Available online at www.sciencedirect.com
ScienceDirect
Discrete Applied Mathematics 157 (2009) 971–981
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Linear codes for high payload steganography

Mahdad Khatirinejad*, Petr Lisoněk

Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Received 10 August 2006; received in revised form 28 November 2007; accepted 12 February 2008
Available online 26 March 2008

Abstract

Steganography is concerned with communicating hidden messages in such a way that no one apart from the sender and the intended recipient can detect the very existence of the message. We study the *syndrome coding method* (sometimes also called the "matrix embedding method"), which uses a linear code as an ingredient. Among all codes of a fixed block length and fixed dimension (and thus of a fixed information rate), an optimal code is one that makes it most difficult for an eavesdropper to detect the presence of the hidden message. We show that the average distance to code is the appropriate concept that replaces the covering radius for this particular application. We completely classify the optimal codes in the cases when the linear code used in the syndrome coding method is a one- or two-dimensional code over GF(2). In the steganography application this translates to cases when the code carries a high payload (has a high information rate).

Keywords: Steganography; Binary linear codes; Covering codes; Syndrome coding method

1. Introduction

Steganography is the scientific discipline concerned with communicating hidden messages in such a way that no one apart from the sender and the intended recipient can detect the *existence* of the message. This process is fundamentally different from *cryptography,* where the existence of a secret message may be suspected by anyone who can observe the scrambled ciphertext while it is communicated.

A common technique in steganography is to *embed* the hidden message into a larger *cover object* (such as a digital image, for example) by slightly distorting the cover object in a way that on one hand makes it possible for the intended recipient to extract the hidden message, but on the other hand makes it very hard for everybody else to detect the distortion of the cover object (i.e., to detect the existence of the hidden message). The amount of noise that is naturally (inherently) present in the cover object determines the amount of distortion that can be introduced into the cover object before the distortion becomes detectable.

Syndrome coding, sometimes also called *matrix embedding* [9,6,7] or *coset encoding* [3], is a steganography method which requires the sender and the recipient to agree in advance on a parity check matrix H; the secret message is then extracted by the recipient as the syndrome (with respect to H) of the received cover object.

* Corresponding author. Fax: +1 778 782 4947.
E-mail addresses: mahdad@math.sfu.ca (M. Khatirinejad), lisonek@math.sfu.ca (P. Lisoněk).

0166-218X/$ - see front matter
doi:10.1016/j.dam.2008.02.004

We restrict our study to the binary case. All codes considered in this article are binary linear codes. Let $C \subseteq \mathbb{F}_2^n$ be the code defined by H, where $\mathbb{F}_2$ denotes the field with two elements.

The syndrome coding method is surveyed in Section 2. In Section 3 we show that the amount of distortion introduced by this method is measured by the average weight of a coset leader for C (which is also the average distance from a vector in $\mathbb{F}_2^n$ to C). We denote this quantity as $R_a(C)$ since it can be viewed as an "averaged" version of the classical concept of the covering radius of C. We say that C is optimal if it minimizes $R_a(C)$ among all codes of the same block length and dimension. The main part of the article is Section 4.2 in which we completely classify the optimal two-dimensional codes using combinatorial counting methods. We conclude by discussing some possible applications of our results in Section 4.3.

2. Linear codes for steganography

Throughout this article we will use some standard concepts and results from coding theory which can be found for example in [10]. By $\mathbb{F}_2^k$ we denote the k-dimensional vector space over $\mathbb{F}_2$. (Depending on the choice which is more convenient in the context, we will use either row vectors or column vectors.) The standard basis of $\mathbb{F}_2^n$ will be denoted by $\{e_1, \ldots, e_n\}$, that is, $(e_i)_j = 1$ if $i = j$ and $(e_i)_j = 0$ otherwise. The $\mathbb{F}_2$-span of $\{v_1, \ldots, v_k\} \subset \mathbb{F}_2^n$ will be denoted by $\langle v_1, \ldots, v_k \rangle$. By $\mathbf{0}$ and $\mathbf{1}$ we denote the all-zero and all-one vector of the appropriate dimension. For simplicity, we write $b_1^{r_1} b_2^{r_2} \ldots b_s^{r_s}$ to represent a vector whose first r_1 coordinates are equal to b_1, next r_2 coordinates are equal to b_2, etc. By $\mathbb{F}_2^{k \times n}$ we will denote the set of all $k \times n$ matrices over $\mathbb{F}_2$.

Syndrome coding, sometimes also called *matrix embedding* [9,6,7] or *coset encoding* [3], is a steganography method that uses linear codes. Suppose that the cover object is a multimedia binary file, say a digital image consisting of n pixels. Suppose that one bit of information is extracted from each pixel of the image, for example the least significant bit (on the grayscale map) of that pixel. Let $E \in \mathbb{F}_2^n$ be the sequence of n bits extracted from the cover object, and let $M \in \mathbb{F}_2^m$ be the message that we want to embed ($m \leq n$). The sender and the recipient agree in advance on a matrix $H \in \mathbb{F}_2^{m \times n}$ of rank m. To perform the hidden message embedding, the sender finds a vector $\delta \in \mathbb{F}_2^m$ such that $H(E + \delta) = M$, that is, $H\delta = M - HE$, and then the sender changes the cover object in the following way: For any $j \in \{1, \ldots, n\}$ the sender leaves the j-th pixel of the image unchanged if $\delta_j = 0$, whereas (s)he flips the least significant bit of the j-th pixel if $\delta_j = 1$. The receiver recovers the hidden message by simply computing $M = HE'$, where E' is the sequence of n bits extracted from the distorted cover object. The total amount of distortion is thus the Hamming weight (number of ones) in the vector $\delta = E' - E$.

It is well known that the set of vectors x satisfying $Hx = M - HE$ is a *coset* of the linear code for which H serves as a *parity check matrix.* Finding a vector of the lowest weight in a coset is the well-known *coset leader problem.* We will assume that, in order to minimize the distortion, the sender will *always* use a coset leader for the vector δ introduced above. The largest weight of any coset leader is the *covering radius* of the code.

A recent historical account of the connection between Steganography and covering codes [3] appears in the introduction to [2] where the references [1,4,8,11] are listed. Another relevant reference is [5].

3. Average distance to a linear code

Let $k := n - m$ and let C be the $[n, k]$ code for which H serves as a parity check matrix, where n, m and H are as in the previous section. Since the message M is typically encrypted before being embedded into the cover object, it is reasonable to assume that M is drawn uniformly at random from $\mathbb{F}_2^m$. Therefore, the expected amount of distortion per one message M is equal to the *average weight of a coset leader*

$$\frac{1}{2^{n-k}} \sum_{u \in L(C)} w(u), \tag{1}$$

where $L(C)$ is a set of coset leaders for C and w denotes the Hamming weight function throughout the article. For every $v \in \mathbb{F}_2^n$, the *distance of v from C* is naturally defined as

$$d(v, C) = \min\{w(v - c) : c \in C\} = w(u),$$