A GAME-THEORETIC ANALYSIS OF CONTENT-ADAPTIVE STEGANOGRAPHY WITH INDEPENDENT EMBEDDING

Pascal Schöttle
University of Münster

Aron Laszka
Budapest University of Technology and Economics

Benjamin Johnson
University of California, Berkeley

Jens Grossklags
Pennsylvania State University

Rainer Böhme
University of Münster

ABSTRACT

We provide a game-theoretic analysis of a scenario from the field of content-adaptive steganography. Alice, a steganographer, wants to embed a secret message into a random binary sequence with a known distribution in which the value of each position is independently but non-identically distributed. Eve, a steganalyst, observes the sequence and wants to determine whether it contains a hidden message. Alice is allowed to flip binary values independently at random, with the constraint that the expected number of changes is a fixed constant. Eve may choose to classify each sequence as either unmodified (cover) or modified (stego). The payoff for Eve in the game is the probability that her classification is correct; and the payoff for Alice is the probability that Eve's classification is incorrect, so that the game is constant-sum.

We show that Eve's best response strategy in this game can be expressed as a linear aggregation threshold formula similar to those used in practical steganalysis. We give a general formula for Alice's best response strategy; and we compute explicit pure strategy equilibria for the special case of changing one bit in a length-two sequence.

1. INTRODUCTION

1.1. Steganography and Steganalysis

Steganography is the art of hiding messages in a communication channel; while its counterpart, steganalysis, is the art of determining whether a communication channel has been modified to contain a message. Much research has been done in this area, most prevalently in the context of digital multimedia [1]. In that context, the objective of a steganographer is to hide a message in a multimedia file, e.g., by adjusting some of the pixels in a JPEG image. The objective of the steganalyst is then to detect whether the image has been modified to encode some message [2].

A useful heuristic when hiding messages is to take advantage of noisy regions in the communication channel. For example, in digital images there are often regions of high color variance where a color change in some pixel is unlikely to be noticed [3]. We say that a steganographic scheme is content-adaptive if it takes advantage of channel noise to decide where the message is embedded.

In studying content-adaptive steganography, it is useful (and common [4]) to give considerable detection power to the steganalyst, including knowledge of the distribution of the unmodified communication channel. For a fixed communication channel, we refer to this distribution as the *cover distribution*, and the set of possible communications as the *cover source*. In digital image steganography, for example, the cover source would be the set of all possible images that might be used to hide a message; and the cover distribution would assign a probability to each such image. In practice, the cover distribution for a given communication channel is almost never known [3]. On the other hand, when we are most interested in developing good content-adaptive embedding strategies, giving more power to the detector is a useful conservative assumption.

1.2. Modeling Approach

To abstract away from the communication channel details, we consider a cover source consisting of binary sequences of some fixed length. In the cover distribution, the value of the sequence at each position is independently but non-identically distributed. The distribution is such that the embedding suitability of a given position does not affect the suitability of any other position, but some positions are more suitable for embedding than others.

In the context of images, one can think of our modeling framework as aggregating noisy regions together, so that a binary choice indicates whether or not the stenographer modified that region. More generally, the model tries to capture the most essential features of content-adaptive embedding strategies by focusing only on the communication channel's noise levels.

A steganographer must choose some positions in which to hide her message. Practical and theoretical considerations serve to motivate appropriate restrictions on this embedding strategy. First, due to the non-deterministic manner in which a real-world encoding scheme converts a message to an embedding [5], we allow the steganographer to use randomized embedding strategies and constrain only the expected value of the embedding size. Second, due to abstractions of noisy regions into independent binary variables, any advantage gained

from introducing new correlations between positions would not translate well to the original motivation. Therefore we require the steganographer's randomized strategy to embed independently in each position. With these changes, the distribution of modified covers (a.k.a. the *stego distribution*), retains the independence property between positions.

1.3. Game Theory

Enter game theory: the study of strategic choices and their consequences. We consider a two-player game between Alice, the steganographer, and Eve, the steganalyst.

Alice wants to hide messages of expected length k into a cover source of length-N binary sequences; so she may choose any set of N probabilities that sum to k. Each a_i represents the probability that Alice changes the value of the sequence at position i.

Eve wants to optimally classify sequences as either containing a message (stego) or not (cover); so she may choose a probability for each length-N binary sequence. Each $e(x)$ represents the probability that Eve classifies the sequence x as stego.

To formalize the game payoff, we suppose that cover sequences and stego sequences are equally likely to be seen by Eve. The game payoff is then determined by the probability over all cover/stego possibilities, binary sequences, embedding probabilities, and classifier probabilities, that Eve correctly determines from which distribution the sequence was drawn. Alice's payoff is the probability that Eve's classifier is incorrect, so that the sum of the two players' payoffs is 1.

1.4. Paper Outline

The rest of the paper is organized as follows. We review related work in Section 2. In Section 3, we present our equilibria results. We show numerical illustrations of these results in Section 4, and we conclude in Section 5.

2. RELATED WORK

The combination of steganography and game theory dates back to 1998 [6]; while the combination of game theory and content-adaptive steganography is more recent, with the first such study appearing in [7]. Our model draws inspiration from the game-theoretic setup in [8], which is also based on content-adaptive steganography. Game-theoretic analyses in this area are well-motivated by the fact that almost all recently published steganographic algorithms are content-adaptive, e. g. [9].

3. MODEL ANALYSIS

In this section, we derive analytical results on the minimax strategies and the existence of pure-strategy Nash equilibria. We begin by introducing some notation to aid the analysis.

3.1. Notation

Let X denote the random variable representing the observed binary sequence, and let Y denote the random variable taking values in $\{C, S\}$ that represents whether the sequence was drawn from the cover (C) or stego (S) distribution.

In the cover distribution, X is determined by a monotonically increasing sequence $\langle f_i \rangle_{i=1}^{N}$ from $(\frac{1}{2}, 1)$, where f_i gives the probability that $X_i = 1$. Since the positions are independent,

$$\Pr[X = x | Y = C] = \prod_{i:x_i=1} f_i \cdot \prod_{i:x_i=0} (1 - f_i) \,. \tag{1}$$

In the stego distribution, Alice flips the value of the sequence in each position i with probability a_i, so that $X_i = 1$ with probability $f_i(1-a_i)+(1-f_i)a_i$. To simplify notation, we introduce the bias sequence $\tilde{f}_i = 2f_i - 1$. Then we have $f_i(1 - a_i) + (1 - f_i)a_i = f_i - a_i\tilde{f}_i$. Since positions in the stego distribution are also independent,

$$\Pr[X = x | Y = S] = \prod_{i:x_i=1} (f_i - a_i\tilde{f}_i) \cdot \prod_{i:x_i=0} (1 - f_i + a_i\tilde{f}_i) \,. \tag{2}$$

3.2. Eve's Best Response

Given a fixed embedding strategy for Alice, Eve must classify each sequence as cover or stego. Since she knows both of these distributions, she can perform a likelihood ratio test to determine her optimal decision [1]. This test gives a deterministic decision rule whenever the two likelihoods are unequal. When they are equal for a given sequence x, Eve's decision at that x does not affect the probability that her classifier is correct; so in this case, any randomized function over cover and stego is a best response.

Theorem 1. *A best response strategy of Eve is given by the decision rule*

$$D_{best}(x) = \begin{cases} C & \text{if } \sum_i w_i x_i > \tau \\ S & \text{if } \sum_i w_i x_i < \tau \\ C \text{ or } S & \text{if } \sum_i w_i x_i = \tau \end{cases} \tag{3}$$

where for each i,

$$w_i = \log \frac{f_i(1 - f_i + a_i\tilde{f}_i)}{(f_i - a_i\tilde{f}_i)(1 - f_i)} \quad \textit{and} \tag{4}$$

$$\tau = \sum_i \log \frac{1 - f_i + a_i\tilde{f}_i}{1 - f_i} \,. \tag{5}$$

Proof. Given a sequence x, Eve's best response selects the most likely distribution from which x was drawn. Her optimal choice can thus be expressed as

$$D_{best}(x) = \begin{cases} C & \text{if } \frac{\Pr[Y=C|X=x]}{\Pr[Y=S|X=x]} > 1 \\ S & \text{if } \frac{\Pr[Y=C|X=x]}{\Pr[Y=S|X=x]} < 1 \\ C \text{ or } S & \text{if } \frac{\Pr[Y=C|X=x]}{\Pr[Y=S|X=x]} = 1. \end{cases}$$

The condition for cover (C) can be expressed using f and a as follows:

$$\begin{aligned}
1 <& \frac{\Pr[Y=C|X=x]}{\Pr[Y=S|X=x]} \\
=& \frac{\Pr[Y=C]\Pr[X=x|Y=C]}{\Pr[Y=S]\Pr[X=x|Y=S]} \\
=& \frac{\frac{1}{2}\cdot\prod_i \Pr[X_i=x_i|Y=C]}{\frac{1}{2}\cdot\prod_i \Pr[X_i=x_i|Y=S]} \\
=& \prod_{i:x_i=1} \frac{f_i}{f_i - a_i\tilde{f}_i} \cdot \prod_{i:x_i=0} \frac{1-f_i}{1-f_i+a_i\tilde{f}_i} \\
0 <& \sum_{i:x_i=1} \log \frac{f_i}{f_i - a_i\tilde{f}_i} + \sum_{i:x_i=0} \log \frac{1-f_i}{1-f_i+a_i\tilde{f}_i} \\
=& \sum_i \left(x_i \log \frac{f_i}{f_i - a_i\tilde{f}_i} + (1-x_i)\log \frac{1-f_i}{1-f_i+a_i\tilde{f}_i} \right) \\
=& \sum_i x_i \log \frac{f_i(1-f_i+a_i\tilde{f}_i)}{(f_i-a_i\tilde{f}_i)(1-f_i)} + \sum_i \log \frac{1-f_i}{1-f_i+a_i\tilde{f}_i} \\
\Leftrightarrow& \sum_i x_i \log \frac{f_i(1-f_i+a_i\tilde{f}_i)}{(f_i-a_i\tilde{f}_i)(1-f_i)} \geq \sum_i \log \frac{1-f_i+a_i\tilde{f}_i}{1-f_i} \,.
\end{aligned}$$

□

3.3. Alice's Best Response

Given a fixed (potentially randomized) classifier for Eve, Alice wants to choose an embedding strategy that maximizes the error probability of this classifier; but since her strategy cannot affect the classifier's false positive rate (on cover inputs), she may concentrate her efforts on maximizing the classifier's false negative rate. Formally, if $e(x)$ is Eve's probability for classifying x as stego, then Alice's best response strategy is to choose an a satisfying $\sum_i a_i = k$ and maximizing

$$\sum_{x\in\{0,1\}^N} (1-e(x))\Pr[X=x|Y=S] =$$

$$\sum_{x\in\{0,1\}^N} (1-e(x)) \prod_{i:x_i=1} (f_i - a_i\tilde{f}_i) \prod_{i:x_i=0} (1-f_i+a_i\tilde{f}_i)\,. \quad (6)$$

To get some leverage from this formula, consider Alice's best response strategy and any pair a_i, a_j that are interior values of $(0,1)$. Alice's payoff cannot increase if she adjusts her strategy by simultaneously increasing a_i and decreasing a_j (or vice versa) by the same small amount ϵ. If we consider the payoff as a function of ϵ in this manner, then for a payoff-maximizing a, the partial derivative with respect to ϵ must be zero at $\epsilon = 0$. This condition can be expressed as a formula, which constrains Alice's best response strategy for each pair (a_i, a_j) taking interior values in $(0,1)$ in terms of the remaining a_m.

$$\begin{aligned}
a_i - a_j = & \frac{\sum_x (1-e(x)) \prod_{m\neq i,j} \Pr[X_m = x_m|Y=S]}{\sum_x (1-e(x)) \prod_{m\neq i,j} \Pr[X_m = x_m|Y=S]} \\
& \cdot\Big((1-2x_i)\tilde{f}_i(x_j\tilde{f}_j + 1 - f_j) \\
& -(1-2x_j)\tilde{f}_j(x_i\tilde{f}_i + 1 - f_i)\Big) \\
& \cdot\Big(\tilde{f}_1\tilde{f}_2(1-2x_i)(1-2x_j)\Big)
\end{aligned} \quad (7)$$

This set of constraints can be solved for at least some small N. However, we conjecture that the general problem of computing Alice's best response strategy is NP-hard, and we leave this as an open problem. We illustrate the structure of the solution in the following subsection by considering the special case of two positions.

3.4. Special Case: $N = 2$, $k = 1$

For this subsection, we restrict our analysis to the case of changing a single bit ($k = 1$) in covers of length two ($N = 2$). Here Alice's strategy (a_1, a_2) can be specified by giving only one of the two probabilities, since $a_2 = 1 - a_1$; consequently, we use only the probability a_1 to specify Alice's strategy in this subsection. Eve's strategy is specified as a vector $(e(00), e(01), e(10), e(11))$.

3.4.1. Alice's Minimax Strategy

Alice's minimax strategy minimizes Eve's payoff assuming Eve is playing a best response strategy. To find this strategy, we divide Alice's strategy space into equivalence classes such that Eve's best response is the same for each element in a class. We begin by giving some lemmas that show the structure of these classes. The proofs use algebra based on the definitions, and are omitted due to shortage of space.

Lemma 1. *Eve always classifies sequence* 00 *as stego and sequence* 11 *as cover.*

Lemma 2. *Eve classifies the sequence* 01 *as cover when* $a_1 \leq \theta_1$*, and she classifies the sequence* 10 *as cover when* $a_1 \geq \theta_2$*, where*

$$\begin{aligned}
\theta_1 = & \frac{(f_1-1)\tilde{f}_2 + \tilde{f}_1(f_2-1)}{2\tilde{f}_1\tilde{f}_2} \\
& + \frac{\sqrt{\Big[(1-f_1)\tilde{f}_2 + \tilde{f}_1(1-f_2)\Big]^2 - 4\tilde{f}_1\tilde{f}_2^2(f_1-1)}}{2\tilde{f}_1\tilde{f}_2}
\end{aligned}$$

and

$$\theta_2 = \frac{f_1\tilde{f}_2 + \tilde{f}_1 f_2 - \sqrt{\Big[f_1\tilde{f}_2 + \tilde{f}_1 f_2\Big]^2 - 4f_1\tilde{f}_1\tilde{f}_2^2}}{2\tilde{f}_1\tilde{f}_2}\,.$$

Lemma 3. *It always holds that $\theta_1 < \theta_2$.*

The following theorem summarizes Eve's best response for the three equivalence classes on Alice's strategy space.

Theorem 2. *Given a fixed strategy for Alice, Eve's optimal decision for each binary sequence x is given by:*

Alice's strategy			*Eve's optimal decision*			
			$x = 00$	*01*	*10*	*11*
	a_1	$\leq \theta_1$	*S*	*C*	*S*	*C*
$\theta_1 \leq$	a_1	$\leq \theta_2$	*S*	*S*	*S*	*C*
$\theta_2 \leq$	a_1		*S*	*S*	*C*	*C*

Proof. It follows immediately from Lemmas 1, 2, and 3. □

Next, for each equivalence class, we consider Alice's payoff, assuming Eve is making an optimal decision.

Lemma 4. *Alice's payoff is increasing for $a_1 \in [0, \theta_1]$ and decreasing for $a_1 \in [\theta_2, 1]$.*

Lemma 5. *The first derivative of Alice's payoff for $a_1 \in [\theta_1, \theta_2]$ is*

$$\frac{\partial u(Alice)}{\partial a_1}\Big|_{\theta_1 \leq a_1 \leq \theta_2} = -4a_1\tilde{f}_1\tilde{f}_2 + 2\Big(f_1\tilde{f}_2 + \tilde{f}_1(f_2 - 1)\Big)$$

and the second derivative is $-4\tilde{f}_1\tilde{f}_2$.

Theorem 3. *Alice's minimax strategy is*

$$a_1 = \begin{cases} a_{max} & when \quad a_{max} \leq \theta_2 \\ \theta_2 & when \quad \theta_2 < a_{max}\ , \end{cases} \tag{8}$$

where a_{max} denotes $\frac{f_1\tilde{f}_2 + \tilde{f}_1(f_2-1)}{2\tilde{f}_1\tilde{f}_2}$.

Proof. From Lemma 4, we have that Alice's minimax strategy satisfies $\theta_1 \leq a_1 \leq \theta_2$. This strategy must be a local maximum for her payoff over $[\theta_1, \theta_2]$. Since the second derivative of the payoff is always below zero in this region, we can find the local maximum by letting the first derivative be equal to zero and solving the equation for a_1, which gives us a_{max}.

- It can be shown that $a_{max} \geq \theta_1$.
- If $a_{max} \leq \theta_2$, the local maximum is attained at a_{max}. Thus, Alice's minimax strategy is $(a_{max}, 1 - a_{max})$.
- If $\theta_2 < a_{max}$, the local maximum is attained at the endpoint θ_2. Thus, Alice's minimax strategy is $(\theta_2, 1 - \theta_2)$.

□

3.4.2. Nash equilibria

We next characterize the equilibria of the game. We start by giving conditions for when there is an equilibrium in which Eve uses a deterministic classifier.

Theorem 4. *A Nash equilibrium with a deterministic strategy for Eve exists if and only if $a_{max} \leq \theta_2$.*

Proof. First, it is easy to see that the strategy pair (S, S, S, C) and $(a_{max}, 1 - a_{max})$ is an equilibrium when $a_{max} \leq \theta_2$ because both strategies are best responses. Second, we have to show that no equilibrium with a deterministic strategy for Eve can exist if $\theta_2 < a_{max}$:

- Alice's best response to the strategy (S, C, S, C) is $a_1 = 1$; however, Eve's best response strategy to $(1, 0)$ is not (S, C, S, C), but (S, S, C, C).
- Alice's response to (S, S, S, C) is $a_1 = a_{max}$; however, since $a_{max} \notin [\theta_1, \theta_2]$, Eve's response is not (S, S, S, C), but either (S, C, S, C) or (S, S, C, C).
- Finally, Alice's best response to (S, S, C, C) is $a_1 = 0$; however, Eve' response to $(0, 1)$ is (S, C, S, C).

□

Next we show that an equilibrium always exists if Eve can use probabilistic strategies.

In the case of $a_{max} \leq \theta_2$, the strategy pair (S, S, S, C), $(a_{max}, 1 - a_{max})$ is an equilibrium. Thus, in this case, Eve can use the probabilistic strategy that chooses the deterministic strategy (S, S, S, C) with probability 1. Consequently, we only have to a find a mixed strategy for Eve in the case of $a_{max} > \theta_2$.

Theorem 5. *Eve's minimax strategy is*

$$e(00) = 1$$
$$e(01) = 1$$
$$e(10) = \begin{cases} 1 & \text{if} \quad a_{max} \leq \theta_2 \\ \frac{\tilde{f}_1}{\sqrt{(f_1 - f_2)^2 - 4f_1\tilde{f}_1\tilde{f}_2(f_2 - 1)}} & \text{otherwise.} \end{cases}$$
$$e(11) = 0$$

Proof. Eve is playing a minimax strategy when she forces Alice to play her minimax strategy as a best response. We obtain the probability for 10 by using brute force and single-variable calculus to compute Alice's best response as a function of this probability and equating it with her minimax strategy. □

4. NUMERICAL ILLUSTRATIONS

In this section, we give numerical illustrations for some results with $N = 2$ and $k = 1$. First, Figure 1 depicts the probability that Eve classifies the sequence 10 as stego in her minimax strategy, as a function of the cover predictability descriptor f. The dotted black line gives the border between the regions $e(10) = 1$ (white area) and $e(10) < 1$, where darker areas indicate lower values.

Figure 2 shows Eve's classification error rates as a function of a_1 for two different examples of f. The example f in Figure 2(a) yields a deterministic strategy equilibrium, while the f in Figure 2(b) yields a randomized strategy equilibrium. Both figures reveal that neither the false positive rate

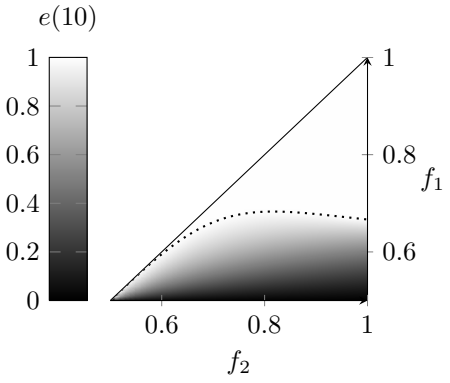
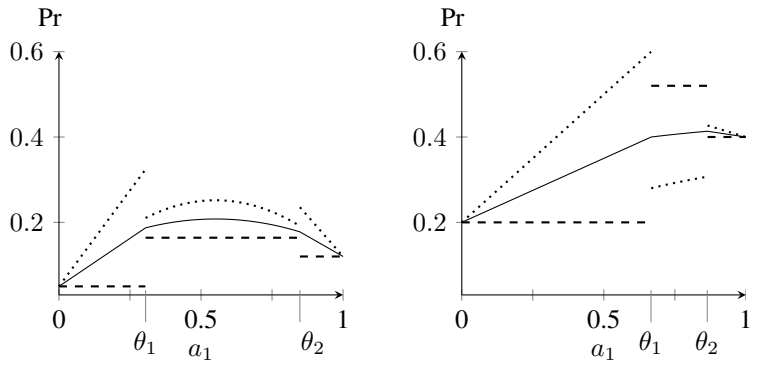


Fig. 1. The value of $e(10)$ in Eve's minimax strategy as function of f.



(a) $f_1 = 0.88, f_2 = 0.95$ (b) $f_1 = 0.6, f_2 = 0.8$

Fig. 2. Eve's false positive rate (dashed line), Eve's false negative rate (dotted line), and Alice's payoff (solid line) as a function of a_1.

nor the false negative rate is continuous, although Alice's payoff (which is half the sum of these rates) is continuous. The discontinuities occur at the two values θ_1 and θ_2 where Eve switches her optimal strategy (see Lemma 2).

5. CONCLUSION

Motivated by the objective of finding good content-adaptive embedding strategies, we analyzed a two-player game in which Alice could flip an expected number of k bits in a cover source consisting of N independently-distributed but not identically-distributed positions; and her objective was to cause maximum failure probability in Eve's optimal classifier. We found that Eve's best response to Alice could be expressed as a linear inequality in the cover positions; and that Alice's best response to Eve was determined by maximizing a weighted sum over all cover realizations. We addressed the basic structure of this problem by tackling the case of two positions, and for that special case we described the game's pure-strategy Nash equilibria.

Many open questions remain. We conjecture that determining Alice's best response strategy in this setting is NP-hard. For reasons not entirely unrelated to this conjecture, we leave more detailed structural analysis of larger covers to future work.

Acknowledgements

We gratefully acknowledge the support of the Penn State Institute for Cyber-Science. The first and last authors would like to thank the German National Science Foundation (DFG) under grant "Sichere adaptive Steganographie", and the second author would like to thank the Campus Hungary Program, respectively, for supporting research visits to the Pennsylvania State University.

6. REFERENCES

[1] Jessica Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[2] Gustavus J. Simmons, "The prisoners' problem and the subliminal channel," in *Advances in Cryptology – CRYPTO '83*, pp. 51–67. Plenum Press, 1983.

[3] Rainer Böhme, *Advanced Statistical Steganalysis*, Springer, 2010.

[4] Tomáš Filler, Andrew D. Ker, and Jessica Fridrich, "The square root law of steganographic capacity for markov covers," in *Media Forensics and Security*. 2009, vol. 7254, p. 725408, SPIE.

[5] Jessica Fridrich, "Minimizing the embedding impact in steganography," in *MM & Sec '06: Proceedings of the 8th Workshop on Multimedia and Security*, 2006, pp. 2–10.

[6] Mark Ettinger, "Steganalysis and game equilibria," in *Information Hiding*. 1998, vol. 1525 of *LNCS*, pp. 319–328, Springer.

[7] Pascal Schöttle and Rainer Böhme, "A game-theoretic approach to content-adaptive steganography," in *Information Hiding*. 2012, vol. 7692 of *LNCS*, pp. 125 – 141, Springer.

[8] Benjamin Johnson, Pascal Schöttle, and Rainer Böhme, "Where to hide the bits?," in *GameSec*. 2012, vol. 7638 of *LNCS*, pp. 1–17, Springer.

[9] Vojtěch Holub and Jessica Fridrich, "Designing steganographic distortion using directional filters," in *IEEE WIFS*, 2012, pp. 234–239.