Steganalysis and Game Equilibria

J. Mark Ettinger [*]
Los Alamos National Laboratory
Mail Stop B-230
Los Alamos, NM 87545
505-665-4789

Abstract

Steganography is the study of methods of concealing data in the noise of another data set. Steganalysis is the field of discovering hidden data and disrupting covert channels. We introduce a two-player, zero-sum, matrix game for the purpose of modeling the contest between a data-hider and a data-attacker. We then solve the game for equilibria, demonstrating that the form of the solution depends on whether the permitted distortion is less than or greater than d_c, the critical distortion. This critical point is a simple function of several parameters which define the game. We then solve two example cases to demonstrate the ideas presented in the general solution. The value of the game is the amount of information that may be reliably stored in the data set.

Keywords: Steganalysis, Cryptanalysis, Equilibria, Shannon Entropy, Binary Symmetric Channel, Active Warden Attack.

1 Introduction

Recently techniques in steganography have received a great deal of attention [1], [6], mostly from members of the signal processing and cryptological communities. Steganography concerns techniques for concealing data in the noise

[*] ettinger@lanl.gov

of other data. Primary applications include digital watermarking for copyright protection of digital data and covert communications. Many methods for hiding data have been proposed [3], [2], [7] but to this point there have been no quantitative methods developed for assessing the security of any of these methods.

Steganography is related to, though distinct from, cryptography. The goal of a secure cryptographic method is to prevent an interceptor from gaining any information about the plaintext from the intercepted ciphertext. The goal of a secure steganographic method is to prevent an observant intermediary from even obtaining knowledge of the mere presence of the secret data. In most applications the embedded message will be encrypted before hiding. Otherwise the natural language structure in the message will be statistically detectable. Also, we are primarily interested in the case where digital data is hidden in a digital data set. Therefore we assume that the hidden message is a pseudorandom bit sequence.

In cryptology, the complementary endeavor to finding secure encryption methods is called cryptanalysis and is concerned with discovering techniques for breaking cryptographic schemes. To this point in time the analogous field of *steganalysis* remains almost completely mathematically undeveloped. In order to begin to remedy this it is important to utilize the proper notion of breaking a steganographic system. The purpose of steganography is to hide data and therefore the primary counterobjectives are to *discover* the presence of hidden data (perhaps in a probabilistic setting) and/or to reduce the channel capacity of the covert channels. The first type of attack is sometimes called a *passive warden attack* and the second type an *active warden attack*. Notice that we do not demand that the hidden message is recovered for a steganalytic attack to be successful, though the question of to what degree this may be accomplished is certainly of interest. The question of then decrypting the recovered message is a classical cryptanalytic question and lies outside the interests of steganalysis proper.

Cryptology has emerged over the years from a collection of ad hoc techniques into a sophisticated discipline utilizing ideas from many areas of mathematics and contributing to the development of mathematics. We believe that steganology holds a similar promise of deep connections with and potential contributions to various branches of mathematics and applied mathematics, especially pattern recognition. With the hope of leveraging some of the success of cryptanalysis, in this paper we borrow several operating assumptions on our way to a mathematical assessment of the security of steganographic algorithms. In particular, we assume that all algorithms are public knowledge and that the security of the system resides in a *secret key*. This assumption, basic to cryptology, will manifest in the formulation of the matrix games that we use to model covert channels.

As we mentioned, there are two different types of attacks on steganographic systems. The first of the two attacks is detecting the presence of the hidden information and the second is interruption of communication by overwriting the hidden information. In this paper we introduce a game-theoretic model of the second type of attack. The data Hider chooses a distribution of locations to hide data in the data set subject to a limit on the amount of distortion he may introduce into the data set. This distortion parameter is thus an measure of the noise in the data set. The Attacker also chooses a distribution of locations to hide pseudorandom noise in an attempt to overwrite the hidden data. In this paper the Attacker is subject to the same distortion limit as the Hider. Associated to any pair of strategies there is an associated payoff which measures the amount of data that is communicated. The Hider desires a strategy which maximizes this payoff whereas the Attacker desires to minimize this payoff. We therefore have a two-player, zero-sum game and we then proceed to solve this game for optimal strategies and payoffs.

This paper is organized as follows. In the next section we review the basic concepts of game theory that we require for our analysis. In Section 3 we define the particular game that models the steganographic scenario. In Section 4 we solve the game for optimal strategies and associated payoffs. In Section 5 we give two examples of the solutions deduced in Section 4 and we conclude in Section 6 with a brief discussion of future research directions.

2 Review of Matrix Games

We now briefly summarize the main ideas and results we will need from the classical theory of games. For further explanation, consult the excellent reference [5]. A *real-valued, two-player, zero-sum game* consists of two sets of pure strategies S^1, S^2 and two payoff functions $P^1 : S^1 \times S^2 \to \mathbf{R}$, $P^2 : S^1 \times S^2 \to \mathbf{R}$ such that $P^1(s,t) = -P^2(s,t)$ where $s \in S^1$ and $t \in S^2$. We will refer to player 1's payoff simply as *the payoff function* and denote it by P with the understanding that the game under consideration is zero-sum.

Without loss of generality assume that Player 1 is trying to maximize the payoff and Player 2 is trying to minimize the payoff. In all of the games studied here the sets of pure strategies will be finite. A *mixed strategy* is a probability distribution over the set of pure strategies. A pure strategy is therefore a mixed strategy with a point mass as the distribution. When the pure strategy sets are finite a mixed strategy may be written as a vector $\mathbf{x} = (x_1, ...x_j)$ where $Prob(S = s_i) = x_i$. The payoff P' for a pair of mixed strategies extends the payoff function P and is defined to be the expected payoff:

$$P'(\mathbf{x}, \mathbf{y}) = \sum_{i,j} x_i y_j P(s_i, t_j). \tag{1}$$

Since $P' = P$ for pure strategies we will simply write P for both functions. An *equilibrium* for a game G is a pair of mixed strategies, $\mathbf{x}^*$ for Player 1 and $\mathbf{y}^*$ for Player 2, such that for all mixed strategies $\mathbf{x}$,$\mathbf{y}$ we have

$$P(\mathbf{x}, \mathbf{y}^*) \leq P(\mathbf{x}^*, \mathbf{y}^*) \leq P(\mathbf{x}^*, \mathbf{y}). \tag{2}$$

Equilibria strategies are the only correct strategies to play because they have the property that neither player may benefit by deviating from an equilibrium strategy. Equilibria strategies therefore represent the solution to the game-theoretic situation. The fundamental theorem of game theory (originally proved by Von Neumann and Morgenstern, [8]) states that every finite, zero-sum game has at least one equilibrium in mixed strategies and that the payoffs at each equilibria are the same. This number is called the *value* of the game.

3 A Steganological Game

Consider the following scenario. An individual wishes to covertly communicate by hiding messages in the noise of some data files. Suppose all data files sent by this individual, say Player 1, pass through some gateway, perhaps a computer server, which is under the control of Player 2. Player 2 wants to automatically introduce noise into *all* data files passing through the gateway in order to disrupt any such covert communication. Player 2 does not examine the files individually to try to find hidden information. Rather, the disruption is completely automatic.

We now introduce a game to model this steganographic scenario and in the next section we solve for the optimal mixed strategies. The scenario is as follows and utilizes the assumptions discussed above. The data is a pseudorandom sequence of bits and is to be hidden in an N-pixel, 2^l-level greyscale image. Both players will modify the values of bits throughout the image under the constraint that the total amount of distortion must remain less than some known constant, say d. The game could easily be modified to permit different levels of distortion for the Hider and Attacker. We use the most simple model of distortion possible in that changing the value of a least significant bit is 1 unit of distortion, modifying a next-to-least significant bit is 2 units of distortion, etc., until finally modifying a most significant bit is 2^{l-1} units of distortion. We emphasize that by using different distortion measures, more sophisticated models of the effects of modifying an image may be accommodated in this game-theoretic framework. See the section on further work for a discussion of variations on this measure.

Player 1 is the data Hider and seeks to maximize the amount of hidden data whereas Player 2 is the Attacker and seeks to minimize the amount of hidden data communicated by introducing noise. A pure strategy for Player 1 is an l-tuple of nonnegative real numbers $\mathbf{x} = (x_0, x_1, ..., x_{l-1})$ such that

$$\sum_{i=0}^{l-1} x_i 2^{i-1} \leq d. \tag{3}$$

Given a strategy $\mathbf{x}$, x_i is the number of bits that the Hider will store in the set of i^{th} lowest order bits in the image. Note that x_i is allowed to be a noninteger, whereas in actual fact one must store an integer number of bits. We will see that in general we require nonintegers in order to obtain equilibria but that utilizing integer valued vectors in practical situations is an acceptable approximation. The locations of the hidden bits within a set are chosen pseudorandomly and are therefore uniformly distributed over all N bits in position i. Because half of the original bits will, on average, agree with the bits to be hidden, hiding x_0 bits in the low order bits will on average result in $x_0/2$ units of distortion. This observation gives rise to the distortion inequality. Similarly, a pure strategy for Player 2 is an l-tuple of nonnegative real numbers $(y_0, y_1, ..., y_{l-1})$ such that

$$\sum_{j=0}^{l-1} y_j 2^j \leq d. \tag{4}$$

The distortion here is different since Player 2 will flip all y_0 of the low order bits. Again the locations are chosen pseudorandomly and are uniformly distributed throughout all N possible choices. Conceptually what we now have are l independent *binary symmetric channels* [4] where the probability for a hidden bit to be flipped in channel C_i is y_i/N. Notice that both pure strategy sets are finite and therefore mixed strategies may be written as vectors, $\mathbf{X}$,$\mathbf{Y}$.

Let us now consider the construction of the payoff function for this game. Suppose Z is a discrete random variable that takes the value z_i with probability p_i, i.e. $P(Z = z_i) = p_i$. The Shannon entropy [4] of Z is given by:

$$H(Z) = \sum_i p_i \log(\frac{1}{p_i}). \tag{5}$$

If the log is taken base 2 than H has units of bits. If Z takes on only two possible values with probabilities p_1 and $1-p_1$ then we abbreviate the entropy by writing

$$H(p_1) = -p_1 \log(p_1) - (1 - p_1)\log(1 - p_1). \tag{6}$$

The channel capacity for a binary symmetric channel with bit error probability p is $1 - H(p)$ [4]. The channel capacity is the asymptotic average number of bits communicated through the channel per bit sent through the channel and this limit is approached by using error-correcting codes. For each channel C_i, $0 \leq i \leq l-1$, the Hider sends x_i bits and the probability for a bit error due to the noise introduction by the Attacker is y_i/N. Therefore the payoff function for the game is given by:

$$P(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{l-1} x_i(1 - H(\frac{y_i}{N})). \tag{7}$$

This represents the total number of bits that are communicated on average utilizing error correcting codes to compensate for the distortion introduced by the Attacker.

4 Equilibria for the Game

We now wish to solve the above game for equilibria. Recall that these are strategies such that assuming the other Player is playing his optimal strategy, the given strategy is optimal, i.e. no improvement is possible by deviating. Since a pure strategy is a tuple, $\mathbf{x} = (x_0, ..., x_{l-1})$, we will write mixed strategies, i.e. probability distributions over the set of pure strategies, as vectors with capital Latin letters, $\mathbf{X}$, and $P(\mathbf{X} = \mathbf{x}) = \mathbf{X}(\mathbf{x})$. It turns out that this game has an equilibrium in pure strategies. Therefore in the following derivation we proceed by assuming the equilibrium we seek consists of pure strategies and our subsequent calculations validate this assumption.

To solve for an equilibrium, recall the payoff function:

$$P(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{l-1} x_i (1 - H(\frac{y_i}{N})). \tag{8}$$

Let

$$M_i = 1 - H(\frac{y_i}{N}) \tag{9}$$

be the channel capacity for the i^{th} channel. Using the fact that neither player desires to deviate from an equilibrium strategy let us solve for an optimal strategy for the Attacker, $\mathbf{y} = (y_0, y_1, ..., y_{l-1})$. Fix all components of $\mathbf{x}$ except for x_j and x_k where $0 \leq j < k \leq l-1$ and consider distributing the remaining distortion

$$d_{jk} = d - \sum_{i \neq j,k} 2^{i-1} x_i \tag{10}$$

between these two components. We may now consider the *partial payoff function* P_{jk} which is the portion of the payoff function, Equation 8, concerning only the two channels C_j and C_k. Consider the following equation for the partial payoff function as a function of x_j:

$$P_{jk}(x_j) = x_j M_j + x_k M_k = x_j M_j + \frac{d_{jk} - 2^{j-1} x_j}{2^{k-1}} M_k. \tag{11}$$

In order to find an equilibrium we must insure that the Hider does not profit from readjusting x_j and x_k for a fixed $\mathbf{y}$. Therefore we have

$$\frac{dP_{jk}}{dx_j}(x_j^*) = M_j - 2^{j-k} M_k = 0 \tag{12}$$

for an equilibrium x_k^* and thus

$$M_j = 2^{j-k} M_k. \tag{13}$$

If there is enough total available distortion, and this point will become more clear below, then this equation must hold for all $0 \leq j < k \leq l-1$ and so for any equilibrium strategy $\mathbf{y}$ we have

$$2^{k-j}(1 - H(\frac{y_j}{N})) = 1 - H(\frac{y_k}{N}). \tag{14}$$

In particular if $k = j+1$ then we have $2M_j = M_{j+1}$. For fixed d and N Equations 4 and 14 determine the equilibrium strategy $\mathbf{y}^*$ and this can be solved numerically.

We now sketch the form of these solutions. Notice that in order to satisfy $2M_k = M_{k+1}$ for all $0 \leq k \leq l-2$ requires a critical amount of allowable distortion $d_c(l, N)$. For suppose that $M_{l-1} = 1$, i.e. the Attacker introduces no noise in channel C_{l-1}, the most expensive channel in which to introduce noise. Then the above channel capacity constraints reduce to $M_{l-2} = 1/2$, $M_{l-3} = 1/4$,...., and in general

$$M_i = 2^{-(l-1-i)}. \tag{15}$$

Define a pseudoinverse to the binary symmetric channel capacity function $Cap(p) = 1 - H(p)$ as $Cap^{-1}(M) =$ unique p such that $Cap(p) = M$ and $0 \leq p \leq 1/2$. Then $Cap^{-1} : [0,1] \rightarrow [0,1/2]$ and we have

$$d_c(l, N) = \sum_{i=0}^{l-1} 2^i N Cap^{-1}(2^{-(l-1-i)}). \tag{16}$$

This is the minimal amount of distortion necessary in order for the Attacker to simultaneously satisfy Equation 14 for each i. If $d \geq d_c(l, N)$ then the Attacker's strategy is to distribute d among all channels subject to these constraints. As mentioned previously, these constraints determine the noise allocation subject to the distortion limitation. Notice also that

$$d_{total} = \sum_{i=0}^{l-1} 2^i N/2 \tag{17}$$

is the amount of distortion necessary to reduce all channels to zero capacity. This is equivalent to flipping each bit of each pixel with probability 1/2, i.e. randomizing the image. Of course in a practical situation such extreme distortion would be prohibited.

Now what is the proper strategy if $d < d_c$? Since it costs twice as much for the Hider to place data in C_{i+1} as it does in C_i, there is an implicit reduction in the channel capacity of the higher order channels. Define a quantity called the *effective channel capacity*

$$M'_i = 2^{-i} M_i. \tag{18}$$

which is the reciprocal of the number of units of distortion required on average to communicate one logical bit of information in channel C_i. Notice that if $y_i = 0$ then $M'_i = 2^{-i}$. Also note that Equation 14 is equivalent to

$$M'_i = M'_j \tag{19}$$

for all i, j. If $d < d_c$ then the Attacker's optimal strategy consists of equalizing as many of the lowest order effective channel capacities as allowed by the distortion limit. In other words, the Attacker first introduces noise into C_0, reducing M'_0 until reaching the point $M'_0 = M'_1 = \frac{1}{2}$. If he has not reached the distortion limit then he continues to introduce noise into C_0 and now also C_1, maintaining the relation $M'_0 = M'_1$ until he reaches the point $M'_0 = M'_1 = M'_2 = \frac{1}{4}$ and so on until the limit d is reached.

Let us now solve for the Hider's optimal strategy. For a fixed $\mathbf{x}$ consider the partial payoff functions as functions of y_j:

$$P_{jk}(y_j) = x_j(1 - H(\frac{y_j}{N})) + x_k(1 - H(\frac{\frac{d_{jk} - 2^j y_j}{2^k}}{N})) \tag{20}$$

for $0 \leq j < k \leq l - 1$ where

$$d_{jk} = d - \sum_{i \neq j,k} 2^i y_i \tag{21}$$

is again a constant of remaining distortion. In order for $\mathbf{y}^*$ to be an equilibrium strategy it is necessary the Attacker does not profit from readjusting and therefore we have

$$\frac{dP_{jk}}{dy_j}(y_j^*) = 0. \tag{22}$$

Taking the derivative and carrying through the calculation yields:

$$\frac{x_j}{x_k} = \frac{2^{j-k} \log \frac{p_k^*}{1-p_k^*}}{\log \frac{p_j^*}{1-p_j^*}} \tag{23}$$

where $p_i^* = y_i^*/N$. For fixed N, d and from $\mathbf{y}^*$ derived above we can also solve this numerically. This process yields pure strategies $\mathbf{x}^*$ and $\mathbf{y}^*$ with nonnegative, possibly noninteger coordinates. The total number of logical bits communicated is given by

$$P(\mathbf{x}^*, \mathbf{y}^*) = \sum_{i=0}^{l-1} x_i^*(1 - H(\frac{y_i^*}{N})). \tag{24}$$

The fact that the components may not be integers requires comment. Recall that a pure strategy is an l-tuple of nonnegative *real* numbers, representing the number of bits to be stored in a particular channel. However a noninteger coordinate in a tuple would evidently have no practical interpretation. For example, for the strategy $(5, 10.5, ..., \pi)$ we have no way of actually storing 10.5 bits in the N next-to-least significant bits or π bits in the N most significant bits. In practice this issue is insignificant as the channel capacities are continuous functions of the strategy components. Therefore neither player can profit significantly from the other player's need to approximate a noninteger strategy with integer components. Asymptotically as N goes to infinity the difference in the payoffs resulting from using the exact equilibrium and an integer approximation goes to zero.

5 Two Numerical Examples

Let us consider two concrete examples of the preceding game. Numerical solutions to all sets of nonlinear equations were obtained by the use of simple Mathematica programs. In the first example we will have $d < d_c$ and in the second example we have $d > d_c$. Consider an 8-bit, 256 level greyscale image with 10^6 pixels. For the first example suppose we set the distortion to be equivalent to freely replacing the two least significant bits in each pixel. Then $d = 3 \times 10^6$. Solving Equation 16 yields $d_c(8, 10^6) \approx 2.4 \times 10^7$. Solving Equations 14 and 4 numerically yields the following solutions for the optimal Attacker strategy and is presented in Table 1. Solving Equations 23 and 4 numerically yields the following solutions for the optimal Hider strategy and is presented in Table 2. These solutions and Equation 24 give

$$P(\mathbf{x}^*, \mathbf{y}^*) = 346264. \tag{25}$$

Table 1: Optimal Strategy for Attacker ($d = 3 \times 10^6$, $N = 10^6$)

Channel	0	1	2	3	4	5	6	7
y_i	359526	302710	224980	123138	9376	0	0	0
p_i	.359	.303	.225	.123	.01	0	0	0
M_i	0.058	0.115	.231	.462	.923	1	1	1
M_i'	.058	.058	.058	.058	.058	$2^{-5} = .031$	2^{-6}	2^{-7}

Table 2: Optimal Strategy for Hider ($d = 3 \times 10^6$, $N = 10^6$)

Channel	0	1	2	3	4	5	6	7
x_i	97114	134406	181347	228526	192530	0	0	0

For the second example let $d = 5 \times 10^7$. This level of distortion corresponds to treating between 5 and 6 of the lowest bits in each byte as replaceable. Solving Equations 14 and 4 numerically yields the following solutions for the optimal Attacker strategy and is presented in Table 3. Solving Equations 23 and 4 numerically yields the solutions for the optimal Hider strategy and is presented in Table 4. These solutions and Equation 24 give

$$P(\mathbf{x}^*, \mathbf{y}^*) = 381120. \tag{26}$$

We now make a final interesting observation. Consider $P(x^*, y^*)$, the total amount of information communicated as a function of d. If $d = 0$ then clearly $P = 0$ because the Hider cannot alter any bits and if $d = d_{total}$ then $P = 0$ because the Attacker can randomize the data set. Numerical analysis suggests that P is concave, achieves a maximum at $d_{max} \approx 37N = 3.7 \times 10^7$ and $P(d_{max}) \approx .405N = 405000$.

Table 3: Optimal Strategy for Attacker ($d = 5 \times 10^7$,$N = 10^6$)

Channel	0	1	2	3	4	5	6	7
y_i	463672	448647	427440	397568	355659	297361	217774	114101
p_i	.463	.449	.427	.398	.356	.297	.218	.114
M_i	.004	.008	.015	.030	.061	.122	.244	.488
M_i'	.004	.004	.004	.004	.004	.004	.004	.004

Table 4: Optimal Strategy for Hider ($d = 5 \times 10^7$,$N = 10^6$)

Channel	0	1	2	3	4	5	6	7
x_i	52624	74324	104830	147456	206254	285075	383418	478426

6 Further Work

One may imagine many variations on the game formulation studied here. For a more general formulation of the problem see [6]. The game presented in the present work is a simultaneous move game and models the situation whereby the Attacker does not have the privilege of witnessing the Hider's move before deciding on a strategy. In a practical situation this may occur if, for example, the Attacker is in control of a server through which a large number of files with potential hidden data pass. The Attacker may set up an automated system to intentionally introduce noise into all data files to disrupt the covert channels. Another scenario occurs if the Attacker examines files individually and then introduces noise based on this analysis. Presumably this would decrease the channel capacity of the covert channels. This latter scenario would be modeled by a game in which the Attacker moves *after* the hider.

Another area which needs further consideration is the distortion measure. In fact, the actual distortion introduced by changing bits is a difficult problem, probably requiring consideration of the details of human vision. Large areas of uniform color are especially sensitive to even small local changes. Therefore our analysis is probably more relevant to images without these large uniformities. Quantitative analysis of these problems will be an important advance in steganalysis.

7 Acknowledgments

I would like to thank Mike Murphy, Mike Neergaard, Ted Handel, Brook Sandford, and the anonymous referees for their corrections and comments.

References

[1] Ross Anderson editor. *Information Hiding. First International Work-*

shop. Cambridge, U.K., May/June, 1996. Proceedings. Lecture Notes In Computer Science Series Number 1174. Springer-Verlag, 1996.

[2] W. Bender, D. Gruhl, N. Morimoto, A. Lu. "Techniques For Data Hiding." IBM Systems Journal Vol. 35, 1996, pp.313 - 336.

[3] I. Cox, J. Kilian, Tom Leighton, Talal Shamoon. "A Secure, Robust Watermark for Multimedia" in [1].

[4] Robert J. McEliece. *The Theory of Information and Coding.* Addison-Wesley, 1977.

[5] Guillermo Owen. *Game Theory.* Academic Press, 1995.

[6] Joseph A. O'Sullivan, Pierre Moulin, J. Mark Ettinger. *Information Theoretic Analysis of Steganography.* 1998 IEEE International Symposium on Information Theory.

[7] M. Sandford, J. Bradley, T. Handel. "The Data Embedding Method." Los Alamos Technical Report 9LA-95-2246UR.

[8] J. Von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior.* Princeton University Press, 1944, 1947.