 9.3.4. *Let m, n, k be positive integers greater than 1 and let r be an integer such that $0 < r \leq k$. If $\mathcal{F}$ is an* SHF$(k + r; n, m, \{1, k\})$. *Then*

$$n \leq \gamma m^2 + O\left(m^{\lceil \frac{\ell}{k} \rceil - 1}\right),$$

where $\gamma = \frac{k+r}{k-r+2}$.

Note that $1 \leq \gamma \leq k$, where $\gamma = 1$ and k only when $r = 1$ and k, respectively. Hence the leading term in Corollary 9.3.4 is generally better than the leading term in Corollary 9.3.3 . Now we prove our new result, Theorem 9.3.1, which eliminates the term $O\left(m^{\lceil \frac{\ell}{k} \rceil - 1}\right)$ from the bound in Corollary 9.3.4.

Proof of Theorem 9.3.1. Let $\mathcal{F} = \{f_1, f_2, ..., f_{k+r} : X \to Y\}$ be an SHF$(k+r; n, m, \{1, k\})$.

Let $S_1, S_2, ..., S_k$ be pairwise disjoint subsets of $[k+r]$, where the cardinality of S_i is 2 for $i \leq r$ and 1 otherwise. Recall the definition of unique and non-unique in Section 9.1. We claim

1. $S_1 \cup S_2 \cup ... \cup S_k = [k + r]$,

2. $U_{S_1} \cup U_{S_2} \cup ... \cup U_{S_k} = X$.

The first assertion is not difficult to see since S_i are pairwise disjoint and $\sum_{i=1}^{k} |S_i| = 2r + 1(k - r) = k + r$. The later one can be seen from the following contradiction.

Assume for a contradiction that $U_{S_1} \cup U_{S_2} \cup ... \cup U_{S_k} \neq X$. Then, there exists $x \in X \setminus (U_{S_1} \cup U_{S_2} \cup ... \cup U_{S_k})$. Hence $x \notin U_{S_i}$ for all $i \in [k]$. Therefore, for every $i \in [k]$, there exists $y^i \in X \setminus \{x\}$ such that $f_j(y^i) = f_j(x)$ for all $j \in S_i$, i.e., none of functions f_j, $j \in S_i$ can separate x and y^i. Let $C_1 = \{x\}$, $C_2 = \{y^1, ..., y^k\}$. We have $|C_1| \leq 1, |C_2| \leq k$ and C_1, C_2 are disjoint. Since $S_1 \cup S_2 \cup ... \cup S_k = [k + r]$, none of function $f_j \in \mathcal{F}$ can separate C_1 and C_2, contradicting the SHF$(k + r; n, m, \{1, k\})$ property of $\mathcal{F}$. Therefore, $U_{S_1} \cup U_{S_2} \cup ... \cup U_{S_k} = X$. This proves our claim.

Let

$$W = \bigcup_{S \subseteq [k+r]: |S|=1} U_S$$

and let $Z = X \setminus W$. For any $I \subseteq [k+r]$, define $\Gamma_I = U_I \cap Z$. For any choice of $S_1, ..., S_k$, we have that $\Gamma_{S_i} = U_{S_i} \cap Z = \emptyset$ whenever $i \geq r + 1$ and

$$\begin{aligned} \Gamma_{S_1} \cup \Gamma_{S_2} \cup ... \cup \Gamma_{S_r} &= (U_{S_1} \cap Z) \cup (U_{S_2} \cap Z) \cup ... \cup (U_{S_r} \cap Z) \\ &= (U_{S_1} \cap Z) \cup (U_{S_2} \cap Z) \cup ... \cup (U_{S_k} \cap Z) \\ &= (U_{S_1} \cup U_{S_2} \cup ... \cup U_{S_k}) \cap Z \\ &= X \cap Z \\ &= Z. \end{aligned} \tag{9.1}$$

By the definition of Z and W, we have

$$\begin{aligned} |Z| &= |X \backslash W| \\ &= |X| - |W| \\ &\geq |X| - (|U_{\{1\}}| + |U_{\{2\}}| + ... + |U_{\{k+r\}}|) \\ &= n - \sum_{i \in [k+r]} |U_{\{i\}}|. \end{aligned} \tag{9.2}$$

We improve our upper bound of $\mathrm{SHF}(k + r; n, m, \{1, k\})$ by providing an upper bound on $|Z|$ by counting the elements of the following set K in two ways:

$$K = \{(x, S) : x \in \Gamma_S, S \subseteq [k + r] \text{ of cardinality } 2\}.$$

There are $\binom{k+r}{2}$ choices for the subset S. For any $x \in Z$, let $\mathcal{J}_x$ be defined by

$$\mathcal{J}_x = \{S \subset [k + r] : |S| = 2 \text{ and } x \notin \Gamma_S\}.$$

Once x is fixed, there are $\binom{k+r}{2} - |\mathcal{J}_x|$ choices for S such that $(x, S) \in K$.

$\mathcal{J}_x$ is r-colliding since if there exist pairwise disjoint subsets $S_1, S_2, ..., S_r \in \mathcal{J}_x$, then $x \notin \Gamma_{S_1} \cup \Gamma_{S_2} \cup ... \cup \Gamma_{S_r}$. This implies $x \notin Z$ by (9.1), contradicting our choice of x. Hence, $\mathcal{J}_x$ is r-colliding. Therefore, by Theorem 9.2.3,

$$|\mathcal{J}_x| \leq \binom{k + r}{2} \frac{2(r - 1)}{k + r}. \tag{9.3}$$

Therefore,

$$\begin{aligned}
|K| &= \sum_{x \in Z} \left(\binom{k+r}{2} - |\mathcal{J}_x| \right) \\
&\geq \sum_{x \in Z} \left(\binom{k+r}{2} - \binom{k+r}{2} \frac{2(r-1)}{k+r} \right) \text{ by (9.3)} \\
&= |Z| \left(\binom{k+r}{2} - \binom{k+r}{2} \frac{2(r-1)}{k+r} \right) \\
&\geq (n - \sum_{i \in [k+r]} |U_{\{i\}}|) \left(\binom{k+r}{2} - \binom{k+r}{2} \frac{2(r-1)}{k+r} \right) \text{ by (9.2)} \\
&= (n - \sum_{i \in [k+r]} |U_{\{i\}}|) \binom{k+r}{2} \left(1 - \frac{2(r-1)}{k+r} \right) \\
&= (n - \sum_{i \in [k+r]} |U_{\{i\}}|) \binom{k+r}{2} \left(\frac{(k+r) - 2(r-1)}{k+r} \right) \\
&= (n - \sum_{i \in [k+r]} |U_{\{i\}}|) \binom{k+r}{2} \left(\frac{k-r+2}{k+r} \right) \\
&= \frac{\binom{k+r}{2}}{\gamma} \left(n - \sum_{i \in [k+r]} |U_{\{i\}}| \right).
\end{aligned}$$

Hence, we have

$$|K| \geq \frac{\binom{k+r}{2}}{\gamma} \left(n - \sum_{i \in [k+r]} |U_{\{i\}}| \right). \tag{9.4}$$

On the other hand, for any fixed S, there are $|\Gamma_S|$ choices for x such that $(x, S) \in K$. Let $S = \{i, j\}$. We have

$$\begin{aligned}
\Gamma_S &= U_S \cap Z \\
&= U_S \backslash W \\
&= U_S \backslash (U_{\{1\}} \cup U_{\{2\}} \cup ... \cup U_{\{k+r\}}) \\
&\subseteq U_S \backslash (U_{\{i\}} \cup U_{\{j\}}).
\end{aligned}$$

Hence, for any x in Γ_S, x is unique under S, but non-unique under $\{i\}$ and $\{j\}$. The

combination of functions f_i and f_j can give up to m^2 different images $(f_i(c), f_j(c))$ for element $c \in X$. However, since $f_i(x)$ and $f_j(x)$ are not unique there are at most

$$(m - |U_{\{i\}}|)(m - |U_{\{j\}}|) = m^2 - m(|U_{\{i\}}| + |U_{\{j\}}|) + |U_{\{i\}}||U_{\{j\}}|$$

possible images $(f_i(x), f_j(x))$ for elements $x \in \Gamma_S$.

Now, we have

$$\begin{aligned}
|K| &= \sum_{S\subseteq[k+r]:|S|=2} |\Gamma_S| \leq \sum_{S\subseteq[k+r]:|S|=2} (m^2 - m(|U_{\{i\}}| + |U_{\{j\}}|) + |U_{\{i\}}||U_{\{j\}}|) \\
&= \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} (m^2 - m(|U_{\{i\}}| + |U_{\{j\}}|) + |U_{\{i\}}||U_{\{j\}}|) \\
&= \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} m^2 - \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} m(|U_{\{i\}}| + |U_{\{j\}}|) + \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| \\
&= \frac{1}{2} 2 \binom{k+r}{2} m^2 - \frac{1}{2} 2(k+r-1)m \sum_{i\in[k+r]} |U_{\{i\}}| + \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| \\
&= \binom{k+r}{2} m^2 - (k+r-1)m \sum_{i\in[k+r]} |U_{\{i\}}| + \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} |U_{\{i\}}||U_{\{j\}}|.
\end{aligned}$$

So,

$$|K| \leq \binom{k+r}{2} m^2 - (k+r-1)m \sum_{i\in[k+r]} |U_{\{i\}}| + \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} |U_{\{i\}}||U_{\{j\}}|. \tag{9.5}$$

From (9.4) and (9.5), we have

$$\begin{aligned}
\frac{\binom{k+r}{2}}{\gamma} \left(n - \sum_{i\in[k+r]} |U_{\{i\}}| \right) \leq & \binom{k+r}{2} m^2 - (k+r-1)m \sum_{i\in[k+r]} |U_{\{i\}}| \\
& + \frac{1}{2} \sum_{\substack{i,j\in[k+r] \\ i\neq j}} |U_{\{i\}}||U_{\{j\}}|.
\end{aligned}$$

Therefore,

$$n \leq \gamma m^2 - \frac{(k+r-1)\gamma m}{\binom{k+r}{2}} \sum_{i\in[k+r]} |U_{\{i\}}| + \frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| + \sum_{i\in[k+r]} |U_{\{i\}}|$$

$$= \gamma m^2 - \left(\left(\frac{(k+r-1)\gamma m}{\binom{k+r}{2}} - 1 \right) \sum_{i\in[k+r]} |U_{\{i\}}| - \frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| \right).$$

We claim that

$$\left(\frac{(k+r-1)\gamma m}{\binom{k+r}{2}} - 1 \right) \sum_{i\in[k+r]} |U_{\{i\}}| - \frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| \geq 0 \tag{9.6}$$

If (9.6) holds, we have $n \leq \gamma m^2 - 0 = \gamma m^2$, which will complete the proof.

Considering each term of (9.6), we have

$$\begin{aligned}
\left(\frac{(k+r-1)\gamma m}{\binom{k+r}{2}} - 1 \right) \sum_{i\in[k+r]} |U_{\{i\}}| &= \left(\frac{(k+r-1)m}{\frac{(k+r)(k+r-1)}{2}} \frac{k+r}{k-r+2} - 1 \right) \sum_{i\in[k+r]} |U_{\{i\}}| \\
&= \left(\frac{2m}{k-r+2} - 1 \right) \sum_{i\in[k+r]} |U_{\{i\}}| \\
&= \frac{2m-(k-r+2)}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}|,
\end{aligned}$$

and since $|U_{\{j\}}| \leq m$ for all $j \in [k+r]$, we have

$$\begin{aligned}
\frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| &\leq \frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}| m \\
&= \frac{\gamma(k+r-1)m}{2\binom{k+r}{2}} \sum_{i\in[k+r]} |U_{\{i\}}| \\
&= \frac{(k+r)(k+r-1)m}{2(k-r+2)\frac{(k+r)(k+r-1)}{2}} \sum_{i\in[k+r]} |U_{\{i\}}| \\
&= \frac{m}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}|.
\end{aligned}$$

Hence

$$\begin{aligned}\left(\frac{(k+r-1)\gamma m}{\binom{k+r}{2}} - 1\right) \sum_{i\in[k+r]} |U_{\{i\}}| - \frac{\gamma}{2\binom{k+r}{2}} \sum_{\substack{i,j\in[k+r]\\ i\neq j}} |U_{\{i\}}||U_{\{j\}}| \\ \geq \frac{2m-(k-r+2)}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}| - \frac{m}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}| \\ = \frac{m-(k-r+2)}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}|.\end{aligned}$$

Since $r \geq 1$, $m \geq k+1$ and $\sum_{i\in[k+r]} |U_{\{i\}}| \geq 0$, we have

$$\begin{aligned}\frac{m-(k-r+2)}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}| &\geq \frac{(k+1)-(k-1+2)}{k-r+2} \sum_{i\in[k+r]} |U_{\{i\}}| \\ &\geq \frac{0}{k-r+2} 0 \\ &= 0.\end{aligned}$$

Hence (9.6) holds and the theorem follows. □

Chapter 10

Secure Frameproof Codes: Separating Hash Families of Type $\{k, k\}$

In this chapter, we focus on improving previously known upper bounds for separating hash families of type $\{k, k\}$; in other words, we improve bounds on the size of secure frameproof codes. The first section is devoted to separating hash families of types $\{2, 2\}$, i.e. 2-SFP codes, of short length (namely, length 4 or less). This is followed by the special case of length 5 for which we reduce the size of upper bound by a factor of $\frac{2}{3}$ compared with the best previously known bound. We then give improved bounds for separating hash families of type $\{k, k\}$ of length ℓ, where $\ell = 2k$. This is followed by the improvement on upper bounds on the size of perfect hash families of type $\{w_1, w_2\}$ of length ℓ, where $(w_1 + w_2) - 1 < \ell \leq 2w_2$. We explore and improve bounds for separating hash families of type $\{k, k\}$ of short length ($\ell \leq k$) in the last section.

10.1 Optimal 2-SFP codes of length 4 or less

Recall the definition of optimal separating hash family from Chapter 8. In this section, we identify optimal separating hash familiies of type $\{2, 2\}$ by studying the bounds on their sizes. Our study focuses on those families of small length, i.e., $N \leq 4$. We omit the case of length 1 since it is trivial.

10.1.1 Length 2

Theorem 10.1.1. *Let m, n be positive integers greater than 1. If there exists an* SHF$(2; n, m, \{2, 2\})$*, then $n \leq m$.*

Proof. Let $\mathcal{F} = \{f_i : X \to Y, i \in \{1, 2\}\}$ be an SHF$(2; n, m, \{2, 2\})$. Assume that $n \geq m + 1$. Then there exist x and y in X such that $x \neq y$ and $f_1(x) = f_1(y)$. Also there exist w and z in X such that $w \neq z$ and $f_2(w) = f_2(z)$. Without loss of generality, assume that $x \neq z$ and $y \neq w$. Let $C_1 = \{x, w\}$, $C_2 = \{y, z\}$. Then both f_1 and f_2 cannot separates images of C_1 and C_2, contradicting the SHF$(2; n, m, \{2, 2\})$ property. Therefore, $n \leq m$. □

It is easy to check that the following matrix gives an SHF$(2; m, m, \{2, 2\})$.

$$\begin{pmatrix} 1 & 2 & \ldots & m \\ 1 & 2 & \ldots & m \end{pmatrix}$$

Hence, the following theorem holds.

Theorem 10.1.2. *Let m be a positive integer greater than 1. An* SHF$(2; m, m, \{2, 2\})$ *exists and is optimal.*

10.1.2 Length 3

The next theorem is deduced from Theorem 8.2.1. Substituting N, w_1, w_2 in Theorem 8.2.1 by 3, 2, 2, respectively, we obtain the following result.

Corollary 10.1.3. *Let m, n be positive integers greater than 1. If there exists an* $\mathrm{SHF}(3; n, m, \{2, 2\})$*, then $n \leq 3m$.*

The results of Bazrafshan and van Trung [6] show that the optimal bounds for the values of m from 2 to 7 are at most $2m$, hence the bound above is not always tight.

10.1.3 Length 4

Theorem 10.1.4. *Let m, n be positive integers greater than 1. If there exists an* $\mathrm{SHF}(4; n, m, \{2, 2\})$*, then $n \leq m^2$.*

Proof. Let $\mathcal{F} = \{f_i : X \to Y, i \in \{1, 2, 3, 4\}\}$ be an $\mathrm{SHF}(4; n, m, \{2, 2\})$. Assume that $n \geq m^2 + 1$. Since there are at most m^2 different ordered pairs $(f_1(a), f_2(a))$ for $a \in X$, there exist x and y in X such that $x \neq y$ and $(f_1(x), f_2(x)) = (f_1(y), f_2(y))$. Likewise, there exist w and z in X such that $w \neq z$ and $(f_3(w), f_4(w)) = (f_3(z), f_4(z))$. Without loss of generality, assume that $x \neq z$ and $y \neq w$. Let $C_1 = \{x, w\}$, $C_2 = \{y, z\}$. Then none of $f_i \in \mathcal{F}$ can separate C_1 and C_2, contradicting the $\mathrm{SHF}(4; n, m, \{2, 2\})$ property. Therefore, $n \leq m^2$. □

We present an $\mathrm{SHF}(4; 9, 3, \{2, 2\})$:

$$\begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 & 3 \\ 1 & 2 & 3 & 2 & 3 & 1 & 3 & 1 & 2 \\ 1 & 2 & 3 & 3 & 1 & 2 & 2 & 3 & 1 \end{pmatrix}$$

It is easily observed that in the above matrix every two columns agree in exactly one position. Hence, the above matrix represents an $\mathrm{SHF}(4; 9, 3, \{2, 2\})$. This shows that the above bound is tight and optimal for $m = 3$.

10.2 2-SFP codes of Length 5

This section contains one of our main results of this chapter: a new upper bound on the size of $\mathrm{SHF}(5; n, m, \{2, 2\})$. We begin by reviewing some previously known upper

bounds on the size of $\mathrm{SHF}(5; n, m, \{2, 2\})$. We then explain our contribution.

The best previously known upper bound can be obtained directly from Theorem 8.2.1 by substituting w_1 and w_2 by 2. When $N = 5$ we have:

Corollary 10.2.1. *Let m, n be positive integers greater than 1. If there exists an $\mathrm{SHF}(5; n, m, \{2, 2\})$, then $n \leq 3m^2$.*

Before showing our improved bounds, we state another useful result by Bazrafshan and van Trung in [7].

Theorem 10.2.2 ([7])**.** *If there exists an $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ with $w_2 \geq 2$, then there exists an $\mathrm{SHF}(N-1; n', m, \{w_1, w_2 - 1\})$ with $n' \geq n - m$.*

The next corollary follows naturally from Theorem 10.2.2 and Theorem 8.2.1 .

Corollary 10.2.3. *Let m, n be positive integers greater than 1. If there exists an $\mathrm{SHF}(5; n, m, \{2, 2\})$, then $n \leq 2m^2 + m$.*

Proof. Let $\mathcal{F}$ be an $\mathrm{SHF}(5; n, m, \{2, 2\})$.
By Theorem 10.2.2, there exists an $\mathrm{SHF}(4; n', m, \{1, 2\})$ with $n' \geq n - m$. By Theorem 8.2.1, we have $n' \leq 2m^2$. This implies $n \leq n' + m \leq 2m^2 + m$. Therefore, $n \leq 2m^2 + m$ as required. □

Hence, we have got an improved upper bound on the size of $\mathrm{SHF}(5; n, m, \{2, 2\})$. Nevertheless, this is not the bound we desire. Our ultimate aim in this section is to improve this bound further to $n \leq 2m^2$.

The following theorem provides a slightly better bound than that given in Corollary 10.2.3. Even though the theorem only brings the upper bound down by 2, the technique using in the proof is a key technique leading us to the main contribution of this section in which we drastically reduce the upper bound on the size of $\mathrm{SHF}(5; n, m, \{2, 2\})$ to $2m^2$. However, the details of the proof are very similar to the proof of the next theorem. Therefore, if the reader is eager to read the main result, we advise to skip reading the proof of this theorem.

Theorem 10.2.4. *Let m, n be positive integers greater than 1. If there exists an* $\mathrm{SHF}(5; n, m, \{2, 2\})$*, then $n \leq 2m^2 + m - 2$.*

Proof. Let $\mathcal{F}$ be an $\mathrm{SHF}(5; n, m, \{2, 2\})$. Assume for a contradiction that $n \geq 2m^2 + m - 1$.

For $r \in [m-1]$, let $P(r)$ be the following statement;

There are r disjoint sets of 4 distinct elements a, b, c, x in X that satisfy all the following 4 conditions:

1. $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.
2. $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.
3. $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.
4. No element p in X other than a, b, c, and x is such that $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$.

We will prove the $P(r)$ holds for all $r \in [m-1]$, by induction on r. We will then show that $P(m-1)$ leads to a contradiction, establishing the theorem.

The base case: $r = 1$. As $n \geq 2m^2 + m - 1$ and $m \geq 2$, there are at least 9 elements in X. Recall the definition of V_I from Definition 9.1. For any $I = \{i, j\} \subseteq [5]$, there are at most m^2 different ordered pairs $u_I = (f_i(u), f_j(u))$ for $u \in X$. So there can be at most $m^2 - 1$ unique elements u_I under $\{i, j\}$. Hence,

$$\begin{aligned} |V_{\{i,j\}}| &\geq (2m^2 + m - 1) - (m^2 - 1) \\ &= m^2 + m. \end{aligned}$$

Therefore,

$$\begin{aligned}|V_{\{1,2\}} \cap V_{\{3,4\}}| &\geq |V_{\{1,2\}}| + |V_{\{3,4\}}| - n \\ &\geq (m^2+m) + (m^2+m) - (2m^2+m-1) \\ &= m+1.\end{aligned}$$

This implies there are at least $m+1$ elements in X that are non-unique under both $\{1,2\}$ and $\{3,4\}$. Then, there exist two distinct elements a and x in $V_{\{1,2\}} \cap V_{\{3,4\}}$ such that $a_5 = x_5$. We will now show that $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.

Assume $a_{\{1,2\}} = x_{\{1,2\}}$. Since $|V_{\{3,4\}}| \geq m^2+m \geq 6$, we can easily find two distinct elements b and y in $V_{\{3,4\}} \backslash \{a, x\}$ such that $b_{\{3,4\}} = y_{\{3,4\}}$. Then $C_1 = \{a, b\}$ and $C_2 = \{x, y\}$ violates the SHF$(5; n, m, \{2, 2\})$ property. A similar argument derives a contradiction from the equality $a_{\{3,4\}} = x_{\{3,4\}}$. Hence, we have shown that there exist two distinct elements a and x in $V_{\{1,2\}} \cap V_{\{3,4\}}$ that $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$, and $a_{\{3,4\}} \neq x_{\{3,4\}}$.

Let $b \in V_{\{1,2\}} \backslash \{a\}$ such that $b_{\{1,2\}} = a_{\{1,2\}}$ and let $y \in V_{\{3,4\}} \backslash \{x\}$ such that $y_{\{3,4\}} = x_{\{3,4\}}$. Consider $C_1 = \{a, y\}, C_2 = \{b, x\}$. From our choice of b and y, we have $a \neq x$, $a \neq b$, $x \neq y$, and

$$\begin{aligned}a_{\{1,2\}} &= b_{\{1,2\}}, \\ y_{\{3,4\}} &= x_{\{3,4\}}, \\ a_5 &= x_5.\end{aligned}$$

which contradicts the SHF$(5; n, m, \{2, 2\})$ property if $b \neq y$. So b and y must be equal. Hence, there exists an element $b \in X \backslash \{a, x\}$ such that $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.

On the other hand, since both a and x are in $V_{\{1,2\}} \cap V_{\{3,4\}}$, there exist $z \in V_{\{1,2\}} \backslash \{x\}$ such that $z_{\{1,2\}} = x_{\{1,2\}}$ and there exists $c \in V_{\{3,4\}} \backslash \{a\}$ such that $c_{\{3,4\}} = a_{\{3,4\}}$. Consider $C_1 = \{a, z\}, C_2 = \{c, x\}$. From our choice of c and z, we have $a \neq x$, $a \neq c$,

$x \neq z$, and

$$z_{\{1,2\}} = x_{\{1,2\}},$$

$$a_{\{3,4\}} = c_{\{3,4\}},$$

$$a_5 = x_5.$$

which contradicts the SHF$(5; n, m, \{2, 2\})$ property if $c \neq z$. So c and z must be equal. Hence, there exists an element $c \in X\backslash\{a, x\}$ such that $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.

By looking at the first two coordinates, we can see that $b \neq c$ since $b_{\{1,2\}} = a_{\{1,2\}} \neq x_{\{1,2\}} = c_{\{1,2\}}$. Now, we have shown that there exist four distinct elements that satisfy the first three conditions of our inductive hypothesis $P(1)$. The next step is to show that the fourth condition holds. Consider the four distinct elements a, b, c, and x: forming a table where the entry indexed by row y and column f_i is $f_i(y)$, we may write

	f_1	f_2	f_3	f_4	f_5
a	a_1	a_2	a_3	a_4	a_5
x	x_1	x_2	x_3	x_4	a_5
b	a_1	a_2	x_3	x_4	$*$
c	x_1	x_2	a_3	a_4	$*$

where $a_1, a_2, a_3, a_4, a_5, x_1, x_2, x_3, x_4 \in Y$ such that $a_1a_2 \neq x_1x_2$, and $a_3a_4 \neq x_3x_4$, and $*$ can be any alphabet symbol in Y.

If there exists an element d, e, f, g or $h \in X\backslash\{a, b, c, x\}$ of the following forms, then we can always form C_1 and C_2 that violates the SHF$(5; n, m, \{2, 2\})$ property:

	f_1	f_2	f_3	f_4	f_5
d	a_1	a_2	$*$	$*$	$*$
e	x_1	x_2	$*$	$*$	$*$
f	$*$	$*$	a_3	a_4	$*$
g	$*$	$*$	x_3	x_4	$*$
h	$*$	$*$	$*$	$*$	a_5

The choice of C_1 and C_2 that violates the SHF$(5; n, m, \{2, 2\})$ property in each case is as follows: choose $C_1 = \{a, b\}, C_2 = \{x, d\}$ when d exists; choose $C_1 = \{a, e\}, C_2 = \{x, c\}$ when e exists; choose $C_1 = \{a, c\}, C_2 = \{x, f\}$ when f exists; choose $C_1 = \{a, g\}, C_2 = \{x, b\}$ when g exists; and choose $C_1 = \{a, x\}, C_2 = \{b, h\}$ when h exists.

Hence, none of the elements p in X other than a, b, c, and x has $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$. Note that there are at most $m^2 \geq 4$ different ordered pairs for $u \in X$ under $\{1, 2\}$ and $\{3, 4\}$, and at most $m \geq 2$ different alphabet symbols under $\{5\}$. Therefore, the fourth condition holds. This implies $P(1)$ is true and the proof is complete in the case $m = 2$.

Inductive step: Assume $m \geq 3$. Let $k \in [m - 2]$. Suppose $P(k)$ holds. Therefore, there are k disjoint sets of 4 distinct elements a, b, c, x in X that satisfy all the following 4 conditions:

1. $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.
2. $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.
3. $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.
4. No element p in X other than a, b, c, and x is such that $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$.

We will show that $P(k + 1)$ holds too.

Remove those $4k$ elements mentioned in $P(k)$ from X to produce the set X_k. Since $|X_k| = n - 4k = (2m^2 + m - 1) - 4k$, and there are at most $m^2 - 2k$ different ordered

pairs for $u \in X_k$ under $\{1, 2\}$ and $\{3, 4\}$, there can be at most $m^2 - 2k - 1$ unique elements in X_k under $\{1, 2\}$ and $\{3, 4\}$. Hence,

$$\begin{aligned}
|X_k| &= (2m^2 + m - 1) - 4k \\
&\geq (2m^2 + m - 1) - 4(m - 2) \\
&= 2m^2 - 3m + 7 \\
&\geq 16, \\
|V_{\{1,2\}} \cap X_k| &\geq (2m^2 + m - 1 - 4k) - (m^2 - 2k - 1) \\
&= m^2 + m - 2k \\
&\geq m^2 + m - 2(m - 2) \\
&= m^2 - m + 4 \\
&\geq 10,
\end{aligned}$$

and

$$\begin{aligned}
|V_{\{3,4\}} \cap X_k| &\geq (2m^2 + m - 1 - 4k) - (m^2 - 2k - 1) \\
&= m^2 + m - 2k \\
&\geq m^2 + m - 2(m - 2) \\
&= m^2 - m + 4 \\
&\geq 10.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_k| &\geq |V_{\{1,2\}} \cap X_k| + |V_{\{3,4\}} \cap X_k| - (n - 4k) \\
&\geq (m^2 + m - 2k) + (m^2 + m - 2k) - (2m^2 + m - 1 - 4k) \\
&= m + 1.
\end{aligned}$$

This implies that there are at least $m+1$ elements in X_k that are non-unique under both $\{1,2\}$ and $\{3,4\}$. Then, there exists two distinct elements a' and x' in $V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_k$ such that $a'_5 = x'_5$.

Just as in the base case, we may argue that $a'_{\{1,2\}} \neq x'_{\{1,2\}}$ and $a'_{\{3,4\}} \neq x'_{\{3,4\}}$. Moreover, we may similarly show that there exists an element $b' \in X_k \backslash \{a', x'\}$ such that $b'_{\{1,2\}} = a'_{\{1,2\}}$ and $b'_{\{3,4\}} = x'_{\{3,4\}}$, and there exists an element $c' \in X_k \backslash \{a', x'\}$ such that $c'_{\{1,2\}} = x'_{\{1,2\}}$ and $c'_{\{3,4\}} = a'_{\{3,4\}}$.

By looking at the first two coordinates, we can see that $b' \neq c'$ since $b'_{\{1,2\}} = a'_{\{1,2\}} \neq x'_{\{1,2\}} = c'_{\{1,2\}}$. Now, we have shown that there exists another four distinct elements a', b', c', d' not contained in the first k sets, that satisfy the first three conditions. The next step is to show that the fourth condition also holds. Consider, the four distinct elements a', b', c', and x' in the following form,

$$\begin{array}{c|ccccc} & f_1 & f_2 & f_3 & f_4 & f_5 \\ a' & a'_1 & a'_2 & a'_3 & a'_4 & a'_5 \\ x' & x'_1 & x'_2 & x'_3 & x'_4 & a'_5 \\ b' & a'_1 & a'_2 & x'_3 & x'_4 & * \\ c' & x'_1 & x'_2 & a'_3 & a'_4 & * \end{array}$$

where $a'_1, a'_2, a'_3, a'_4, a'_5, x'_1, x'_2, x'_3, x'_4 \in Y$ such that $a'_1 a'_2 \neq x'_1 x'_2$, and $a'_3 a'_4 \neq x'_3 x'_4$, and $*$ can be any alphabet symbol in Y.

If there exists an element d, e, f, g or $h \in X \backslash \{a', b', c', x'\}$ of the following forms, then we can always form C_1 and C_2 that violates the SHF$(5; n, m, \{2, 2\})$ property:

$$\begin{array}{c|ccccc} & f_1 & f_2 & f_3 & f_4 & f_5 \\ d & a'_1 & a'_2 & * & * & * \\ e & x'_1 & x'_2 & * & * & * \\ f & * & * & a'_3 & a'_4 & * \\ g & * & * & x'_3 & x'_4 & * \\ h & * & * & * & * & a'_5 \end{array}$$

The choice of C_1 and C_2 that violates the $\mathrm{SHF}(5; n, m, \{2, 2\})$ property in each case is as follows: choose $C_1 = \{a', b'\}, C_2 = \{x', d\}$ when d exists; choose $C_1 = \{a', e\}, C_2 = \{x', c'\}$ when e exists; choose $C_1 = \{a', c'\}, C_2 = \{x', f\}$ when f exists; choose $C_1 = \{a', g\}, C_2 = \{x', b'\}$ when g exists; and choose $C_1 = \{a', x'\}, C_2 = \{b', h\}$ when h exists.

Hence, none of the elements p in X other than a', b', c', and x has $p_{\{1,2\}} = a'_{\{1,2\}}$, $p_{\{3,4\}} = a'_{\{3,4\}}$, $p_{\{1,2\}} = x'_{\{1,2\}}$, $p_{\{3,4\}} = x'_{\{3,4\}}$, or $p_5 = a'_5 = x'_5$. Note that in X_k there are at most $m^2 - 2k \geq m^2 - 2(m-2) \geq 7$ different ordered pairs left for $u \in X_k$ under $\{1, 2\}$ and $\{3, 4\}$, and at most $m - k \geq m - (m-2) = 2$ different alphabet symbols under $\{5\}$ left for $u \in X_k$. Therefore, the fourth condition holds. Hence, we have shown that $P(k+1)$ is also true.

Thus the statement $P(r)$ is true for all $r \in [m-1]$. To be precise, since $P(m-1)$ holds, there exist $m-1$ disjoint sets $\{a, b, c, x\} \subseteq X$ satisfying the four conditions of the inductive hypothesis. We now derive our contradiction. We eliminate two elements a and x from each of these sets, a total of $2m-2$ elements from X to produce the set X'. This results in prohibiting $m-1$ alphabet symbols from the image of f_5. Hence, the $n-(2m-2)$ elements of X' share the same alphabet symbol in the last coordinate. Then

$$\begin{aligned} |V_{\{1,2\}} \cap X'| &\geq |X'| - (m^2 - 1) \\ &= (n - (2m-2)) - (m^2 - 1) \\ &= (2m^2 + m - 1 - (2m-2)) - (m^2 - 1) \\ &= m^2 - m + 4 \\ &\geq 6, \end{aligned}$$

and similarly $|V_{\{3,4\}} \cap X'| \geq 6$. Hence there exist two distinct elements p and q in $V_{\{1,2\}} \cap X'$ such that $p_{\{1,2\}} = q_{\{1,2\}}$, and two distinct elements s and t in $V_{\{3,4\}} \cap X' \setminus \{p, q\}$

such that $s_{\{3,4\}} = t_{\{3,4\}}$. Now

$$p_{\{1,2\}} = q_{\{1,2\}},$$
$$s_{\{3,4\}} = t_{\{3,4\}},$$
$$p_5 = q_5 = s_5 = t_5.$$

Since p, q, s and t are pairwise distinct elements, $C_1 = \{p, s\}$ and $C_2 = \{q, t\}$ contradict the SHF$(5; n, m, \{2, 2\})$ property.

Therefore, the assumption $n \geq 2m^2 + m - 1$ is false and so the theorem follows. □

We state and prove the next theorem as a tool for our final result, which we reduce the upper bounds of SHF$(5; n, m, \{2, 2\})$ further by $m - 2$.

The proof of the following theorem is very similar to the proof of the previous theorem. The only difference is that we first remove an element of X with a certain property before proceeding with the same technique to derive a contradiction: proving an existence of $m - 1$ disjoint sets of 4 distinct elements a, b, c, x in X_0 satisfying our conditions.

Theorem 10.2.5. *Let m, n be positive integers greater than 1.*
Let $\mathcal{F}$ be an SHF$(5; n, m, \{2, 2\})$. *If $n \geq 2m^2$, then $|f_i^{-1}(f_i(x))| > 1$ for all $f_i \in \mathcal{F}$ and all $x \in X$.*

Proof. Let $\mathcal{F}$ be an SHF$(5; n, m, \{2, 2\})$ with $n = 2m^2 + n_0$, where n_0 is a non-negative integer. Assume for a contradiction that there exists an element v in X and $i \in [N]$ such that v is unique under $\{i\}$. Without loss of generality, fix $i = 1$. Define $X_0 = X \backslash \{v\}$.

We now proceed with the similar arguments as in the proof of Theorem 10.2.4. The only difference is that the size of $V_{\{1,2\}}$ in X_0 is smaller due to the removal of the element v.

We firstly use induction to show that there are $m - 1$ disjoint sets of 4 distinct elements a, b, c, x in X_0 that satisfy all the following 4 conditions:

1. $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.
2. $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.
3. $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.
4. No element $p \in X_0$ other than a, b, c, and x is such that $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$.

For $r \in [m-1]$, let $P(r)$ be the following statement: There are r disjoint sets of 4 distinct elements a, b, c, x in X_0 that satisfy all the following 4 conditions:

1. $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.
2. $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.
3. $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.
4. No element p in X_0 other than a, b, c, and x is such that $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$.

We will prove that $P(r)$ holds for all $r \in [m-1]$, by induction on r. We will then show that $P(m-1)$ leads to a contradiction, establishing the theorem.

The base case: $r = 1$. As $n = 2m^2 + n_0$, $n_0 \geq 0$ and $m \geq 2$, we have $|X_0| \geq 7$. Recall the definition of V_I from Definition 9.1. Considering X_0, there are at most $m(m-1) = m^2 - m$ different ordered pairs $(f_1(u), f_2(u))$ for $u \in X_0$ under $\{1, 2\}$ and at most m^2 different ordered pairs $(f_3(u), f_4(u))$ for $u \in X_0$ under $\{3, 4\}$. This implies there can be at most $m^2 - m - 1$ and $m^2 - 1$ unique elements under $\{1, 2\}$ and $\{3, 4\}$, respectively. Hence,

$$\begin{aligned} |V_{\{1,2\}} \cap X_0| &\geq (2m^2 + n_0 - 1) - (m^2 - m - 1) \\ &= m^2 + m + n_0 \\ &\geq 6, \end{aligned}$$

and

$$\begin{aligned} |V_{\{3,4\}} \cap X_0| &\geq (2m^2 + n_0 - 1) - (m^2 - 1) \\ &= m^2 + n_0 \\ &\geq 4. \end{aligned}$$

Since $|X_0| = n - 1$, $V_{\{1,2\}} \cap X_0$ and $V_{\{3,4\}} \cap X_0$ can share at most $n - 1$ elements in common. Therefore,

$$\begin{aligned} |V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_0| &\geq |V_{\{1,2\}} \cap X_0| + |V_{\{3,4\}} \cap X_0| - (n-1) \\ &\geq (m^2 + m + n_0) + (m^2 + n_0) - (2m^2 + n_0 - 1) \\ &= m + n_0 + 1 \\ &\geq m + 1. \end{aligned}$$

That implies there are at least $m + 1$ elements in X_0 that are non-unique under both $\{1, 2\}$ and $\{3, 4\}$. So there exist two distinct elements $a, x \in (V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_0)$ such that $a_5 = x_5$. We will now show that $a_{\{1,2\}} \neq x_{\{1,2\}}$ and $a_{\{3,4\}} \neq x_{\{3,4\}}$.

Assume for a contradiction, that $a_{\{1,2\}} = x_{\{1,2\}}$. Since $|V_{\{3,4\}} \cap X_0| \geq 4$, we can easily find two distinct elements $b, y \in ((V_{\{3,4\}} \backslash \{a, x\}) \cap X_0)$ such that $b_{\{3,4\}} = y_{\{3,4\}}$. Then $C_1 = \{a, b\}$ and $C_2 = \{x, y\}$ violates the SHF$(5; n, m, \{2, 2\})$ property. Thus $a_{\{1,2\}} \neq x_{\{1,2\}}$. The same argument works for the case when $a_{\{3,4\}} = x_{\{3,4\}}$ since $|V_{\{1,2\}} \cap X_0| \geq 6$. Hence, it is justified to say there exist two distinct elements a and x in $V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_0$ that $a_5 = x_5$, $a_{\{1,2\}} \neq x_{\{1,2\}}$, and $a_{\{3,4\}} \neq x_{\{3,4\}}$.

Let $b \in (V_{\{1,2\}} \backslash \{a\}) \cap X_0$ such that $b_{\{1,2\}} = a_{\{1,2\}}$ and let $y \in (V_{\{3,4\}} \backslash \{x\}) \cap X_0$ such that $y_{\{3,4\}} = x_{\{3,4\}}$. Consider $C_1 = \{a, y\}, C_2 = \{b, x\}$. From the choices of b and

y, we have $a \neq x$, $a \neq b$, $x \neq y$, and

$$a_{\{1,2\}} = b_{\{1,2\}},$$
$$y_{\{3,4\}} = x_{\{3,4\}},$$
$$a_5 = x_5,$$

which contradicts the SHF$(5; n, m, \{2, 2\})$ property if $b \neq y$. So b and y must be the same element. Hence, there exists an element $b \in X_0 \backslash \{a, x\}$ such that $b_{\{1,2\}} = a_{\{1,2\}}$ and $b_{\{3,4\}} = x_{\{3,4\}}$.

On the other hand, since both a and x are in $V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_0$, there exist $z \in (V_{\{1,2\}} \backslash \{x\}) \cap X_0$ such that $z_{\{1,2\}} = x_{\{1,2\}}$. Let $c \in (V_{\{3,4\}} \backslash \{a\}) \cap X_0$ be such that $c_{\{3,4\}} = a_{\{3,4\}}$. Consider $C_1 = \{a, z\}$ and $C_2 = \{c, x\}$. From our choice of c and z, we have $a \neq x$, $a \neq c$, $x \neq z$, and

$$z_{\{1,2\}} = x_{\{1,2\}},$$
$$a_{\{3,4\}} = c_{\{3,4\}},$$
$$a_5 = x_5,$$

which contradicts the SHF$(5; n, m, \{2, 2\})$ property if $c \neq z$. So c and z must be equal. Hence, there exists an elements $c \in X_0 \backslash \{a, x\}$ such that $c_{\{1,2\}} = x_{\{1,2\}}$ and $c_{\{3,4\}} = a_{\{3,4\}}$.

By looking at the first two coordinates, we can see that $b \neq c$ since $b_{\{1,2\}} = a_{\{1,2\}} \neq x_{\{1,2\}} = c_{\{1,2\}}$. Now, we have shown that there exist four distinct elements that satisfy the first three conditions of our inductive hypothesis $P(1)$. The next step is to show that the fourth condition holds. Consider the four distinct elements a, b, c, and x: forming a table where the entry indexed by row u and column f_i is $f_i(u)$, we may write:

$$\begin{array}{cccccc} & f_1 & f_2 & f_3 & f_4 & f_5 \\ a & a_1 & a_2 & a_3 & a_4 & a_5 \\ x & x_1 & x_2 & x_3 & x_4 & a_5 \\ b & a_1 & a_2 & x_3 & x_4 & * \\ c & x_1 & x_2 & a_3 & a_4 & * \end{array}$$

where $a_1, a_2, a_3, a_4, a_5, x_1, x_2, x_3, x_4 \in Y$ such that $a_1a_2 \neq x_1x_2$, and $a_3a_4 \neq x_3x_4$, and $*$ can be any alphabet symbol in Y. As before

If elements of any of the following forms happen to be in $X_0\backslash\{a, b, c, x\}$, then we can always form C_1 and C_2 that violates the SHF$(5; n, m, \{2, 2\})$ property:

$$\begin{array}{cccccc} & f_1 & f_2 & f_3 & f_4 & f_5 \\ d & a_1 & a_2 & * & * & * \\ e & x_1 & x_2 & * & * & * \\ f & * & * & a_3 & a_4 & * \\ g & * & * & x_3 & x_4 & * \\ h & * & * & * & * & a_5 \end{array}$$

The choice of C_1 and C_2 that violates the SHF$(5; n, m, \{2, 2\})$ property are as follows: choose $C_1 = \{a, b\}, C_2 = \{x, d\}$ if d exists; choose $C_1 = \{a, e\}, C_2 = \{x, c\}$ if e exists; choose $C_1 = \{a, c\}, C_2 = \{x, f\}$ if f exists; choose $C_1 = \{a, g\}, C_2 = \{x, b\}$ if g exists; and choose $C_1 = \{a, x\}, C_2 = \{b, h\}$ if h exists.

Hence, if $p \in X_0\backslash\{a, b, c, x\}$ then none of the following hold: $p_{\{1,2\}} = a_{\{1,2\}}$, $p_{\{3,4\}} = a_{\{3,4\}}$, $p_{\{1,2\}} = x_{\{1,2\}}$, $p_{\{3,4\}} = x_{\{3,4\}}$, or $p_5 = a_5 = x_5$. Therefore, the fourth condition holds. This implies $P(1)$ is true and the proof is complete for the case $m = 2$.

Inductive hypothesis: Assume, $m \geq 3$. Let $k \in [m - 2]$. Suppose $P(k)$ holds. Therefore, there are k disjoint sets of 4 distinct elements a, b, c, x in X that satisfy all the 4 conditions. We will show that $P(k + 1)$ holds too.

Remove those $4k$ elements from X_0 and denote the set of the rest of the elements by X_k. Now $|X_k| = n - 1 - 4k = (2m^2 + n_0 - 1) - 4k$; there are at most $m(m - 1) - 2k = m^2 - m - 2k$ different ordered pairs for $u \in X_k$ under $\{1, 2\}$ and at most $m^2 - 2k$

different ordered pairs for $u \in X_k$ under $\{3, 4\}$. This implies there can be at most $m^2 - m - 2k - 1$ and $m^2 - 2k - 1$ unique elements $u \in X_k$ under $\{1, 2\}$ and $\{3, 4\}$, respectively. Hence,

$$\begin{aligned}
|X_k| &= (2m^2 + n_0 - 1) - 4k \\
&\geq (2m^2 + n_0 - 1) - 4(m - 2) \\
&= 2m^2 - 4m + n_0 + 7 \\
&\geq 13, \\
|V_{\{1,2\}} \cap X_k| &\geq (2m^2 + n_0 - 1 - 4k) - (m^2 - m - 2k - 1) \\
&= m^2 + m + n_0 - 2k \\
&\geq m^2 + m + n_0 - 2(m - 2) \\
&= m^2 - m + n_0 + 4 \\
&\geq 10,
\end{aligned}$$

and

$$\begin{aligned}
|V_{\{3,4\}} \cap X_k| &\geq (2m^2 + n_0 - 1 - 4k) - (m^2 - 2k - 1) \\
&= m^2 + n_0 - 2k \\
&\geq m^2 + n_0 - 2(m - 2) \\
&= m^2 - 2m + n_0 + 4 \\
&\geq 7.
\end{aligned}$$

Therefore,

$$\begin{aligned}|V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_k| &\geq |V_{\{1,2\}} \cap X_k| + |V_{\{3,4\}} \cap X_k| - (n-1-4k) \\ &\geq (m^2+m+n_0-2k) + (m^2+n_0-2k) - (2m^2+n_0-1-4k) \\ &= m+n_0+1 \\ &\geq m+1.\end{aligned}$$

That implies there are at least $m+1$ elements in X_k that are non-unique under both $\{1,2\}$ and $\{3,4\}$. Then, there exists two distinct elements $a', x' \in (V_{\{1,2\}} \cap V_{\{3,4\}} \cap X_k)$ that $a'_5 = x'_5$.

Just as in the base case, we may argue that $a'_{\{1,2\}} \neq x'_{\{1,2\}}$ and $a'_{\{3,4\}} \neq x'_{\{3,4\}}$. Moreover, we may similarly show that there exists an element $b' \in X_k \backslash \{a', x'\}$ such that $b'_{\{1,2\}} = a'_{\{1,2\}}$ and $b'_{\{3,4\}} = x'_{\{3,4\}}$, and there exists an element $c' \in X_k \backslash \{a', x'\}$ such that $c'_{\{1,2\}} = x'_{\{1,2\}}$ and $c'_{\{3,4\}} = a'_{\{3,4\}}$.

By looking at the first two coordinates, we can see that $b' \neq c'$ since $b'_{\{1,2\}} = a'_{\{1,2\}} \neq x'_{\{1,2\}} = c'_{\{1,2\}}$. Hence, we have shown that there exists another four distinct elements, apart from the first k sets, that satisfy the first three conditions.

With a similar argument as in the base case, we may also argue that the fourth condition also holds, and thus, we have shown that $P(k+1)$ is also true.

Therefore the statement $P(r)$ is true for all $r \in [m-1]$. To be precise, there exist $m-1$ disjoint sets satisfying those 4 conditions.

Eliminate two elements a and x from each of our $m-1$ disjoint sets $\{a, b, c, x\} \in X_0$, which means $2m-2$ elements from X_0 to produce a set X'. This results in prohibiting $m-1$ alphabet symbols from the image $f_5(X')$ of f_5. Hence, the remaining $n-(2m-2)$

elements in X' share the same alphabet symbol in the last coordinate. Now,

$$\begin{aligned}
|V_{\{1,2\}} \cap X'| &\geq |X'| - (m^2 - m - 1) \\
&= (n - 1 - (2m - 2)) - (m^2 - m - 1) \\
&= (2m^2 + n_0 - 1 - (2m - 2)) - (m^2 - m - 1) \\
&= m^2 - m + n_0 + 4 \\
&\geq 6,
\end{aligned}$$

and

$$\begin{aligned}
|V_{\{3,4\}} \cap X'| &\geq |X'| - (m^2 - 1) \\
&= (n - 1 - (2m - 2)) - (m^2 - 1) \\
&= (2m^2 + n_0 - 1 - (2m - 2)) - (m^2 - 1) \\
&= m^2 - 2m + n_0 + 4 \\
&\geq 7.
\end{aligned}$$

Hence there exist two distinct elements p and q in $V_{\{1,2\}} \cap X'$ such that $p_{\{1,2\}} = q_{\{1,2\}}$, and two distinct elements s and t in $V_{\{3,4\}} \cap X' \backslash \{p, q\}$ such that $s_{\{3,4\}} = t_{\{3,4\}}$. Since

$$\begin{aligned}
p_{\{1,2\}} &= q_{\{1,2\}}, \\
s_{\{3,4\}} &= t_{\{3,4\}}, \\
p_5 = q_5 &= s_5 = t_5,
\end{aligned}$$

and since p, q, s and t are pairwise distinct , $C_1 = \{p, s\}$ and $C_2 = \{q, t\}$ contradict the $\mathrm{SHF}(5; n, m, \{2, 2\})$ property.

This contradiction show that an element $v \in X$ which is unique under $\{i\}$ does not exist. Therefore, if $n \geq 2m^2$, then $|f_i^{-1}(f_i(x))| > 1$ for all $f_i \in \mathcal{F}$, and all $x \in X$, as required. □

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

Theorem 10.2.6. *Let m, n be positive integers greater than 1. If there exists an* $\mathrm{SHF}(5; n, m, \{2, 2\})$*, then $n \leq 2m^2$.*

There exist constructions in various literatures that give $\mathrm{SHF}(5; m^2, m, \{2, 2\})$. However, there are no known constructions provide $\mathrm{SHF}(5; 2m^2, m, \{2, 2\})$. Hence, it still cannot be concluded how good the bounds above are.

Proof of Theorem 10.2.6. Let $\mathcal{F}$ be an $\mathrm{SHF}(5; n, m, \{2, 2\})$. Assume for a contradiction that $n = 2m^2 + n_0$, where n_0 is a positive integer.

Recall the definition of V_I and U_I from Definition 9.1. For any $I \subseteq [5]$, let $D_I = \{x \in X : |\{z \in X : z_I = x_I\}| = 2\}$, and let $T_I = \{x \in X : |\{z \in X : z_I = x_I\}| \geq 3\}$. Hence $D_I \cup T_I = V_I$.

Let $I_{\max}$ be a subset of $[5]$ of cardinality 2 that maximises the size of U_I. Let $j \in [5] \backslash I_{\max}$, and let $J = [5] \backslash (I_{\max} \cup \{j\})$. Since $n > 2m^2$, by the pigeonhole principle $T_{I_{\max}} \neq \emptyset$.

We first show that $|U_J| \geq |T_{I_{\max}}|$ by proving the existence of an injective function from $T_{I_{\max}}$ to U_J. Then we reach the contradiction by considering the size of $U_{I_{\max}}, D_{I_{\max}}$ and $T_{I_{\max}}$.

Claim. $|U_J| \geq |T_{I_{\max}}|$

First, we will show that for any $a \in X$, if there exists an $x \in T_{I_{\max}} \backslash \{a\}$ such that $x_{\{j\}} = a_{\{j\}}$, then $a \in U_J$.

To show this, let $a \in X$ and let $x \in T_{I_{\max}} \backslash \{a\}$ such that $x_{\{j\}} = a_{\{j\}}$. Assume for contradiction that $a \notin U_J$. Let y and z be two distinct elements in $T_{I_{\max}} \backslash \{x\}$ such that $y_{I_{\max}} = z_{I_{\max}} = x_{I_{\max}}$. Assume that $a \notin U_J$. Then, there exists $b \in X \backslash \{a\}$ such that $b_J = a_J$. Since y and z are two distinct elements at least one of them is not equal to b, say y.

Choose $C_1 = \{x, b\}$ and $C_2 = \{y, a\}$, we have $x \neq y$, $x \neq a$, $b \neq y$ and $b \neq a$. So

$C_1 \cap C_2 = \emptyset$ and

$$x_{I_{\max}} = y_{I_{\max}},$$
$$b_J = a_J,$$
$$x_{\{j\}} = a_{\{j\}}.$$

which contradicts the SHF$(5; n, m, \{2, 2\})$ property. Thus $a \in U_J$

In order to prove our claim, define a mapping $\varphi : T_{I_{\max}} \to X$ as follows:

$$\varphi(x) = \begin{cases} x & \text{if } x_{\{j\}} = y_{\{j\}} \text{ for some } y \in T_{I_{\max}} \backslash \{x\} \\ v & \text{if } x_{\{j\}} \neq y_{\{j\}} \text{ for all } y \in T_{I_{\max}} \backslash \{x\} \text{ and } x_{\{j\}} = v_{\{j\}} \text{ for some } v \in X \backslash T_{I_{\max}} \end{cases},$$

for any $x \in T_{I_{\max}}$. It follows from Theorem 10.2.5 that if $x_{\{j\}} \neq y_{\{j\}}$ for all $y \in T_{I_{\max}} \backslash \{x\}$ there always exists at least one $v \in X \backslash T_{I_{\max}}$ that $v_{\{j\}} = x_{\{j\}}$. If there are more than one such v, we can just pick one. It is not difficult to see that φ is also injective.

Now we will show that $\varphi(T_{I_{\max}}) \subseteq U_J$. For any $y \in \varphi(T_{I_{\max}})$, there exists an element $x \in T_{I_{\max}}$ that $x_{\{j\}} = y_{\{j\}}$, which makes $y \in U_J$ by the first sentence after the statement of the claim. Therefore φ is an injective function from $T_{I_{\max}}$ to U_J, and that makes $|U_J| \geq |T_{I_{\max}}|$. So our claim follows.

Next, we derive our contradiction by considering the size of $U_{I_{\max}}, D_{I_{\max}}$ and $T_{I_{\max}}$. Let $|U_{I_{\max}}| = \lambda$. There are only $m^2 - \lambda$ different ordered pairs $f_{I_{\max}}(x)$ left for elements x in $V_{I_{\max}}$. Since, we have to reserve at least one ordered pair $f_{I_{\max}}(x)$ for elements x in $T_{I_{\max}}$, there are at most $m^2 - \lambda - 1$ different ordered pairs $f_{I_{\max}}(x)$ remaining for elements x in $D_{I_{\max}}$. This makes

$$|D_{I_{\max}}| \leq 2(m^2 - \lambda - 1).$$

Therefore

$$\begin{aligned} |T_{I_{\max}}| &= n - |U_{I_{\max}}| - |D_{I_{\max}}| \\ &\geq 2m^2 + n_0 - \lambda - 2(m^2 - \lambda - 1) \\ &= n_0 + \lambda + 2. \end{aligned}$$

From $|U_J| \geq |T_{I_{\max}}|$ and the maximality of $I_{\max}$, we have

$$\begin{aligned} \lambda &= |U_{I_{\max}}| \\ &\geq |U_J| \\ &\geq |T_{I_{\max}}| \\ &\geq n_0 + \lambda + 2. \end{aligned}$$

which makes $n_0 \leq -2$, contradicting our initial assumption. Therefore, $n < 2m^2$ as required. □

10.3 k-SFP of Length $2k$

This section contains our improved bounds on the size of separating hash families of type $\{k, k\}$ of length N when $N = 2k$, Theorem 10.3.3, which is equivalent to proving Theorem 10.3.4 in secure frameproof codes language. This is then followed by our improved bounds on the size of separating hash families of type $\{w_1, w_2\}$ of length N when $(w_1 + w_2) - 1 < N \leq 2w_2$, Theorem 10.3.5.

First we show:

Theorem 10.3.1. *Let m, n be positive integers, and let w_1, w_2 be positive integers such that $w_1 \leq w_2$ and $w_1 + w_2 < m$. If there exists an* $\mathrm{SHF}(N + 2; n, m, \{w_1 + 1, w_2 + 1\})$ *where $n \geq m^2$, then there exists an* $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$.

Proof. Let $\mathcal{F} = \{f_1, f_2, ..., f_{N+2} : X \to Y\}$ be an $\mathrm{SHF}(N + 2; n, m, \{w_1 + 1, w_2 + 1\})$.

Assume for a contradiction that there is no SHF$(N; n, m, \{w_1, w_2\})$.

Let $\mathcal{F}' = \mathcal{F}\backslash\{f_1, f_2\}$. We have that $|\mathcal{F}'| = N$. Let C_1, C_2 be two disjoint subsets of X such that $|C_1| \leq w_1$ and $|C_2| \leq w_2$. By our assumption, there is no SHF$(N; n, m, \{w_1, w_2\})$, hence none of the functions $f \in \mathcal{F}'$ can separate C_1 and C_2.

Claim. There exist two distinct elements $x \in X\backslash C_2$ and $y \in X\backslash(C_1 \cup \{x\})$, such that $C_1 \cup \{x\}$ and $C_2 \cup \{y\}$ cannot be separated by f_1 and f_2.

Consider the two following statements:

1. There exists an element $x \in X\backslash C_2$ such that $f_1(x) \in f_1(C_2)$. So f_1 cannot separate $C_1 \cup \{x\}$ and C_2.

2. There exists an element $y \in X\backslash(C_1 \cup \{x\})$ such that $f_2(y) \in f_2(C_1)$. So f_2 cannot separate C_1 and $C_2 \cup \{y\}$. .

If both of the statements above are true, our claim follows.

If the first statement is not true, any symbol in $f_1(C_2)$ is not the image of an element outside C_2 under f_1. Therefore, under f_1, there are at most $m - 1$ symbols left for $n - |C_2|$ elements in $X\backslash(C_1 \cup C_2)$. So, there are at most $m(m-1)$ distinct ordered ordered pairs $(f_1(c), f_2(c))$ for the $n - |C_1| - |C_2|$ elements c in $X\backslash(C_1 \cup C_2)$. Since $\left\lceil \frac{n-|C_1|-|C_2|}{m(m-1)} \right\rceil \geq \left\lceil \frac{n-(w_1+w_2)}{m^2-m} \right\rceil \geq \left\lceil \frac{n-(m-1)}{m^2-m} \right\rceil \geq \left\lceil \frac{m^2-m+1}{m^2-m} \right\rceil \geq 2$, by the pigeonhole principle, there are at least 2 elements x and y such that $(f_1(x), f_2(x)) = (f_1(y), f_2(y))$. Let $C_1' = C_1 \cup \{x\}$ and $C_2' = C_2 \cup \{y\}$, then C_1' and C_2' that cannot be separated by f_1 and f_2. Hence we have justified the claim when statement 1 is false. Similarly, the claim holds when statement 2 is false.

We now have C_1' and C_2' are disjoint, $|C_1'| \leq w_1 + 1$ and $|C_2'| \leq w_2 + 1$. Since C_1' and C_2' cannot be separated by f_1 and f_2, we have C_1', C_2' cannot be separated by any function $f \in \mathcal{F}$, contradicting the SHF$(N + 2; n, m, \{w_1 + 1, w_2 + 1\})$ property. Hence, there exists an SHF$(N; n, m, \{w_1, w_2\})$. □

Theorem 10.3.1 can be generalised as follows:

Theorem 10.3.2. *Let m, n be positive integers, and let w_1, w_2 be positive integers such that $w_1 \leq w_2$ and $w_1 + w_2 < m$. If there exists an* $\mathrm{SHF}(N + 2s; n, m, \{w_1 + s, w_2 + s\})$ *where $n \geq m^2$, then there exists an* $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$.

Proof. The statement follows by induction on s, with Theorem 10.3.1 providing the inductive step.

□

By substituting N, w_1, w_2 and s in Theorem 10.3.2 by $2, 1, 1$ and $k-1$, respectively, we obtain the following theorem. Note that this theorem also includes Theorem 10.1.4 as a special case when $k = 2$.

Theorem 10.3.3. *Let m, n, k be positive integers greater than 1, where $m > k$.*
If $\mathcal{F}$ is an $\mathrm{SHF}(2k; n, m, \{k, k\})$. *Then*

$$n \leq m^2.$$

Proof. Assume that there exists an $\mathrm{SHF}(2k; n, m, \{k, k\})$ where $n > m^2$. By Theorem 10.3.2, there exists an $\mathrm{SHF}(2; n, m, \{1, 1\})$, which contradicts Theorem 8.2.5. Hence the theorem follows. □

Theorem 10.3.3 can be written in secure frameproof codes language as follows:

Theorem 10.3.4. *Let C be an m-ary k-SFP code of length ℓ where $\ell = 2k$ and $m > k$. Then*

$$n \leq m^2$$

Theorem 10.3.2 can be used to improve the bounds for $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ when $(w_1 + w_2) - 1 < N \leq 2w_2$. The result is as follows:

Theorem 10.3.5. *Let m, n, N be positive integers greater than 1. Let w_1, w_2, u be positive integers such that $1 \leq w_1 \leq w_2$ and $u = w_1 + w_2$.*

If there exists an $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ *where* $(u-1) < N \leq 2w_2$ *and* $m > (w_2 - w_1) + 1$. *Then*

$$n \leq \gamma m^2,$$

where $\gamma = \frac{N-2(w_1-1)}{N-2(w_1-1)-2r+2}$ *and* $r = N - (u-1)$.

Before giving the proof of Theorem 10.3.5, we here show that the bound in Theorem 10.3.5 is better than any previously known bounds (see Theorem 8.2.1). When $w_1 = 1$, Theorem 10.3.5 gives the same bound as Theorem 9.3.1. Hence Theorem 10.3.5 is a generalised version of Theorems 9.3.1 and 10.3.3. Now, $r = N - (u-1) \geq 1$, so $N - 2(w_1 - 1) - 2r + 2 \geq N - 2(w_1 - 1)$. Hence

$$\gamma \geq 1 \tag{10.1}$$

and $\gamma = 1$ only when $r = 1$, i.e, $\gamma = 1$ only when $N = u$. Note that $r = N - (u-1) \leq 2w_2 - (w_1 + w_2 - 1) = w_2 - w_1 + 1 \leq w_2 + w_1 - 1 = u - 1$. Hence $N = r + (u-1) \leq 2(u-1)$, and it becomes equality only when $r = u - 1$. We have that

$$\begin{aligned}
\gamma &= \frac{N - 2(w_1 - 1)}{N - 2(w_1 - 1) - 2r + 2} \\
&= \frac{((u-1) + r) - 2(w_1 - 1)}{((u-1) + r) - 2(w_1 - 1) - 2r + 2} \\
&\leq \frac{(u-1) + r}{((u-1) + r) - 2r + 2} \text{ since } 2(w_1 - 1) \geq 0 \\
&\leq \frac{2(u-1)}{(u-1) - r + 2} \text{ since } r \leq u - 1 \\
&\leq \frac{2(u-1)}{2} \text{ since } r \leq u - 1 \\
&= u - 1.
\end{aligned}$$

Therefore $\gamma \leq u - 1$ with equality only when $2(w_1 - 1) = 0$ and $r = u - 1$, in other words, when $w_1 = 1$ and $N = 2(u-1) = 2(w_1 + w_2 - 1) = 2w_2$. Hence, our leading

term is always at least as good as any previously known bounds.

Proof of Theorem 10.3.5. Let $\mathcal{F}$ be an SHF$(N; n, m, \{w_1, w_2\})$ where $(u-1) < N \leq 2w_2$. By (10.1), $\gamma m^2 \geq m^2$. By substituting s in Theorem 10.3.2 with $w_1 - 1$, there exists an SHF$(N'; n, m, \{1, k\})$ where $N' = N - (w_1 - 1)$ and $k = w_2 - (w_1 - 1)$.

Let $r = N' - k = (N - 2(w_1 - 1)) - (w_2 - (w_1 - 1) = N - (u - 1)$. Now $N - (u-1) > (u-1) - (u-1) = 0$ and $N - (u-1) < 2w_2 - (u-1) = w_2 - (w_1 - 1) = k$. So we have $0 < r \leq k$. Therefore

$$n \leq \gamma m^2,$$

where $\gamma = \frac{k+r}{k-r+2} = \frac{N'}{N'-2r+2} = \frac{N-2(w_1-1)}{N-2(w_1-1)-2r+2}$. □

10.4 k-SFP of Short Length

This section gives our improved bounds on the size of separating hash families of type $\{k, k\}$ of length ℓ, when $\ell \leq k$. We start from stating the best previously known bound; we then give an improved bounds when $\ell \leq k$, and a further improvement when $\ell = k = 3$, $\ell = k = 4$ and $\ell = k = 5$.

Recall Theorem 3.2.3 which can be obtained by substituting w_1 and w_2 in Theorem 8.2.1 by k. When $\ell \leq k$ we have:

Corollary 10.4.1. *Let k, m, n and ℓ be positive integers greater than 1, where $\ell \leq k$. If there exists an* SHF$(\ell; n, m, \{k, k\})$*, then $n \leq (2k-1)m$.*

The next theorem is our improved bound when $\ell \leq k$.

Theorem 10.4.2. *Let k, m, n and ℓ be positive integers where $\ell \leq k$. If there exists an* SHF$(\ell; n, m, \{k, k\})$*, then $n \leq m + 2\ell - 3$.*

Which can be rephrased as:

Theorem 10.4.3. *Let C be an m-ary k-SFP code of length ℓ, where $\ell \leq k$. Then*

$$n \leq m + 2\ell - 3$$

Proof. Let $\mathcal{F}$ be an SHF$(\ell; n, m, \{k, k\})$. Assume for a contradiction that $n \geq m+2\ell-2$. We try to show that there exist $C_1, C_2 \subset X$ of cardinality at most k such that $C_1 \cap C_2 = \emptyset$ but none of f in $\mathcal{F}$ can separate C_1 and C_2.

We firstly use induction to show that there are $2\ell - 2$ distinct elements x_1, y_1, x_2, y_2, ... , $x_{\ell-1}$, $y_{\ell-1}$ such that $f_i(x_i) = f_i(y_i)$ for all $i \in [\ell - 1]$.

For $r \in [\ell - 1]$, let $P(r)$ be the following statement: There are $2r$ distinct elements x_1, y_1, x_2, y_2, ... , x_r, $y_r \in X$ such that $f_i(x_i) = f_i(y_i)$ for all $i \in [r]$.

The base case: $r = 1$. Since $n \geq m + 2\ell - 2 > m$ there exist x_1 and y_1 in X such that $f_1(x_1) = f_1(y_1)$. This implies $P(1)$ is true and the induction proof is complete for the case $\ell = k = 2$.

Inductive hypothesis: Assume, $\ell \geq 3$. Let $c \in [\ell-2]$. Suppose $P(c)$ holds. Therefore, there are $2c$ distinct elements x_1, y_1, x_2, y_2, ... , x_c, y_c such that $f_i(x_i) = f_i(y_i)$ for all $i \in [c]$. We will show that $P(c + 1)$ holds too.

Let $X_0 = X \backslash \{y_1, x_2, y_2, ..., x_c, y_c\}$ Now $|X_0| = n-2c \geq n-(2(\ell-2)) = n-(2\ell-4) \geq (m + 2\ell - 2) - (2\ell - 4) = m + 2 > m$. So, there exist x_{c+1} and y_{c+1} in X_0 such that $f_{c+1}(x_{c+1}) = f_{c+1}(y_{c+1})$. Therefore, there are $2c + 2$ distinct elements x_1, y_1, x_2, y_2, ... , x_{c+1}, y_{c+1} such that $f_i(x_i) = f_i(y_i)$ for all $i \in [c + 1]$. Hence, we have shown that $P(c + 1)$ is also true.

Thus the statement $P(r)$ is true for all $r \in [\ell - 1]$. To be precise, there exist $2\ell - 2$ distinct elements x_1, y_1, x_2, y_2, ... , $x_{\ell-1}$, $y_{\ell-1}$ such that $f_i(x_i) = f_i(y_i)$ for all $i \in [\ell-1]$.

Now we consider $X' = X \backslash \{y_1, x_2, y_2, ..., x_{\ell-1}, y_{\ell-1}\}$. We have $|X'| \geq (m+2\ell-2) - (2\ell - 3) = m + 1 > m$. So, there is at least one pair of elements x_ℓ and y_ℓ in X' such that $f_\ell(x_\ell) = f_\ell(y_\ell)$.

If $y_\ell = x_1$, let $C_1 = \{x_1, x_2, ..., x_{\ell-1}\}$ and $C_2 = \{y_1, y_2, ..., y_{\ell-1}, x_\ell\}$. If $y_\ell \neq x_1$,

let $C_1 = \{x_1, x_2, ..., x_\ell\}$ and $C_2 = \{y_1, y_2, ..., y_\ell\}$. We have $|C_1| \leq k$, $|C_2| \leq k$ and $C_1 \cap C_2 = \emptyset$. However, none of $f_i \in \mathcal{F}$ can separate C_1 and C_2. This contradicts our assumption that $\mathcal{F}$ is SHF$(\ell; n, m, \{k, k\})$. Therefore $n \leq m + 2\ell - 3$. □

For any hash family $\mathcal{F}$, let $G(\mathcal{F}) = (V, E)$ be an edge-colored graph corresponding to $\mathcal{F}$, where $V = X$ and for any $x \neq y$ in X, $(x, y) \in E$ if $f_i(x) = f_i(y)$ for some $i \in [\ell]$. Label each edge (x, y) of graph by i when $f_i(x) = f_i(y)$. Draw multiple edges if there is more than one function $f_i \in \mathcal{F}$ such that $f_i(x) = f_i(y)$. For any edge-colored graph G with exactly k colors/labels, let $\mathfrak{H}(G)$ be the set of all subgraphs of G that are induced from a set of k edges with distinct labels.

Theorem 10.4.4. *Let X and Y be two finite sets such that $|X| = n$ and $|Y| = m$ and let k and ℓ be integers such that $k \geq \ell \geq 2$. Let $\mathcal{F}$ be a family of functions $\{f_i : X \to Y, i \in [\ell]\}$. Then $\mathcal{F}$ is an* SHF$(\ell; n, m, \{k, k\})$ *if and only if all subgraphs $H \in \mathfrak{H}(G(\mathcal{F}))$ are non-2-vertex-colorable graphs.*

Proof. Let $\mathcal{F}$ be an SHF$(\ell; n, m, \{k, k\})$. Let $H \in \mathfrak{H}(G(\mathcal{F}))$.

Assume that H is a 2-vertex-colorable graph, properly colored red and blue. Let C_1 be the set of red vertices and let C_2 be the set of blue vertices. Then $C_1, C_2 \subseteq X$ are such that $|C_1| \leq k$ and $|C_2| \leq k$, and $C_1 \cap C_2 = \emptyset$. However, $f_i(C_1) \cap f_i(C_2) \neq \emptyset$ for all $f_i \in \mathcal{F}$ since the edge with label i in H is adjacent to a red and a blue vertex. This violates the SHF$(\ell; n, m, \{k, k\})$ property, hence H is a non-2-vertex-colorable graph.

Conversely, assume that all subgraphs $H \in \mathfrak{H}(G(\mathcal{F}))$ are non-2-vertex-colorable graphs. Assume for a contradiction that $\mathcal{F}$ is not an SHF$(\ell; n, m, \{k, k\})$. So there exist $C_1, C_2 \subseteq X$ such that $|C_1| \leq k$ and $|C_2| \leq k$, and $C_1 \cap C_2 = \emptyset$, and for each $i \in [\ell]$ there exist $x_i \in C_1$ and $y_i \in C_2$ such that $f_i(x_i) = f_i(y_i)$. Let H be a subgraph of $G(\mathcal{F})$ with vertices $x_1, y_1, ..., x_\ell, y_\ell$ and edges $\{x_i, y_i\}$ for all $i \in [\ell]$. Thus $H \in \mathfrak{H}(G(\mathcal{F}))$. Color the vertices in C_1 red and color the vertices in C_2 blue. Since $C_1 \cap C_2 = \emptyset$, H is 2-vertex-colorable, which contradicts the assumption. Therefore $\mathcal{F}$ is an SHF$(\ell; n, m, \{k, k\})$.

Thus, we can conclude that $\mathcal{F}$ is an SHF$(\ell; n, m, \{k, k\})$ if and only if all subgraphs

$H \in \mathfrak{H}(G(\mathcal{F}))$ are non-2-vertex-colorable graphs. □

From this point onwards, unless stated otherwise, by 2-colorable we mean 2-vertex-colorable.

Theorem 10.4.5. *If there exists an* SHF$(3; n, m, \{3, 3\})$*, then* $n \leq m + 1$.

Proof. Let $\mathcal{F}$ be an SHF$(3; n, m, \{3, 3\})$. Assume for a contradiction, that $n \geq m + 2$. Suppose $f \in \mathcal{F}$ is not surjective, say $a \notin f(X)$. Let $z \in X$ be such that $f(z) = f(z')$ for some $z' \in X \backslash \{z\}$. We may modify f to increase the size of $f(X)$ by defining

$$f'(x) = \begin{cases} f(x) & \text{for } x \neq z \\ a & \text{for } x = z \end{cases}.$$

Then $(\mathcal{F} \backslash \{f\}) \cup \{f'\}$ is also an SHF$(3; n, m, \{3, 3\})$. So, without loss of generality, we may assume that all functions in $\mathcal{F}$ are surjective.

Since $n \geq m + 2$, for each $i \in [\ell]$ there are at least 2 pairs of elements of X, $\{x_i, y_i\}$ and $\{u_i, v_i\}$ where $f_i(x_i) = f_i(y_i)$ and $f_i(u_i) = f_i(v_i)$. Note that these 2 pairs are not necessary disjoint.

Let $H \in \mathfrak{H}(G(\mathcal{F}))$. By Theorem 10.4.4, H is not a 2-colorable graph. Since H is a non-2-colorable graph with 3 edges, H must be a triangle. If there are at least two edges in $G(\mathcal{F})$ with label 1, consider H' a subgraph of $G(\mathcal{F})$ with 3 edges, such that edges labeled 2 and 3 are the same as H, but edge with label 1 is different. Then H' is a 2-colorable graph and $\mathcal{F}$ is not a SHF$(\ell; n, m, \{3, 3\})$. Hence there is a unique edge in $G(\mathcal{F})$ with label 1. Therefore $n \leq m + 1$, by the definition of $G(\mathcal{F})$. □

Example 23. The following matrix gives an SHF$(3; m + 1, m, \{3, 3\})$ when $m > 1$.

$$\begin{pmatrix} 0 & 0 & 1 & 2 & 3 & \ldots & m-2 & m-1 \\ 1 & 0 & 0 & 2 & 3 & \ldots & m-2 & m-1 \\ 0 & 1 & 0 & 2 & 3 & \ldots & m-2 & m-1 \end{pmatrix}$$

Hence, the following corollary holds.

Corollary 10.4.6. *Let m be a positive integer greater than 1. An* $\mathrm{SHF}(3; m+1, m, \{3, 3\})$ *exists and is optimal.*

Theorem 10.4.7. *If there exists an* $\mathrm{SHF}(4; n, m, \{4, 4\})$*, then* $n \leq m + 1$.

Proof. Let $\mathcal{F}$ be an $\mathrm{SHF}(4; n, m, \{4, 4\})$. Assume for a contradiction, that $n \geq m + 2$. Suppose $f \in \mathcal{F}$ is not surjective, say $a \notin f(X)$. Let $z \in X$ be such that $f(z) = f(z')$ for some $z' \in X \backslash \{z\}$. We may modify f to increase the size of $f(X)$ by defining

$$f'(x) = \begin{cases} f(x) & \text{for } x \neq z \\ a & \text{for } x = z \end{cases}.$$

Then $(\mathcal{F} \backslash \{f\}) \cup \{f'\}$ is also an $\mathrm{SHF}(4; n, m, \{4, 4\})$. So, without loss of generality, we may assume that all functions in $\mathcal{F}$ are surjective.

Since $n \geq m + 2$, for each $i \in [\ell]$ there are at least 2 pairs of elements of X, $\{x_i, y_i\}$ and $\{u_i, v_i\}$ where $f_i(x_i) = f_i(y_i)$ and $f_i(u_i) = f_i(v_i)$. Note that these 2 pairs are not necessary disjoint.

Let $H \in \mathfrak{H}(G(\mathcal{F}))$. By Theorem 10.4.4, H is not a 2-colorable graph. Moreover, since $n \geq m + 2$ and there exist at least two edges for each label $i \in [\ell]$. Since H is a non-2-colorable graph with 4 edges, H must contain a triangle; so H must have one of the three forms in Figure 10.1. Since H contains exactly one edge per label, for convenience we label H as in Figure 10.1.

If H is of the form in Figures 10.1a or 10.1b, by choosing any other edge with label 3 we find a new subgraph of $G(\mathcal{F})$ with no triangle. This new subgraph is 2-colorable which contradicts the $\mathrm{SHF}(4; n, m, \{4, 4\})$ property of $\mathcal{F}$. Therefore, H can only have the form in Figure 10.1c.

For each label 1, 2 or 3, removing the edge with that label and adding another edge from $G(\mathcal{F})$ with the same label must produce a non-2-colorable graph. So, $G(\mathcal{F})$ must

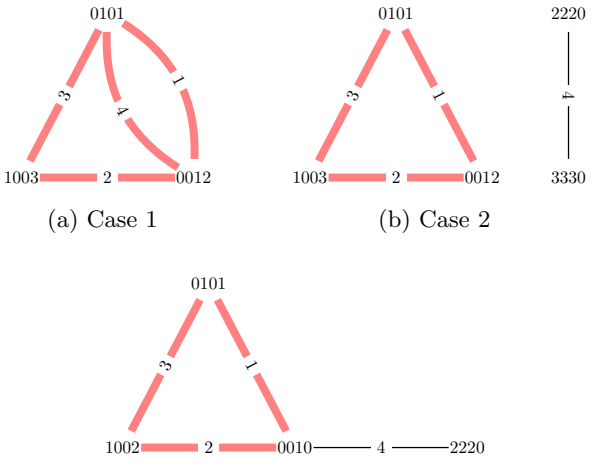


Figure 10.1: Possible subgraphs H for a SHF$(4; n, m, \{4, 4\})$

contain a subgraph of the form given in Figures 10.2a or 10.2b. However, each case the graph contains a 2-colorable subgraph H'; see Figures 10.3a or 10.3b, respectively.

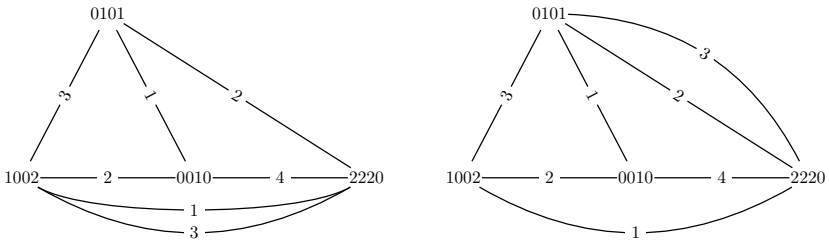
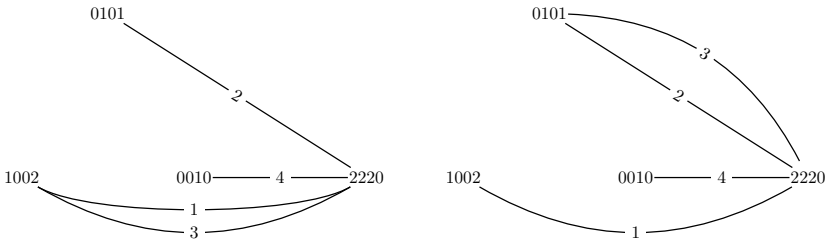


Figure 10.2: Possible subgraph of $G(\mathcal{F})$ in the proof of Theorem 10.4.7

So $G(\mathcal{F})$ always contains a 2-colorable subgraph $H \in \mathfrak{H}(G(\mathcal{F}))$. This contradiction shows that $n \leq m + 1$. □

Example 24. The following matrix gives an SHF$(4; m + 1, m, \{4, 4\})$ when $m > 2$.

$$\begin{pmatrix} 0 & 0 & 1 & 2 & 3 & \dots & m-2 & m-1 \\ 1 & 0 & 0 & 2 & 3 & \dots & m-2 & m-1 \\ 1 & 2 & 0 & 0 & 3 & \dots & m-2 & m-1 \\ 1 & 0 & 2 & 0 & 3 & \dots & m-2 & m-1 \end{pmatrix}$$



(a) Case 1 (b) Case 2

Figure 10.3: Possible subgraph H' of G(F) in the proof Theorem 10.4.7

Proof. The only subgraph $H \in \mathfrak{H}(G(\mathcal{F}))$ contains is the graph given in Figure 10.4. This is a non-2-colorable graph. By Theorem 10.4.4, $\mathcal{F}$ is an SHF$(4; m+1, m, \{4, 4\})$.

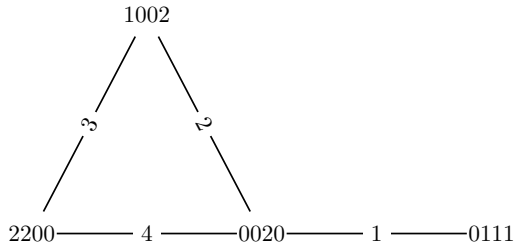


Figure 10.4: Subgraph H of $G(\mathcal{F})$ from Example 24

□

Hence, the following corollary holds.

Corollary 10.4.8. *Let m be a positive integer greater than 1. An* SHF$(4; m+1, m, \{4, 4\})$ *exists and is optimal.*

The next example pushes the lower bound of SHF$(k; n, m, \{k, k\})$ up to $m+1$; this is straightforward generalisation of Examples 23 and 24.

Example 25. The following matrix gives an SHF$(k; m+1, m, \{k, k\})$ when $m > k-2$.

$$\begin{pmatrix}
0 & 0 & 1 & 2 & 3 & 4 & \dots & k-5 & k-4 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 0 & 0 & 2 & 3 & 4 & \dots & k-5 & k-4 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 0 & 0 & 3 & 4 & \dots & k-5 & k-4 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 3 & 0 & 0 & 4 & \dots & k-5 & k-4 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 3 & 4 & 0 & 0 & \dots & k-5 & k-4 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 2 & 3 & 4 & 5 & 6 & \dots & 0 & 0 & k-3 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 3 & 4 & 5 & 6 & \dots & k-3 & 0 & 0 & k-2 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 3 & 4 & 5 & 6 & \dots & k-3 & k-2 & 0 & 0 & k-1 & k & \dots & m-2 & m-1 \\
1 & 2 & 3 & 4 & 5 & 6 & \dots & k-3 & 0 & k-2 & 0 & k-1 & k & \dots & m-2 & m-1
\end{pmatrix}$$

This is very straightforward since the only possible subgraph H of $G(\mathcal{F})$ is a triangle connected to a straight line, which is non-2-colorable. Hence, $\mathcal{F}$ is an SHF$(k; m+1, m, \{k,k\})$ by Theorem 10.4.4.

The next example is of an SHF$(\ell; n, m, \{5,5\})$ of size $n = m+2$ showing that Corollaries 10.4.6 and 10.4.8 cannot be generalised in a straightforward way.

Example 26. The following matrix gives an SHF$(5; m+2, m, \{5,5\})$ when $m > 2$.

$$\begin{pmatrix}
0 & 0 & 1 & 2 & 1 & 3 & 4 & \dots & m-2 & m-1 \\
1 & 0 & 0 & 1 & 2 & 3 & 4 & \dots & m-2 & m-1 \\
2 & 1 & 0 & 0 & 1 & 3 & 4 & \dots & m-2 & m-1 \\
1 & 2 & 1 & 0 & 0 & 3 & 4 & \dots & m-2 & m-1 \\
0 & 1 & 2 & 1 & 0 & 3 & 4 & \dots & m-2 & m-1
\end{pmatrix}$$

Proof. The graph $G(\mathcal{F})$ is given in Figure 10.5. For each label i, the edges labelled i occur when f_i maps two elements in X to the alphabet symbols 0 or 1. Since the graph is symmetric with respect to labels and there exists a one-to-one function mapping from the outer cycle (f_i maps two elements to 0) to the inner star (f_i maps two elements to 1) we can reduce the subgraphs $H \in \mathfrak{H}(G(\mathcal{F}))$ we need to check 4 cases: see Figures 10.6a, 10.6b,10.6c and 10.6d, respectively. The bold edges in Figure 10.6 are the selected

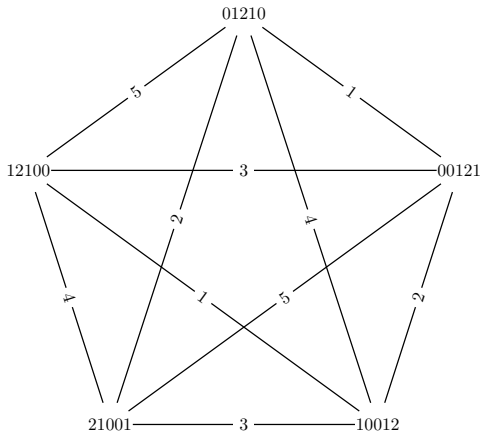


Figure 10.5: Graph of SHF$(5; m+2, m, \{5, 5\})$

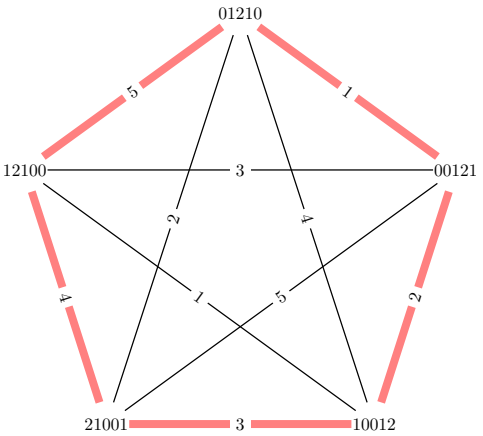
edges in H, where the red edges emphasise an odd-length cycle in H. Since, all possible subgraphs $H \in \mathfrak{H}(G(\mathcal{F}))$ contain at least one odd-length cycle, they are non-2-colorable. Therefore $\mathcal{F}$ is an SHF$(5; m+2, m, \{5, 5\})$.

□

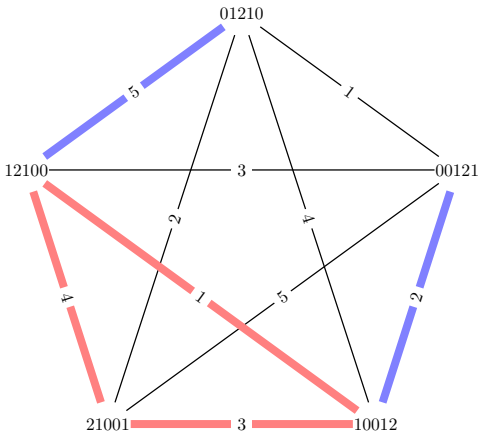
The next theorem gives a new upper bound on the size of SHF$(5; n, m, \{5, 5\})$. We achieve the result by deriving a contradiction from all possible cases of non-2-colorable subgraphs $H \in \mathfrak{H}(G(\mathcal{F}))$: proving an existence of a 2-colorable subgraph $H' \in \mathfrak{H}(G(\mathcal{F}))$ in each graph $G(\mathcal{F})$ containing H. The proof consists of plenty of cases and figures, hence it is quite long. An alternative approach is using the subgraphs in Figure 10.7 as an input to a graph searching algorithm as explained in Appendix A. We implemented this algorithm in C to verify the theorem for the most complicated cases.

Theorem 10.4.9. *If there exists an* SHF$(5; n, m, \{5, 5\})$*, then* $n \leq m + 2$.

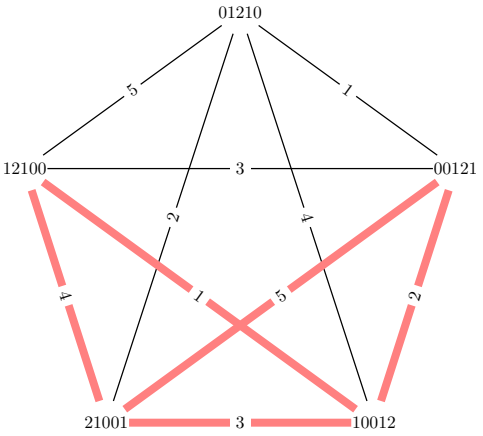
Proof. Let $\mathcal{F}$ be an SHF$(5; n, m, \{5, 5\})$. Assume for a contradiction, that $n \geq m + 3$. Suppose $f \in \mathcal{F}$ is not surjective, say $a \notin f(X)$. Let $z \in X$ be such that $f(z) = f(z')$



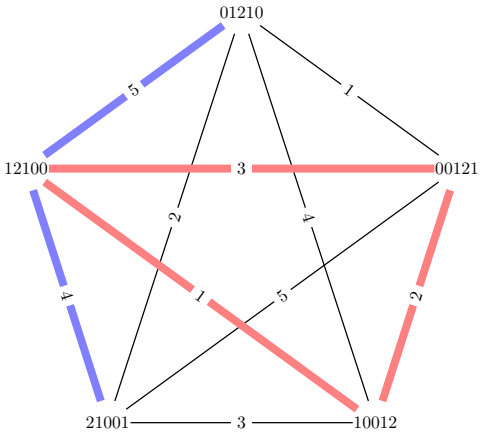
(a) 5 0's



(b) 4 0's and 1 1's



(c) 3 0's and 2 adjacent 1's



(d) 3 0's and 2 non-adjacent 1's

Figure 10.6: Subgraph H of $G(\mathcal{F})$ from Example 26

for some $z' \in X \backslash \{z\}$. We may modify f to increase the size of $f(X)$ by defining

$$f'(x) = \begin{cases} f(x) & \text{for } x \neq z \\ a & \text{for } x = z \end{cases}.$$

Then $(\mathcal{F} \backslash \{f\}) \cup \{f'\}$ is also an SHF$(5; n, m, \{5, 5\})$. So, without loss of generality, we may assume that all functions in $\mathcal{F}$ are surjective.

Since $n \geq m+3$, for each $i \in [\ell]$ there are at least 3 pairs of elements of X, $\{x_i^1, y_i^1\}$,

$\{x_i^2, y_i^2\}$ and $\{x_i^3, y_i^3\}$ where $f_i(x_i^1) = f_i(y_i^1)$, $f_i(x_i^2) = f_i(y_i^2)$ and $f_i(x_i^3) = f_i(y_i^3)$. Note that these 3 pairs are not necessary disjoint.

Let $H \in \mathfrak{H}(G(\mathcal{F}))$. By Theorem 10.4.4, H is not a 2-colorable graph. Moreover, since $n \geq m + 3$ and there exist at least three edges for each label $i \in [\ell]$. Since H is a non-2-colorable graph with 5 edges, H must contain an odd-length cycle; so H must have one of the fifteen forms in Figure 10.7. The bold edges in each subfigure represent its odd-length cycle(s). Since H contains exactly one edge per label, for convenience we label H as in Figure 10.7.

Cases 10.7a to 10.7e: If H is of any form in Figures 10.7a to 10.7e, by choosing any other edge with label 1 we find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle. This new subgraph is 2-colorable which contradicts the SHF$(5; n, m, \{5, 5\})$ property of $\mathcal{F}$. Therefore, H cannot be in these five forms.

Case 10.7f: If H is of the form in Figure 10.7f, by choosing any other edge with label 2, apart from the blue edge in Figure 10.8a, we find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle. This is always possible since there are at least 3 edges for label 2. This new subgraph is 2-colorable which contradicts the SHF$(5; n, m, \{5, 5\})$ property of $\mathcal{F}$. Therefore, H cannot be in this form.

Cases 10.7g, 10.7h, 10.7l, 10.7m and 10.7n: As in the previous case, we may argue that H cannot be in this five forms, subjecting to the blue edges in Figure 10.8b, 10.8c, 10.8d, 10.8e and 10.8f, respectively.

Case 10.7i: For each label 1, 2 or 3, removing the edge with that label and adding another edge from $G(\mathcal{F})$ with the same label must produce a non-2-colorable graph. So, with respect to the edges labeled 1, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.9b, 10.9d or 10.9e; with respect to the edges labeled 2, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.9a, 10.9b or 10.9c; with respect to the edges labeled 3, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.9a, 10.9d or 10.9e. All the subgraphs induced from the combinations of these blue edges are illustrated in Figures 10.10, 10.11 and 10.12.

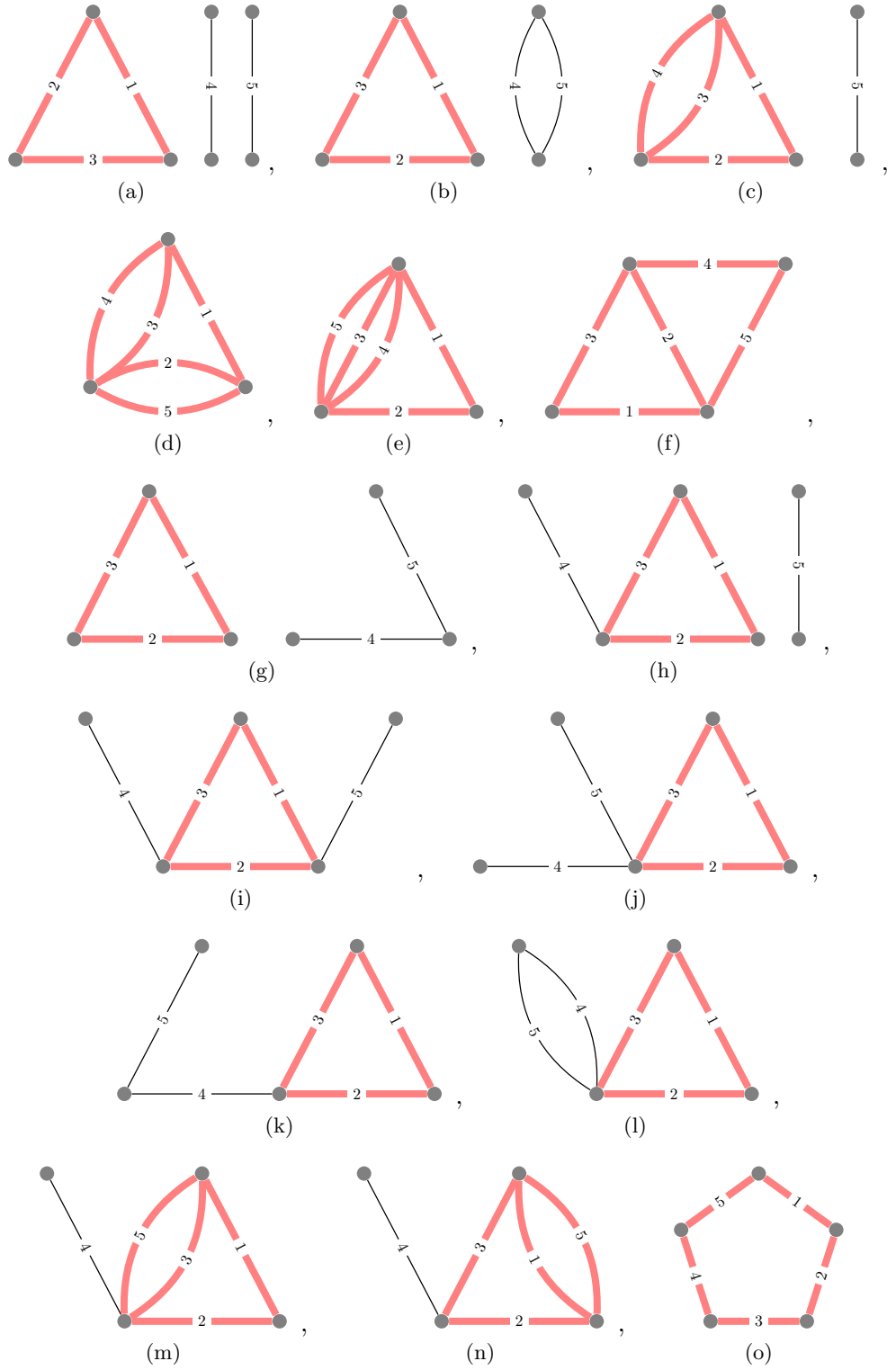


Figure 10.7: Possible subgraphs H for a SHF$(5; n, m, \{5, 5\})$

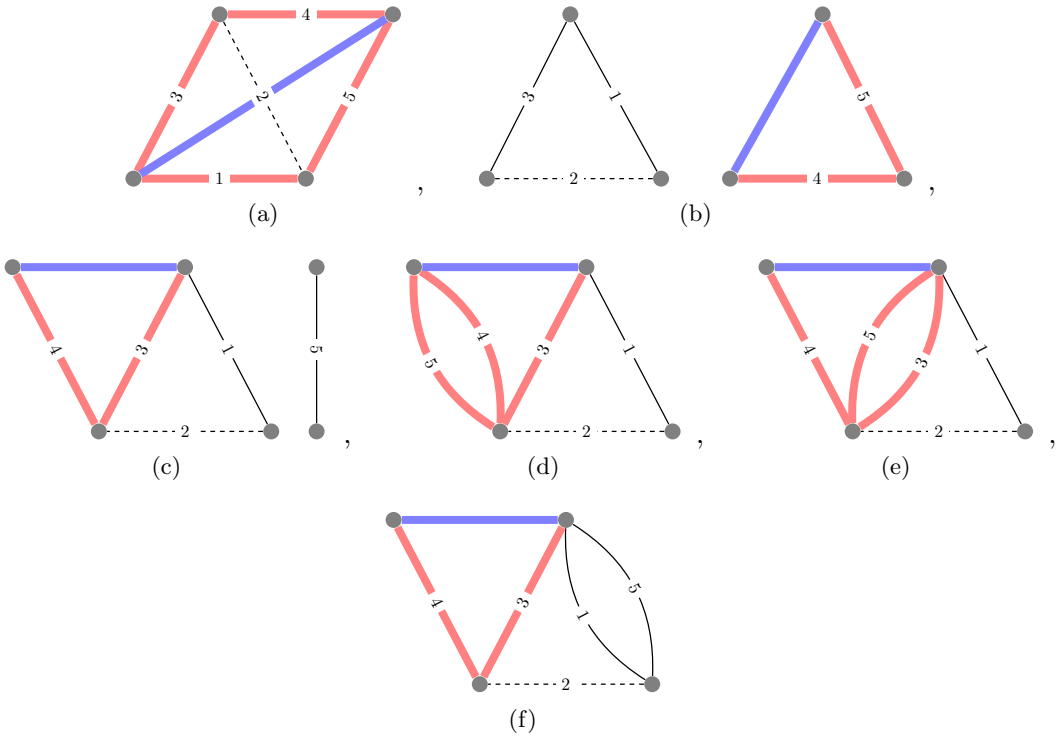


Figure 10.8: Subgraph of $G(\mathcal{F})$ in Cases 10.7f to 10.7h and Cases 10.7l to 10.7n

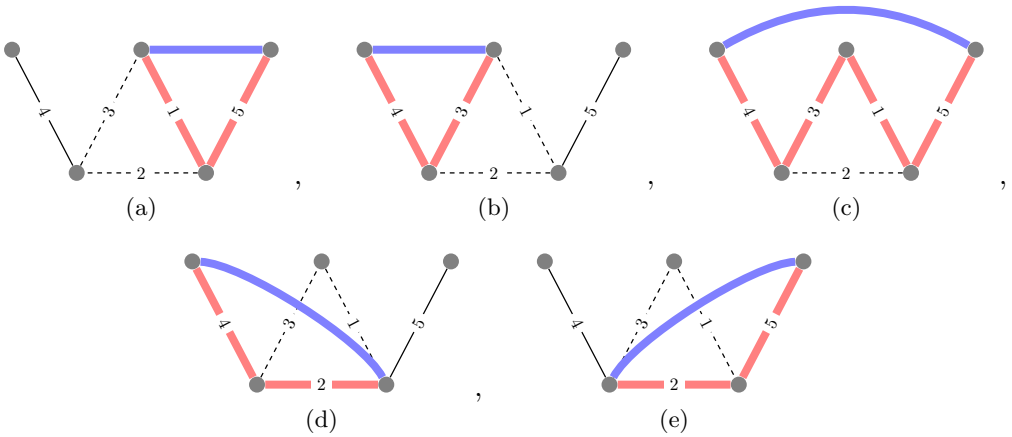


Figure 10.9: Subgraph of $G(\mathcal{F})$ from Case 10.7i

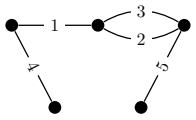
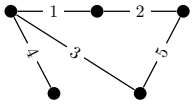
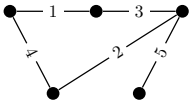
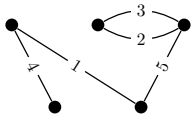
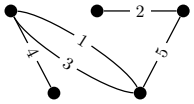
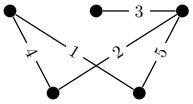
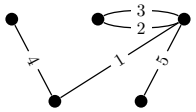
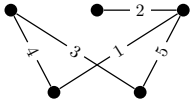
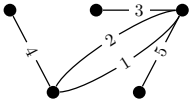
Edge labeled 1	Edge labeled 3		
	Figure 10.9a	Figure 10.9d	Figure 10.9e
Figure 10.9b	 (a)	 (b)	 (c)
Figure 10.9d	 (d)	 (e)	 (f)
Figure 10.9e	 (g)	 (h)	 (i)

Figure 10.10: Possible subgraph of $G(\mathcal{F})$ form Case 10.7i with edge labeled 2 from Figure 10.9a

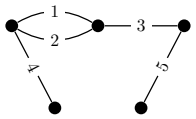
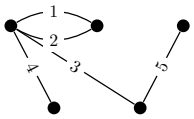
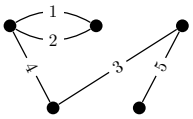
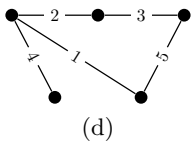
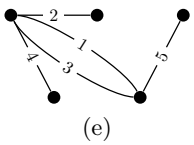
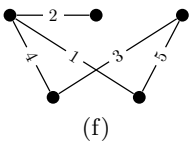
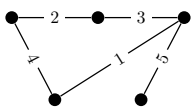
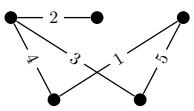
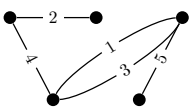
Edge labeled 1	Edge labeled 3		
	Figure 10.9a	Figure 10.9d	Figure 10.9e
Figure 10.9b	 (a)	 (b)	 (c)
Figure 10.9d	 (d)	 (e)	 (f)
Figure 10.9e	 (g)	 (h)	 (i)

Figure 10.11: Possible subgraph of $G(\mathcal{F})$ form Case 10.7i with edge labeled 2 from Figure 10.9b

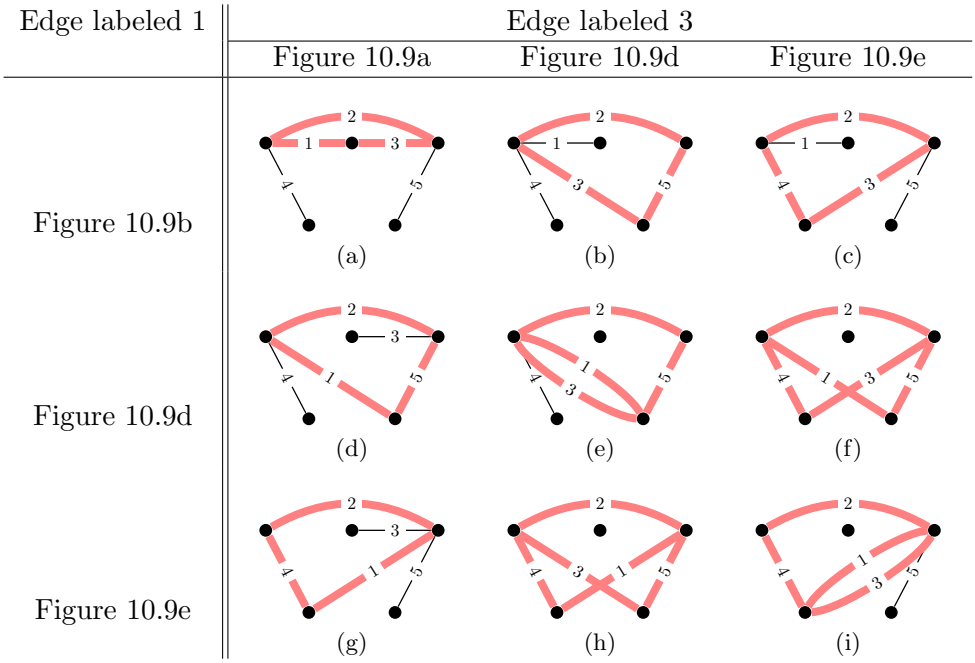


Figure 10.12: Possible subgraph of $G(\mathcal{F})$ form Case 10.7i with edge labeled 2 from Figure 10.9c

Only the edge labeled 2 in Figure 10.9c guarantees an odd-length cycle for any choices of edges labeled 1 and 3. Hence, by choosing any other edge with label 2, apart from the blue edge in Figure 10.9c, we can find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle. This is always possible since there are at least 3 edges for label 2. This new subgraph is 2-colorable which contradicts the SHF$(5; n, m, \{5, 5\})$ property of $\mathcal{F}$. Therefore, H cannot be in this form.

Case 10.7j: For each label 1, 2 or 3, removing the edge with that label and adding another edge from $G(\mathcal{F})$ with the same label must produce a non-2-colorable graph. So, with respect to the edges labeled 1, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.13a, 10.13b or 10.13c; with respect to the edges labeled 2, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.13a, 10.13b or 10.13d; with respect to the edges labeled 3, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.13b, 10.13c or 10.13e. All the subgraphs induced from the combinations of these blue edges are illustrated in Figures 10.14, 10.15 and 10.16.

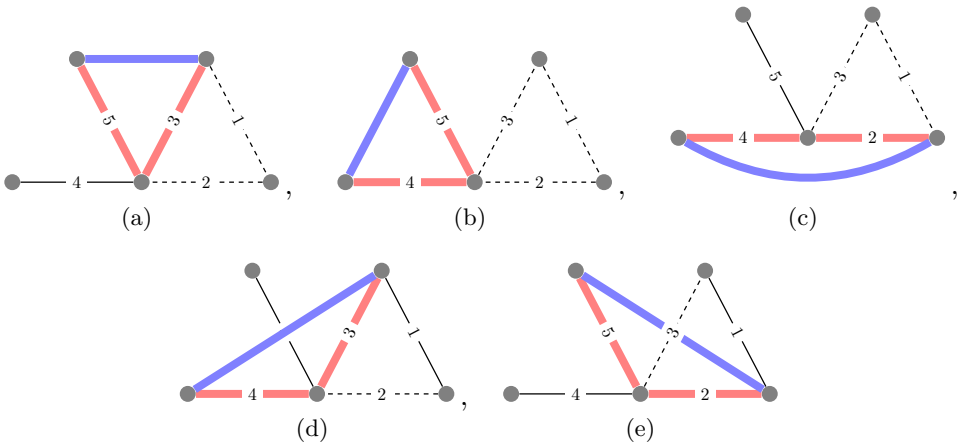


Figure 10.13: Subgraph of $G(\mathcal{F})$ form Case 10.7j

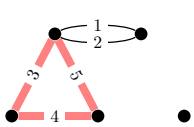
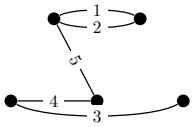
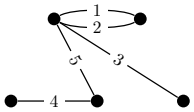
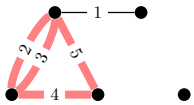
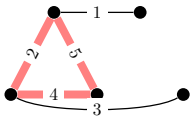
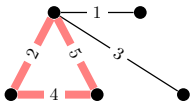
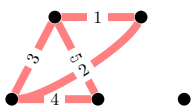
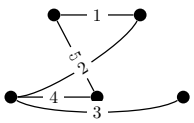
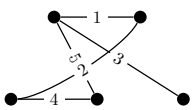
Edge labeled 2	Edge labeled 3		
	Figure 10.13b	Figure 10.13c	Figure 10.13e
Figure 10.13a	 (a)	 (b)	 (c)
Figure 10.13b	 (d)	 (e)	 (f)
Figure 10.13d	 (g)	 (h)	 (i)

Figure 10.14: Possible subgraph of $G(\mathcal{F})$ form Case 10.7j with edge labeled 1 from Figure 10.13a

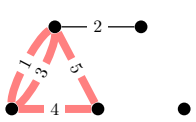
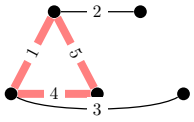
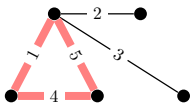
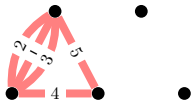
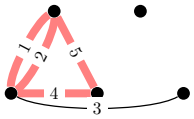
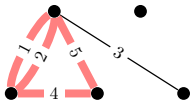
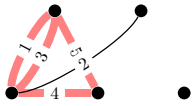
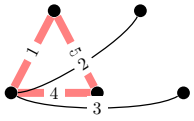
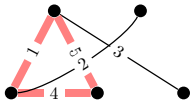
Edge labeled 2	Edge labeled 3		
	Figure 10.13b	Figure 10.13c	Figure 10.13e
Figure 10.13a	 (a)	 (b)	 (c)
Figure 10.13b	 (d)	 (e)	 (f)
Figure 10.13d	 (g)	 (h)	 (i)

Figure 10.15: Possible subgraph of $G(\mathcal{F})$ form Case 10.7j with edge labeled 1 from Figure 10.13b

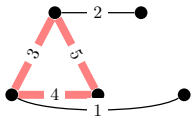
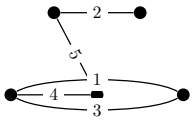
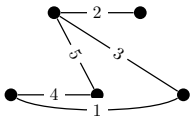
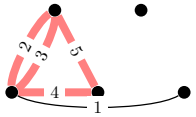
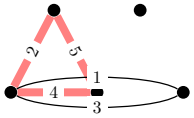
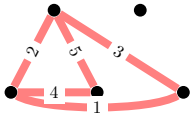
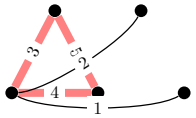
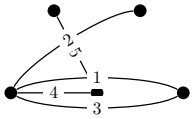
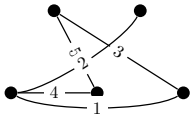
Edge labeled 2	Edge labeled 3		
	Figure 10.13b	Figure 10.13c	Figure 10.13e
Figure 10.13a	 (a)	 (b)	 (c)
Figure 10.13b	 (d)	 (e)	 (f)
Figure 10.13d	 (g)	 (h)	 (i)

Figure 10.16: Possible subgraph of $G(\mathcal{F})$ form Case 10.7j with edge labeled 1 from Figure 10.13c

Only the edge labeled 1 in Figure 10.13b guarantees an odd-length cycle for any choices of edges labeled 2 and 3. Hence, by choosing any other edge with label 1, apart from the blue edge in Figure 10.13b, we can find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle. This is always possible since there are at least 3 edges for label 1. This new subgraph is 2-colorable which contradicts the SHF$(5; n, m, \{5, 5\})$ property of $\mathcal{F}$. Therefore, H cannot be in this form.

Case 10.7k: If H is of the form in Figure 10.7k, by choosing any other edge with label 2, apart from the blue edges in Figure 10.17, we find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle.

Since there are at least 3 edges with label 2, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.18a, 10.18b or 10.18c. If $G(\mathcal{F})$ contains a subgraph of the form in Figures 10.18a or 10.18b, we have that $f_2(x) = f_2(y) = f_2(z)$. Thus, $G(\mathcal{F})$ contains a subgraph of the form given in Figure 10.19. If $G(\mathcal{F})$ contains a subgraph of the form in Figures 10.18c, we have that $f_2(x) = f_2(y) = f_2(z)$. Since there are at least m elements in $X\backslash\{x, y, z\}$, there exists either an element $u \in X\backslash\{x, y, z\}$ such that $f_2(u) = f_2(x)$ or a pair of elements $u, v \in X\backslash\{x, y, z\}$ such that $f_2(u) = f_2(v)$. This ensures there exists one more edge labeled 2 in the graph $G(\mathcal{F})$. Since the only possible edge labeled 2 that has not been included in the graph is of the form in Figure 10.17b, $G(\mathcal{F})$ must also contain a subgraph of the form given in Figure 10.19.

Similarly, by choosing any other edge with label 3, apart from the blue edges in Figure 10.20, we find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle. Since there are at least 3 edges with label 3, with similar arguments as above, we may conclude that $G(\mathcal{F})$ contains a subgraph of the form given in Figure 10.21. Therefore, $G(\mathcal{F})$ contains a subgraph of the form given in Figure 10.22 as the merging between Figures 10.19 and 10.21.

However, the graph in Figure 10.22 contains a subgraph H'' leading to a 2-colorable subgraph: as in Figure 10.23, any other choices for edge labeled 4 would produce a new subgraph H' of $G(\mathcal{F})$ with no odd-length cycle. This is always possible since there are

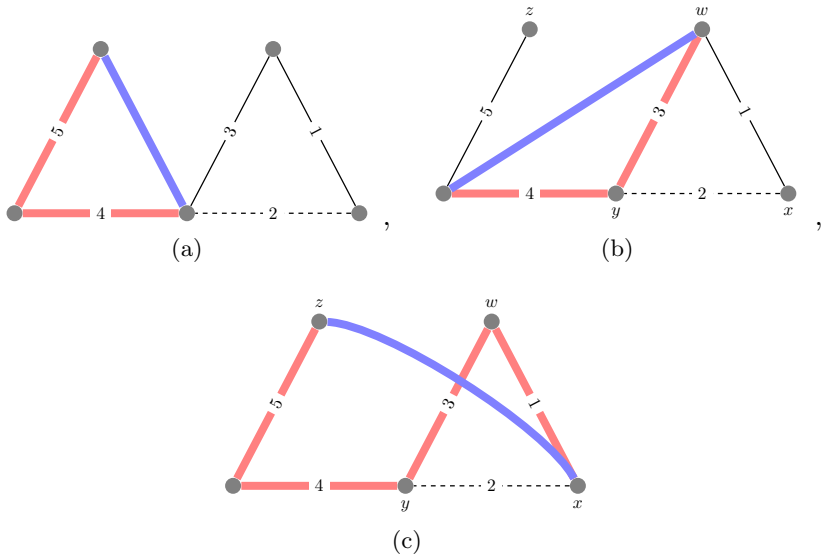


Figure 10.17: Possible subgraphs of $G(\mathcal{F})$ from Case 10.7k

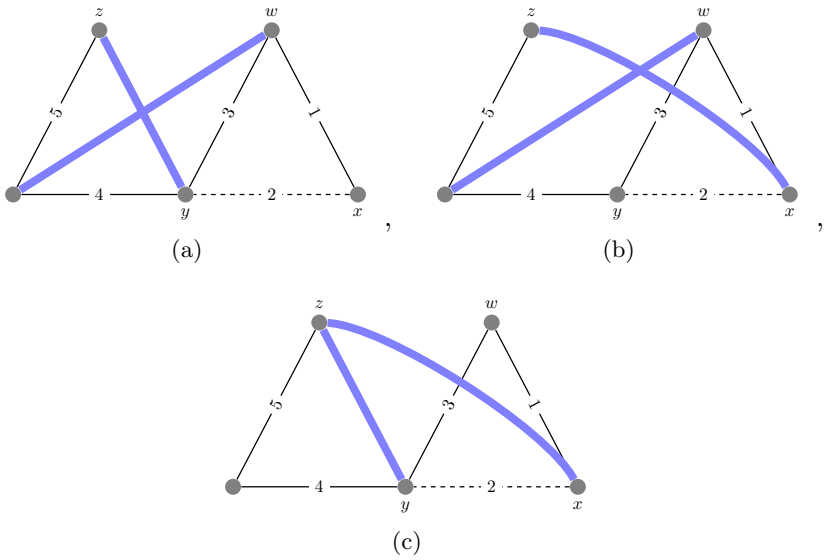
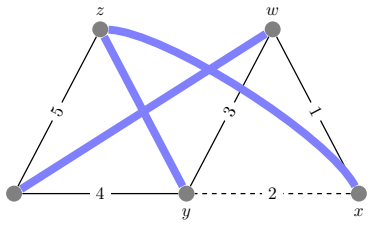


Figure 10.18: Possible edges labeled 2 from Case 10.7k



(a)

Figure 10.19: Subgraph of $G(\mathcal{F})$ from Case 10.7k

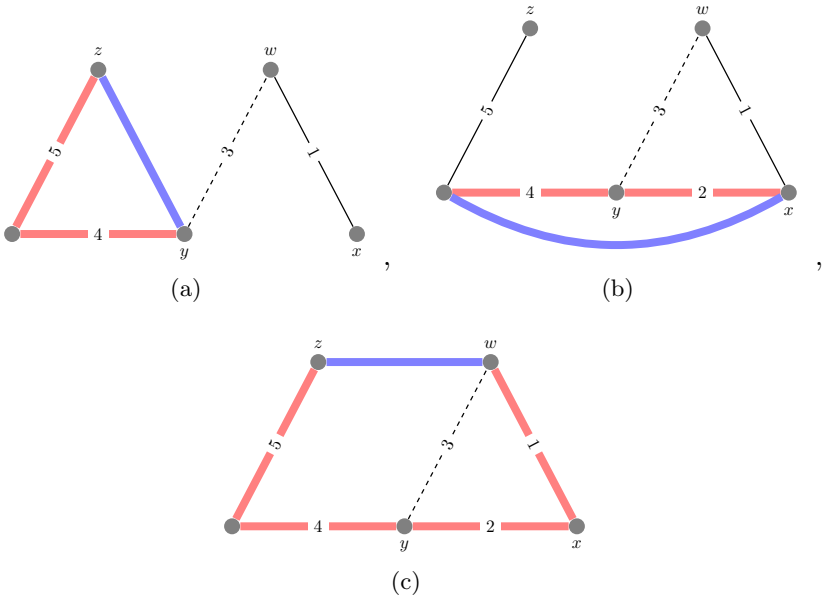
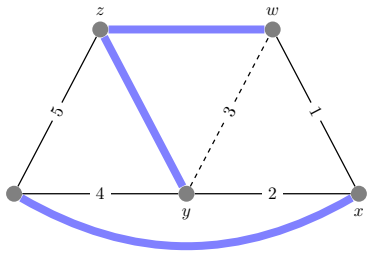
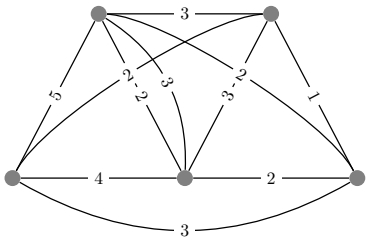


Figure 10.20: Possible subgraphs of $G(\mathcal{F})$ from Case 10.7k



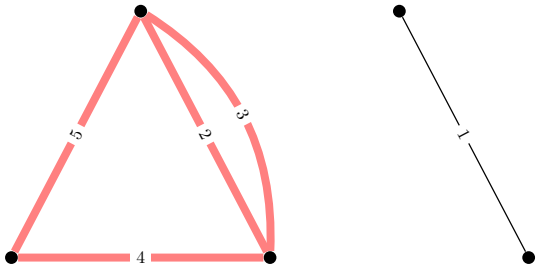
(a)

Figure 10.21: Subgraph of $G(\mathcal{F})$ from Case 10.7k



(a)

Figure 10.22: Subgraph of $G(\mathcal{F})$ from Case 10.7k



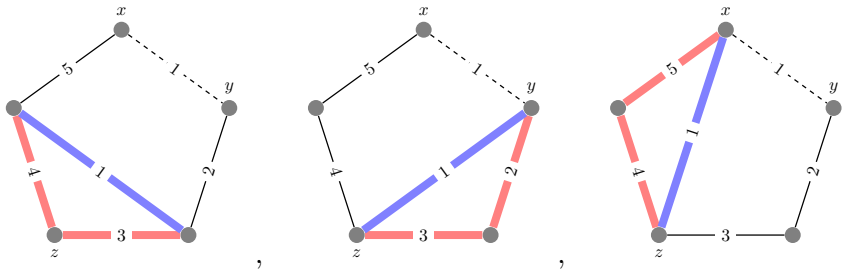
(a)

Figure 10.23: Subgraph H'' of $G(\mathcal{F})$ form Case 10.7k

at least 3 edges for each label. This new subgraph is 2-colorable which contradicts the SHF$(5; n, m, \{5, 5\})$ property of $\mathcal{F}$. Therefore, H cannot be in this form.

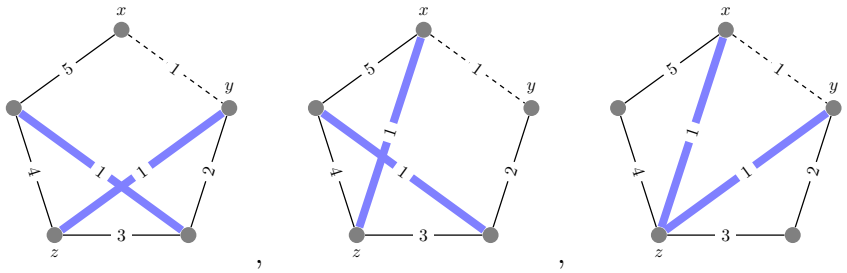
Case 10.7o: If H is of the form in Figure 10.7o, by choosing any other edge with label 1, apart from the blue edges in Figure 10.24, we find a new subgraph of $G(\mathcal{F})$ with no odd-length cycle.

Since there are at least 3 edges with label 1, $G(\mathcal{F})$ must contain a subgraph of the form given in Figures 10.25a, 10.25b or 10.25c. If $G(\mathcal{F})$ contains a subgraph of the



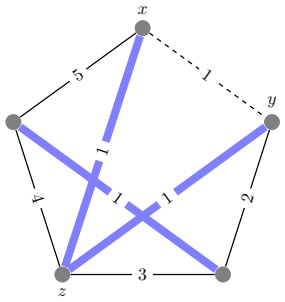
(a) (b) (c)

Figure 10.24: Possible subgraphs of $G(\mathcal{F})$ from Case 10.7o



(a) (b) (c)

Figure 10.25: Possible edges labeled 1 from Case 10.7o



(a)

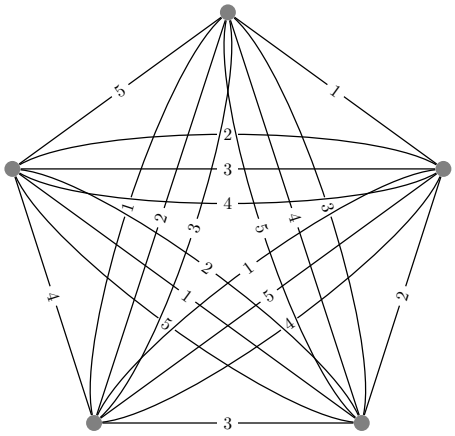
Figure 10.26: Subgraph of $G(\mathcal{F})$ from Case 10.7o

form in Figures 10.25a or 10.25b, we have that $f_1(x) = f_1(y) = f_1(z)$. Thus, $G(\mathcal{F})$ contains a subgraph of the form given in Figure 10.26. If $G(\mathcal{F})$ contains a subgraph of the form in Figures 10.25c, we have that $f_1(x) = f_1(y) = f_1(z)$. Since there are at least m elements in $X\backslash\{x, y, z\}$, there exists either an element $u \in X\backslash\{x, y, z\}$ such that $f_1(u) = f_1(x)$ or a pair of elements $u, v \in X\backslash\{x, y, z\}$ such that $f_1(u) = f_1(v)$. This ensures there exists one more edge labeled 1 in the graph $G(\mathcal{F})$. Since the only possible edge labeled 1 that has not been included in the graph is of the form in Figure 10.24a, $G(\mathcal{F})$ must also contain a subgraph of the form given in Figure 10.26.

Since the graph is symmetric with respect to labels, $G(\mathcal{F})$ must contain a subgraph of the form given in Figure 10.27.

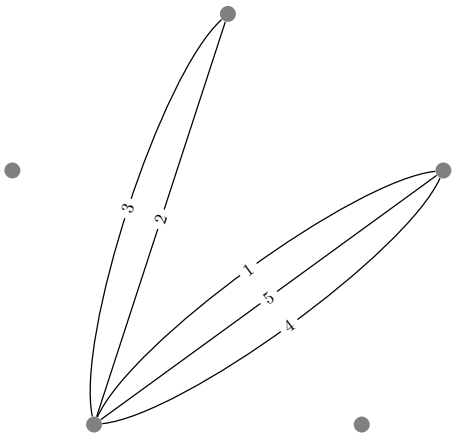
However, the graph in Figure 10.27 contains a 2-colorable subgraph H'; see Figure 10.28.

So $G(\mathcal{F})$ always contains a 2-colorable subgraph $H \in \mathfrak{H}(G(\mathcal{F}))$. This contradiction



(a)

Figure 10.27: $G(\mathcal{F})$ from Case 10.7o



(a)

Figure 10.28: Subgraph H' of $G(\mathcal{F})$ from Case 10.7o

shows that $n \leq m + 2$. □

Hence, the following corollary holds.

Corollary 10.4.10. *Let m be a positive integer greater than 1. An* SHF$(5; m + 2, m, \{5, 5\})$ *exists and is optimal.*

The next example pushes the lower bound of SHF$(k; n, m, \{k, k\})$ up to $m+2$; this is straightforward generalisation of Examples 26.

Example 27. The following matrix gives an SHF$(k; m+2, m, \{k, k\})$ when $m > k-3$.

$$\begin{pmatrix}
0 & 0 & 1 & 1 & 2 & 3 & 4 & \dots & k-8 & k-7 & k-6 & k-5 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 0 & 0 & 1 & 1 & 3 & 4 & \dots & k-8 & k-7 & k-6 & k-5 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 0 & 0 & 1 & 1 & 4 & \dots & k-8 & k-7 & k-6 & k-5 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 0 & 0 & 1 & 1 & \dots & k-8 & k-7 & k-6 & k-5 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 0 & 0 & 1 & \dots & k-8 & k-7 & k-6 & k-5 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & 0 & 0 & 1 & 1 & k-4 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & k-4 & 0 & 0 & 1 & k-3 & 1 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & k-4 & 1 & 0 & 0 & 1 & k-3 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & k-4 & k-3 & 1 & 0 & 0 & 1 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & k-4 & 1 & k-3 & 1 & 0 & 0 & k-2 & k-1 & \dots & m-2 & m-1 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots & k-4 & 0 & 1 & k-3 & 1 & 0 & k-2 & k-1 & \dots & m-2 & m-1
\end{pmatrix}$$

This is quite straightforward since $G(\mathcal{F})$ is a K_5 connected to two straight lines. Since there are 5 labels involve in the K_5 part, labeled as in Example 26, all subgraphs H of $G(\mathcal{F})$ must contain an odd-length cycle, which is non-2-colorable. Hence, $\mathcal{F}$ is an SHF$(k; m+2, m, \{k, k\})$ by Theorem 10.4.4.

Chapter 11

Open Problems

In this chapter, we gather some open problems arising from this thesis.

1. Let $g \leq q$, and let C be a q-ary t-TA code of length ℓ. Does there always exist a q-ary (T,t)-TA code C' of length ℓ of cardinality at least a half of the original code C, containing g groups? If this is not true in general, is it true when g is small ($g \leq q$)?

2. Are there any constructions of two-level fingerprinting codes that are better than the trivial construction, when the number of groups is greater than the alphabet size?

3. Are there any upper bounds on the size of two-level codes that have a smaller leading term than the bounds for one-level codes?

4. Can we find explicit constructions for (K,k)-SFP and (K,k)-TA?

5. Is it possible to reduce the upper bound of k-FP to $\left(\frac{\ell}{\ell-(r-1)\lceil\frac{\ell}{k}\rceil}\right) q^{\lceil\frac{\ell}{k}\rceil}$, where r is a unique positive integer in $\{1,2,...,k\}$ such that $r = \ell \mod k$?

Bibliography

[1] N. Alon, E. Fischer, and M. Szegedy. Parent-identifying codes. *Journal of Combinatorial Theory, Series A*, 95(2):349 – 359, 2001. 34

[2] N. Alon and U. Stav. New bounds on parent-identifying codes: The case of multiple parents. *Combinatorics, Probability & Computing*, pages 795–807, 2004. 35, 36

[3] N. Anthapadmanabhan and A. Barg. Two-level fingerprinting: Stronger definitions and code constructions. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 2528 –2532, june 2010. 15

[4] N. P. Anthapadmanabhan and A. Barg. Two-level fingerprinting codes, 2009. `http://www.citebase.org/abstract?id=oai:arXiv.org:0905.0417`. 13, 16, 39, 42, 75

[5] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky, and G. Zemor. A hypergraph approach to the identifying parent property: The case of multiple parents. *SIAM Journal on Discrete Mathematics*, 14(3):423–431, 2001. 36

[6] M. Bazrafshan and T. van Trung. On optimal bounds for separating hash families. Preprint, 2009. 106

[7] M. Bazrafshan and T. van Trung. Bounds for separating hash families. *Journal of Combinatorial Theory, Series A*, 118(3):1129 – 1135, 2011. 33, 34, 86, 97, 107

[8] M. Bazrafshan and T. van Trung. Improved bounds for separating hash families. *Designs, Codes, and Cryptography*, pages 1–14, 2012. 92

[9] S. R. Blackburn. Combinatorial schemes for protecting digital content. *Surveys in combinatorics 2003*, 307:43–78, 2003. 12, 21, 24, 32

[10] S. R. Blackburn. Frameproof codes. *SIAM Journal on Discrete Mathematics*, 16(3):499–510, 2003. 31, 32, 70, 93, 97

[11] S. R. Blackburn. An upper bound on the size of a code with the k-identifiable parent property. *Journal of Combinatorial Theory, Series A*, 102(1):179 – 185, 2003. 35

[12] S. R. Blackburn, M. Burmester, Y. Desmedt, and P. R. Wild. Efficient multiplicative sharing schemes. In *Proceedings of the 15th annual international conference on Theory and application of cryptographic techniques*, EUROCRYPT'96, pages 107–118, Berlin, Heidelberg, 1996. Springer-Verlag. 16

[13] S. R. Blackburn, T. Etzion, and S.-L. Ng. Traceability codes. *Journal of Combinatorial Theory, Series A*, 117(8):1049–1057, Nov. 2010. 37, 38, 44, 80

[14] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 47:1897–1905, 1998. 11, 22, 30, 33

[15] B. Chor, A. Fiat, and M. Naor. Tracing traitors. In *CRYPTO '94: Proceedings of the 14th Annual International Cryptology Conference on Advances in Cryptology*, pages 257–270, London, UK, 1994. Springer-Verlag. 22, 36

[16] B. Chor, A. Fiat, M. Naor, and B. Pinkas. Tracing traitors. *IEEE Transactions on Information Theory*, 46:893–910, 1994. 36

[17] G. Cohen and S. Encheva. Efficient constructions of frameproof codes. *Electron. Lett.*, 36:1840–1842, 2000. 32

[18] G. Cohen and S. Encheva. Some new p-ary two-secure frameproof codes. *Applied Mathematics Letters*, 14(2):177 – 182, 2001. 34

[19] G. Cohnen, S. Encheva, S. Litsyn, and H. Schaathun. Intersecting codes and separating codes. *Discrete Applied Mathematics*, 128(1):75 – 83, 2003. International Workshop on Coding and Cryptography (WCC2001). 34

[20] A. Fiat and M. Naor. Broadcast encryption. In *Proceedings of the 13th annual international cryptology conference on Advances in cryptology*, CRYPTO '93, pages 480–491, New York, NY, USA, 1994. Springer-Verlag New York, Inc. 12, 16

[21] H. D. L. Hollman, J. H. van Lint, J.-P. Linnart, and L. M. G. M. Tolhuizen. On codes with the identifiable parent property. *Journal of Combinatorial Theory, Series A*, 82(2):121 – 133, 1998. 22, 34, 35, 85

[22] K. Kim. *Perfect Hash Families: Constructions and Applications.* Dissertation in master of mathematics, Department of Combinatorics and Optimization, University of Waterloo, 2003. 16

[23] P. C. Li, G. H. J. van Rees, and R. Wei. Constructions of 2-cover-free families and related separating hash families. *Journal of Combinatorial Designs*, 14(6):423–440, 2006. 91

[24] L. Liu and H. Shen. Explicit constructions of separating hash families from algebraic curves over finite fields. *Designs, Codes and Cryptography*, 41:221–233, 2006. 10.1007/s10623-006-9004-y. 88

[25] Y. Meng Chee and X. Zhang. Improved Constructions of Frameproof Codes. *ArXiv e-prints*, June 2012. 96

[26] A. Panoui. *Wide-Sense Fingerprinting Codes and Honeycomb Arrays.* Ph.D. thesis in Mathematics, Department of Mathematics Royal Holloway, University of London, 2012. 14, 30

[27] R. Safavi-Naini and H. Wang. New constructions for multicast re-keying schemes using perfect hash families. In *Proceedings of the 7th ACM conference on Computer and communications security*, CCS '00, pages 228–234, New York, NY, USA, 2000. ACM. 16

[28] P. Sarkar and D. R. Stinson. Frameproof and IPP codes. In *Proceedings of the Second International Conference on Cryptology in India: Progress in Cryptology*, INDOCRYPT '01, pages 117–126, London, UK, 2001. Springer-Verlag. 16, 32

[29] J. N. Staddon, D. R. Stinson, and R. Wei. Combinatorial properties of frameproof and traceability codes. *IEEE Transactions on Information Theory*, 47:1042–1049, 2000. 16, 24, 25, 28, 30, 31, 33, 34, 35, 37

[30] D. Stinson, R. Wei, and K. Chen. On generalized separating hash families. *Journal of Combinatorial Theory, Series A*, 115(1):105 – 120, 2008. 16, 83, 85

[31] D. Stinson and G. M. Zaverucha. Some improved bounds for secure frameproof codes and related separating hash families. *IEEE Transactions on Information Theory*, 54:2508–2514, 2008. 16, 33

[32] D. R. Stinson. On some methods for unconditionally secure key distribution and broadcast encryption. *Designs, Codes and Cryptography*, 12(3):215–243, Nov. 1997. 16

[33] D. R. Stinson, T. van Trung, and R. Wei. Secure frameproof codes, key distribution patterns, group testing algorithms and related structures. *Journal of Statistical Planning and Inference*, 86:595–617, 1997. 16, 22, 33, 81, 83, 84

[34] D. R. Stinson and R. Wei. Combinatorial properties and constructions of traceability schemes and frameproof codes. *SIAM Journal on Discrete Mathematics*, 11:41–53, 1998. 16, 30

[35] D. R. Stinson, R. Wei, and L. Zhu. New constructions for perfect hash families and related structures using combinatorial designs. *Journal of Combinatorial Designs*, 8:189–200, 1999. 88

[36] D. R. Stinson and G. M. Zaverucha. New bounds for generalized separating hash families, 2007. 85

[37] D. Tonien and R. Safavi-Naini. Explicit construction of secure frameproof codes. *International Journal of Pure and Applied Mathematics*, 6(3):343–360, 2005. 34

[38] N. R. Wagner. Fingerprinting. In *IEEE Symposium on Security and Privacy'83*, pages 18–22, 1983. 11

[39] Y. Yemane. *Codes with k-identifiable parent property.* Ph.D. thesis in Mathematics, Department of Mathematics Royal Holloway, University of London, 2002. 36

[40] G. Zaverucha. *Hash Families and Cover-Free Families with Cryptographic Applications.* Ph.D. thesis in computer science, School of Computer Science, University of Waterloo, 2010. 16

Appendix A

Appendix

This appendix contains an alternative proof of Theorem 10.4.9. The algorithm explained in this Appendix is general enough to be used in finding bounds for any $\mathrm{SHF}(k; n, m, \{k, k\})$. However, with a larger k, the computational cost for the algorithm might become unaffordable.

A.1 Algorithm

Recall that from Theorem 10.4.4 a separating hash family $\mathcal{F}$ can be represented by a graph $G(\mathcal{F})$ such that all $H \in \mathfrak{H}(G(\mathcal{F}))$ are non-2-colorable. In other words, given a graph G with such property, one may construct the corresponding hash family. From a non-2-colorable graph H with k edges, the following algorithm determines whether there exists a graph $G(\mathcal{F})$ containing H, such that all subgraphs H' of $G(\mathcal{F})$ in $\mathfrak{H}(G(\mathcal{F}))$ are non-2-colorable. The inputs to the algorithm are the parameters k (number of labels), the graph H with k edges, a set V of vertices for $G(\mathcal{F})$ and c, where $m + c$ is the size of $\mathcal{F}$.

To construct $G(\mathcal{F})$, we notice that it consists of c edges for each label, making a total of kc edges. Since $G(\mathcal{F})$ contains H, the algorithm attempts to add to H $c - 1$ edges for each label, and check whether the resulted graph satisfies the property above. The

remaining concern is to find these additional edges, which can be done in two stages. In the first stage, the algorithm computes for each edge $e_i \in H$, a set of edges E_i, so that by replacing e_i with any edge in E_i, the resulting graph is still non-2-colorable. Apparently, if there are less than $c-1$ such edges for any $e_i \in H$, then there exists no qualified graph $G(\mathcal{F})$ that contains H. This is done via the FindReplacement subroutine.

The second stage involves picking edges in many ways from the sets E_i to enlarge H, hoping to come across a graph $G(\mathcal{F})$ with the required property. By checking all possible ways to pick $c-1$ edges from each set E_i and adding them to H, one can exclusively test if any desirable graph $G(\mathcal{F})$ exists. Such a test for each candidate of G' of $G(\mathcal{F})$ is performed by the function IsNon2Colorable, which test whether all subgraphs of G' in $\mathfrak{H}(G')$ are non-2-colorable.

Further, note that with exhaustive search for all possible candidates $G(\mathcal{F})$, the computational cost for the algorithm might become unaffordable. Instead, we observe that if $\mathfrak{H}(G(\mathcal{F}))$ contains only non-2-colorable subgraphs of $G(\mathcal{F})$, then the same property must applies to $\mathfrak{H}(G')$, for any subgraph G' of $G(\mathcal{F})$. We also observe that while H is being enlarged (by adding edges), in many cases the resulting graph G' loses such property after adding just a few more edges to H. Hence, by applying IsNon2Colorable to G' every time new edges are added, one can eliminate wasted efforts for picking the remaining edges, in case G' has already lost the required property. We execute the algorithm for H as in the cases of Figures 10.7k and 10.7o, containing 231,159,852 and 243 possible candidates for $G(\mathcal{F})$, respectively. However, with the above improvement, to show that no $G(\mathcal{F})$ exists for such H, only 43,929 and 84 calls to IsNon2Colorable were needed.

Finally, since the description of $G(\mathcal{F})$ only involves its edges, it is questionable on how many vertices one would need to construct $G(\mathcal{F})$, as this is a required input for the algorithm. The following lemma gives a reasonable bound on the number of vertices needed to include all possible choices of $G(\mathcal{F})$ containing H.

Lemma A.1.1. *If there exists an* SHF$(k; m + c, m, \{k, k\})$*, for some positive integers* c*, then there exists a graph* $G(\mathcal{F})$ *containing at most* $(k-1)c+1$ *vertices.*

Proof. Let $\mathcal{F}$ be an SHF$(k; m + c, m, \{k, k\})$. It is not too difficult to see that if $G(\mathcal{F})$ is disconnected, we can create a new connected graph that preserves its property (correspond to an SHF$(k; m+c, m, \{k, k\})$ $\mathcal{F}'$) by randomly picking a vertex from each component and merge them into one vertex. This is always possible since joining these vertices does not produce any new cycle or destroy any existing cycle. Hence, without loss of generality, we may assume that $G(\mathcal{F})$ is a connected graph. This ensures that the number of vertices is at most $kc + 1$.

Furthermore, since we can partition the edges of $G'(\mathcal{F})$ into c edge-disjoint subgraphs $G_1, G_2, ..., G_c \in \mathfrak{H}(G(\mathcal{F}))$, $G'(\mathcal{F})$ must contains at least c edge-disjoint odd-length cycles. Each cycle quarantees the reduction of the number of vertices of $G'(\mathcal{F})$ further by one, and thus c vertices in total. Hence, there exists graph $G(\mathcal{F})$ containing at most $(k-1)c+1$ vertices. □

Algorithm A.1 Generate $G(\mathcal{F})$

Input: H, V, $k \in \mathbb{Z}^+$, and $c \in \mathbb{Z}^+$
Output: If exists, $G(\mathcal{F})$.

```
L ⇐ 1
Label each edge of H with a number, from 1 to k        ▷ edge-label the graph H
G ⇐ H
for i = 1 → k do
    e_i ⇐ edge labeled i in H
    E_i ⇐ FindReplacement(H, e_i, V)\{e_i}        ▷ find all possible new edges with label i
    if |E_i| < c − 1 then
        return Null
    end if
end for
for i = 1 → k do
    d_i ⇐ 1
    n_i ⇐ the number of (c − 1)-subsets of E_i
end for
while 1 ≤ L ≤ k do
    if d_L ≤ n_L then
        E'_L ⇐ the d_L-th (c − 1)-subset of E_L
        G ⇐ G ∪ E'_L
        d_L ⇐ d_L + 1
        if IsNon2Colorable(G, k) then
            L ⇐ L + 1
        end if
    else
        d_L ⇐ 1
        L ⇐ L − 1
        G ⇐ G\E'_L
    end if
end while
if L = k + 1 then
    return G
else
    return Null
end if
```

Algorithm A.2 Find the list of edge replacement candidates

Input: A graph H, an edge $e_a \in H$, and a set V of vertices
Output: A set of edges E_a such that $\{e'\} \cup H \backslash \{e\}$ is non-2-colorable for all $e' \in E_a$

```
function FindReplacement(H, e_a, V)
    E_a ⇐ ∅                                        ▷ initialize the set
    for all 2-subset {v_1, v_2} of V do
        e' ⇐ edge between v_1 and v_2
        if {e'} ∪ H\{e_a} is non-2-colorable then
            E_a ⇐ E_a ∪ {e'}                       ▷ add edge e' to the set
        end if
    end for
    return E_a
end function
```

Algorithm A.3 Check if an edge-labeled graph G has $\mathfrak{H}(G)$ containing only non-2-colorable subgraphs

Input: An edge-labeled graph G and a $k \in \mathbb{Z}^+$
Output: TRUE if all graphs $H' \in \mathfrak{H}(G(\mathcal{F}))$ are non-2-colorable, and FALSE otherwise

```
function IsNon2Colorable(G, k)
    L ⇐ 1
    G' ⇐ an empty graph
    for i = 1 → k do
        d_i ⇐ 1
        n_i ⇐ the number of edges of G with label i
    end for
    while 1 ≤ L ≤ k do
        if d_L ≤ n_L then
            e_L ⇐ the d_L-th edge with label L of G
            G' ⇐ G' ∪ {e_L}
            d_L ⇐ d_L + 1
            if L = k then
                if G' is 2-colorable then
                    return FALSE
                end if
            else
                L ⇐ L + 1
            end if
        else
            d_L ⇐ 1
            L ⇐ L − 1
            G' ⇐ G'\{e_L}
        end if
    end while
    return TRUE
end function
```