



Informatik

The Role of Side Information in Steganography

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich
Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von

Pascal Schöttle

aus Freudenstadt

– 2014 –

Dekan:	Prof. Dr. Martin Stein
Erster Gutachter:	Prof. Dr.-Ing. Rainer Böhme
Zweiter Gutachter:	Prof. Dr. Herbert Kuchen
Tag der mündlichen Prüfung:	11. Juli 2014
Tag der Promotion:	11. Juli 2014

*Face your life
Its pain, its pleasure,
Leave no path untaken*

– Neil Gaiman

Abstract

The stated goal of digital steganography is to hide the mere existence of a secret communication in digital media. In this thesis, we motivate the research in digital steganography and give a brief comparison of the basic building blocks of a cryptographic and a steganographic communication system. The most common approach in steganography is to hide the secret messages in empirical cover objects, such as digital images. We formally define the key components of such a system and are the first to define *Steganographic Side Information* (SSI). Side information in steganography differs from side information as it is used in cryptography. Our definition of SSI captures all relevant properties of such information in a steganographic system.

Furthermore, we define *uncertainty*, as the lack of knowledge of an attacker on a steganographic system, termed steganalyst. We give information-theoretic intuitions for these definitions and explain the common usage of SSI. Almost all recently proposed steganographic schemes use some kind of SSI to identify supposedly better suitable areas within cover objects and confine the embedding changes to these areas.

We develop a targeted attack on the strategy that uses only the best suitable positions and call such a strategy naïvely adaptive. With our targeted attack on four widely used variants of SSI, we show that such a scheme is almost perfectly detectable by a steganalyst who can (partially) reconstruct the SSI from the stego object.

Motivated by this result, we argue why the competition between a steganographer who uses SSI in her embedding function and a steganalyst who tries to reconstruct the SSI must be framed with means of game theory, a well established mathematical framework to model two or more rational parties that act strategically. We then present a game-theoretical framework that captures all relevant properties of a steganographic system with SSI available.

We instantiate our framework with five different models and solve each of these models for their game-theoretically optimal strategies. We compare this strategies to their information-theoretically optimal counterparts and observe that the strategies differ, with the exception of degenerate corner cases. Inspired by our solutions, we give a new paradigm for secure adaptive steganography, the so-called *equalizer embedding strategies*.

Keywords: *Steganography, Side Information, Game Theory, Security*

Zusammenfassung

Das erklärte Ziel von digitaler Steganographie ist es, den Umstand einer geheimen Kommunikation in digitalen Medien zu verstecken. In dieser Doktorarbeit wird die Forschung im Bereich der digitalen Steganographie motiviert und ein Vergleich zur besser bekannten Disziplin der Kryptographie gezogen. Der übliche Ansatz im Bereich der Steganographie ist es, die geheime Nachricht in einem empirischen Trägermedium zu verstecken. In dieser Arbeit definieren wir den Begriff der *Steganographischen Seiteninformation* (SSI). Unsere Definition von SSI umfasst alle wichtigen Eigenschaften von Seiteninformation auf dem Gebiet der Steganographie.

Zusätzlich geben wir eine formale Definition von *Unsicherheit*, dem fehlenden Wissen auf der Seite eines Angreifers (Steganalysten). Wir begründen beide Definitionen informationstheoretisch und erklären den Einsatz von SSI in steganographischen Systemen. Fast alle neueren steganographischen Algorithmen nutzen irgendeine Art von SSI um besser geeignete Stellen in den Trägermedien zu identifizieren und die Änderungen beim Einbetten einer geheimen Nachricht auf diese Bereiche zu beschränken.

Wir entwickeln einen gezielten Angriff auf *naive adaptive Steganographie*, die alle Änderungen in den bestmöglichen Stellen konzentriert. Wir zeigen anhand von vier weit verbreiteten SSI-Varianten, dass unser Angriff deren Einsatz nahezu perfekt entdeckt.

Motiviert von diesen Ergebnissen argumentieren wir, dass kein rationaler Steganograph naiv einbetten würde. Wir folgern daraus, dass der Wettbewerb zwischen Steganograph und Steganalysten am besten mit Spieltheorie beschrieben werden kann, einem fest etablierten mathematischen Konzept um die strategische Interaktion mehrerer rationaler Spieler zu modellieren. Wir entwickeln ein spieltheoretisches Rahmenmodell um ein steganographisches System mit SSI zu modellieren.

Wir instanzieren dieses Rahmenmodell mit fünf expliziten Modellen und berechnen die spieltheoretisch optimalen Strategien. Wir vergleichen diese Strategien mit den informationstheoretisch optimalen Strategien und stellen fest, dass sie sich unterscheiden. Daraus schlussfolgern wir, dass ein Steganograph der sich einem rationalen Steganalysten gegenüber sieht, den spieltheoretisch optimalen Strategien folgen sollte. Basierend auf unseren Ergebnissen entwickeln wir eine neue Strategie zur Verteilung der Einbettungsänderungen nach dem Vorbild der spieltheoretisch optimalen Strategien, die sogenannten *Ausgleichseinbettungsstrategien* (equalizer embedding strategies).

Stichworte: *Steganographie, Seiteninformation, Spieltheorie, Sicherheit*

Acknowledgements

First of all, I would like to thank my advisor Rainer Böhme for the opportunity to write my thesis at the IT Security research group at the University of Münster. He drew my interest to the exciting field of digital steganography and information hiding. Rainer was always on the spot to answer my questions, commented thoughtfully on my advances in research and motivated me to pursue the right directions. Furthermore, he commented on all parts of every paper I wrote, and this thesis is no exception. Additionally, he showed me the excitement of kiteboarding, either in combination with lively research discussions or the supervision of students during our kite seminar. Rainer inspired me in more ways than I can express here, and without his support I would not have been able to finish this thesis in the period of time it took me.

Then, I am very grateful to all my colleagues at the IT Security research group and the group for Practical Computer Science. Their personal and professional help was most welcome. Nearly all of them proofread either this thesis or one of the preceding papers. I could always count on a never-ending supply of coffee and a game of foosball, whenever I needed distraction.

Moreover, I am greatly indebted to my family who supported me in every single stage of my life. Started with encouragement for all my travels, through to the financial support of my studies, I could always rely on their support and backing.

This acknowledgments would be far from complete if I did not mention my friends, here in Münster, and all across Germany. From mental support during the exhausting time of writing up the thesis, to complete distraction whenever needed, I do not know where I would be without them. I cannot put it better than:

True friends are the ones who never leave your heart, even if they leave your life for awhile. Even after years apart, you pick up with them right where you left off, and even if they die they're never dead in your heart.

Contents

Contents	i
List of Figures	v
List of Tables	vii
List of Definitions	ix
List of Acronyms	xi
List of Symbols	xiii
1 Introduction	1
1.1 Background, Motivation, and Scope	1
1.2 Outline and Contribution	3
1.3 Notation	4
2 Preliminaries	7
2.1 Principles of Steganography	10
2.1.1 Set-Up of a Steganographic System	10
2.1.2 Embedding Operations	12
2.1.2.1 LSB Replacement	12
2.1.2.2 LSB Matching	13
2.1.3 Embedding Strategies	13
2.1.3.1 (Initial) Sequential Embedding	13
2.1.3.2 Random Uniform Embedding	14
2.1.3.3 Side-Informed Embedding	14
2.2 Security in Steganographic Systems	18
2.2.1 The Prisoners' Problem	18
2.2.2 Theoretical Security Notions	19
2.2.2.1 Information-Theoretic Approach	19
2.2.2.2 Complexity-Theoretic Approach	20
2.2.3 Empirical Security Notions	21
2.2.3.1 Receiver Operating Characteristic	22
2.2.3.2 Single Number Measures	22
2.3 Summary	23
3 Exploiting Side Information in Steganalysis	25
3.1 Side Information in Steganalysis	25
3.1.1 Steganographic Side Information and Uncertainty	25
3.1.1.1 Steganographic Side Information	26

3.1.1.2	Uncertainty	28
3.1.1.3	Common Use of Side Information in Steganography	29
3.1.2	Initial Evidence	30
3.1.2.1	Targeted Attack on PSP Steganography	30
3.1.2.2	The Detectability Profile	31
3.1.3	Formalizing Adaptive Steganography and Steganalysis	32
3.2	Powerful Steganalysis of LSB Replacement	33
3.2.1	Asymptotically Uniformly Most Powerful Test	33
3.2.2	Weighted Stego-Image Steganalysis	36
3.2.2.1	WS Steganalysis for Sequential Embedding	37
3.2.2.2	WS Steganalysis for Naïve Adaptive Embedding	37
3.3	A Targeted Attack on Naïve Adaptive Embedding	38
3.3.1	Overview of Adaptivity Criteria	38
3.3.2	Data and Set-up	39
3.3.3	Evaluation Strategy	40
3.3.4	Attacked Adaptivity Criteria	41
3.3.4.1	Local Variance	41
3.3.4.2	Edges	42
3.3.4.3	Texture	42
3.3.4.4	NUGO (Not so Undetectable steGO)	42
3.3.5	Recoverability of the Adaptivity Criteria	43
3.3.6	Empirical Results – Detecting Naïve Adaptive Embedding	44
3.4	Summary	45
4	Game Theory and Steganography	53
4.1	Motivation	53
4.2	Principles of Game Theory	54
4.2.1	Basic Definitions of Game Theory	54
4.2.2	Solution Concepts	56
4.2.2.1	Dominant Strategy Equilibrium	56
4.2.2.2	Nash Equilibrium	56
4.2.2.3	Maxmin and Minmax Strategy	57
4.2.2.4	Equalizer Strategies	57
4.3	Game-Theoretical Approaches in Steganography	58
4.3.1	Game Theory and Capacity	58
4.3.2	Game Theory and Batch Steganography	59
4.3.3	Game Theory and Detection Performance	59
4.3.4	Game Theory and Adaptive LSB Matching	60
4.4	The Game-Theoretical Framework	60
4.4.1	Basic Definitions	61
4.4.2	Set-Up and Knowledge	62
4.4.3	Strategies	63
4.5	Summary	64

5	Game-Theoretic Insights	67
5.1	Cover Models with Binary Embedding Positions	68
5.1.1	Restricted Steganalyst Model	69
5.1.1.1	Strategies	70
5.1.1.2	Payoff	72
5.1.1.3	Solving the Game	73
5.1.1.4	Numerical Illustration	79
5.1.2	Powerful Steganalyst and Fixed Net Embedding	80
5.1.2.1	Strategies	81
5.1.2.2	Embedding Impact	83
5.1.2.3	Payoff	83
5.1.2.4	Solving the Game	84
5.1.2.5	Solution and Numerical Illustration for $n = 2$ and $k = 1$	87
5.1.3	Powerful Steganalyst and Independent Embedding	91
5.1.3.1	Strategies	91
5.1.3.2	Embedding Impact	91
5.1.3.3	Payoff	91
5.1.3.4	Solving the Game	92
5.1.3.5	Solution and Numerical Illustration for $n = 2$ and $k = 1$	94
5.1.4	Summary	98
5.2	Cover Models with Two Embedding Positions	98
5.2.1	Linear Increasing PMF	99
5.2.1.1	Cover Generation	99
5.2.1.2	Embedding Impact	101
5.2.1.3	Eve's Decision: Optimal Local Detector	102
5.2.1.4	Error Rates and Payoff	103
5.2.1.5	Solving the Game	106
5.2.2	Constant Ratio PMF	109
5.2.2.1	Cover Generation and Justification	109
5.2.2.2	Embedding Impact	110
5.2.2.3	Heterogeneity	111
5.2.2.4	Eve's Decision: Optimal Local Detector	113
5.2.2.5	Error Rates and Payoff	114
5.2.2.6	Solving the Game	117
5.2.3	Imperfect Recoverability	120
5.2.3.1	Imperfect Recovery with Linear Increasing PMF	121
5.2.3.2	Imperfect Recovery with Constant Ratio PMF	123
5.2.4	Numerical Illustrations	125
5.2.4.1	Numerical Illustration for Linear Increasing PMF	125
5.2.4.2	Numerical Illustration for Constant Ratio PMF	126
5.2.4.3	Comparison	129
5.2.5	Type of Game	130
5.2.6	Discussion and Summary	132

5.3	Lessons Learned and Limitations	133
5.3.1	Lessons Learned – Secure Adaptive Steganography	133
5.3.2	Limitations	135
6	Conclusion	139
6.1	Summary of Results	139
6.1.1	Formalizing Side Information in Steganography	139
6.1.2	Game-Theoretical Modeling of Steganography	140
6.2	Outlook and Future Research	141
	Bibliography	143
A	Information-Theoretic Derivations	151
A.1	Derivation of Definition 3.2	151
A.2	Derivation of Remark 3.9	151
B	Game Theory in Related Fields	153
B.1	Multimedia Forensics	153
B.2	Digital Watermarking	154
B.3	Adversarial Classification	155
C	Omitted Proofs	157
C.1	Proof of Lemma 5.11	157
C.2	Proof of Lemma 5.12	157
C.3	Proof of Lemma 5.13	159
C.4	Proof of Lemma 5.14	159
C.5	Proof of Lemma 5.15	161
D	Curriculum Vitae	163

List of Figures

2.1	Comparison of symmetric cryptography and steganography	8
2.2	Cover and possible stego objects in a high-dimensional space	11
2.3	Visualization of naïve adaptive embedding	15
2.4	Comparison of random uniform and naïve adaptive embedding.	16
3.1	Block diagram of a steganographic system with side information	27
3.2	Examples of naïve adaptive embedding and payload $p = 0.3$	43
3.3	Mean absolute error (MAE) of the standard WS variants as function of the embedding rate p for different embedding schemes.	48
3.4	Adaptivity criterion: local variance. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p	49
3.5	Adaptivity criterion: edges. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p	50
3.6	Adaptivity criterion: texture. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p	51
3.7	Adaptivity criterion: NUGO. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p	52
5.1	Block diagram of steganographic communication system with side information and the possibility for Eve to query one position.	70
5.2	Extensive form of the steganography game for $k = 1$	74
5.3	Equilibrium strategies for $\varepsilon \gg 0$ and a linear function f	79
5.4	Equilibrium strategies for $\varepsilon \approx 0$ and a non-linear function f	80
5.5	Eve's error rates as a function of \bar{a}_0	90
5.6	The value of $e(10)$ in Eve's minimax strategy as function of f	97
5.7	Eve's error rates as a function of \bar{a}_0 with independent embedding.	97
5.8	Extensive form of the instantiated adaptive steganography game.	100
5.9	Cover generation model – linear PMF.	101
5.10	Example histograms of the cover source for $n = \ell = 2$. Compare the more suitable (brighter bars) to the less suitable (darker bars) position for a: (a) homogeneous ($t_0 = t_1 = 1.3$); (b) heterogeneous ($t_0 = 1.1$, $t_1 = 2$) cover source. The arrows indicate which values are exchanged by the LSBR embedding operation.	113
5.11	Equilibrium strategies – linear PMF	126
5.12	KLD in equilibrium – linear PMF	127
5.13	Equilibrium payoff – linear PMF	127
5.14	Entropy of cover generation – constant ratio PMF	128
5.15	Optimal adaptive embedding strategy – constant ratio PMF	129
5.16	Equilibrium strategy and payoff – constant ratio PMF	129
5.17	Comparison of change probabilities for different embedding strategies.	136

LIST OF FIGURES

5.18 Comparison of Eve's advantage for different embedding strategies. . . . 137

List of Tables

2.1	Error probabilities for different detector outputs	21
3.1	Overview of different adaptivity criteria for content-adaptive embedding and their respective embedding operations.	39
3.2	Recovery rate, calculated according to Definition 3.7.	44
3.3	Summary of detection results for 10 000 images of the BOSSBase	46
3.4	Summary of detection results for 700 images of the Dresden Image Base	47
5.1	Payoff for (Eve, Alice)	71
5.2	Game outcome for binary covers of length n	72
5.3	Game outcome for 2 positions, correct recovery and linear PMF	105
5.4	Game outcome for 2 positions, correct recovery and constant ratio PMF	117
5.5	Game outcome for 2 positions, incorrect recovery and linear PMF	121
5.6	Game outcome for 2 positions, incorrect recovery and constant ratio PMF	123
5.7	Payoff matrices of the bi-matrix games	131

List of Definitions

2.1	Cover	11
2.2	Embedding Operation	12
2.3	Embedding Strategy	12
2.4	Stego Object	12
2.5	Naïve Adaptive Embedding	15
2.6	Entropy	19
2.7	Conditional Entropy	19
2.8	Mutual Information	19
2.9	Kullback-Leibler Divergence	19
2.10	Perfect Steganographic Security	20
2.11	Steganographic Decision Problem	21
3.1	Steganographic Side Information	26
3.2	Uncertainty	28
3.3	Uncertainty with Regard to Positions	28
3.4	Perfect Recoverability	32
3.5	Order Recoverability	33
3.6	AUMP Test	34
3.7	Recovery Rate	40
4.1	Two-Player Game	54
4.2	Zero-Sum Game	55
4.3	Pure Strategy	55
4.4	Mixed Strategy	55
4.5	Support	55
4.6	Best Response Strategy	55
4.7	Dominant and Dominated Strategy	55
4.8	Incomplete Information	56
4.9	Imperfect Information	56
4.10	Dominant Strategy Equilibrium	56
4.11	Nash Equilibrium	56
4.12	Maxmin and Minmax Strategy	57
4.13	Equalizer Strategies	57
4.14	Homogeneous vs Heterogeneous Cover Source	61
4.15	Suitability	61
4.16	Adaptivity Criterion	61
4.17	Canonical Embedding Strategies	63
4.18	Canonical Detection Strategies	63
4.19	Information-Theoretic Optimal Strategies	64

LIST OF DEFINITIONS

5.1	Eve's Local Advantage	75
5.2	Eve's Total advantage	75
5.3	Recovery Rate for Two Embedding Positions	120
5.4	Optimal Embedding Strategies	133
5.5	Equalizer Embedding Strategy	134

List of Acronyms

(in alphabetical order)

AER	average error rate (under equal priors)
AUMP	asymptotically uniformly most powerful
DIB	Dresden image database
DR	decision rule
DSE	dominant strategy equilibrium
EER	equal error rate
FP_{50}	false positive rate at 50% detection rate
HUGO	highly undetectable stego
IQR	interquartile range
JPEG	Joint Photographic Experts Group
KDD	knowledge discovery and data mining
KLD	Kullback-Leibler divergence
LRT	likelihood ratio test
LSB	least significant bit
LSBM	LSB matching
LSBR	LSB replacement
MAE	mean absolute error
MAP	maximum a posteriori
ML	machine learning
NUGO	not so undetectable stego
PDF	probability density function
PET	privacy enhancing technology
PGM	portable greymap format
PMF	probability mass function
PRNG	pseudorandom number generator
PSP	preserving statistical properties
PQ	perturbed quantization
PVD	pixel value differencing
ROC	receiver operating characteristic
RS	regular/singular analysis
SP	sample pair analysis
SSI	steganographic side information
STC	syndrome trellis-codes
TN	true negative
TP	true positive
WPC	wet paper codes
WS	weighted stego-image

List of Symbols

(in the order of appearance)

\mathbf{k}	a secret key
\mathcal{K}	the key space
m	a message
\mathcal{M}	the message space
$\mathbf{x}^{(0)}$	a cover object
$x_i^{(0)}$	a position in $\mathbf{x}^{(0)}$
n	the number of positions in $\mathbf{x}^{(0)}$
\mathcal{P}_0	the cover distribution
$\mathbf{X}^{(0)}$	the cover source
$\mathbf{x}^{(1)}$	a stego object
$x_i^{(1)}$	a position in $\mathbf{x}^{(1)}$
\mathcal{P}_1	the stego distribution
$\mathbf{X}^{(1)}$	the stego objects
$\mathbf{x}^{(p)}$	a stego image with payload p
$\zeta(\cdot)$	an adaptivity criterion
p	embedding rate
$\mathbf{z}^{(0)}$	a raw grayscale image
Q	an integer scalar quantizer
k	bit-length of the (encoded) message m
H_0	Hypothesis 0
H_1	Hypothesis 1
α	false positive rate
β	false negative rate
τ	a threshold
\hat{p}	estimation of embedding rate p
Θ	a source of steganographic side information
ρ_i	a detectability measure for position i
$\mathbf{y}^{(0)}$	a ordered cover $\mathbf{x}^{(0)}$
$\hat{\mathbf{x}}^{(0)}$	an estimation of the cover $\mathbf{x}^{(0)}$
$\hat{\zeta}$	the steganalyst's estimation of the values of ζ
$\hat{\mathbf{y}}^{(1)}$	the stego object ordered by recoverable suitability
$\mathbf{x}^{(p,\lambda)}$	a weighted stego image
$\bar{\mathbf{x}}^{(p)}$	stego image with every element's LSB flipped
\mathbb{S}_i	a finite set of strategies available to player i
u_i	a payoff function for player i
s_i^*	a best response strategy
\mathbb{N}	set of all natural numbers
\mathbb{R}	set of all real numbers

\mathbb{Z}	set of all integer numbers
\mathbf{a}	Alice's choice
\mathbf{e}	Eve's choice
$\bar{\mathbf{a}}$	Alice's mixed strategy
$\bar{\mathbf{e}}$	Eve's mixed strategy
\bar{a}_i	probability that Alice embeds in position i
\bar{e}_i	probability that Eve looks at position i
$\chi(\bar{\mathbf{a}}, \bar{\mathbf{e}})$	payoff function in mixed strategies
r	recovery rate

Chapter 1

Introduction

1.1 Background, Motivation, and Scope

Steganography literally means “covered writing” and expresses any kind of concealed communication. The scientific research on steganography belongs to the field of data hiding and is closely related to, but distinct from, the areas of digital watermarking and multimedia forensics. The research field on digital steganography itself is relatively young. Its first formal security definition, a term called *undetectability*, dates back to the year 1983 and was formulated by the cryptographer Gustavus Simmons [83]. It is not surprising that it was a cryptographer who formulated this definition, as both cryptography and steganography try to achieve secure communication. The main difference between the two disciplines is that cryptography aims to protect the content of a communication, while steganography wants to hide the mere existence of a non-obvious communication.

To achieve this in practice, steganography *embeds* the secret messages in inconspicuous cover objects which have to appear plausible on the communication channel chosen. By this, we have an additional input parameter in comparison to a cryptographic system, the empirical cover object. Although information-theoretic [11] and complexity-theoretic [47] security definitions similar to such definitions in cryptography exist, they only make sense in specific set-ups.

The common practice in the research field of steganography, similar to other domains in the information hiding research, is to call a steganographic algorithm secure as long as no successful attacking algorithm against it is known.

The use of steganography is meaningful in several aspects. First of all, as even the circumstance of the communication is hidden, it can be used as a means for privacy enhancing technologies (PETs). Although in the classical model, we still have two communicating parties, we could think of a one-sided communication. For example, one of the parties uploads the cover to a website and many people can download it but only the person who knows that there is some hidden content and shares a secret key with the uploader can extract the message. Thus, no connection between sender and recipient can be attested.

Then, several countries restrict the use of cryptographic algorithms and here, using steganography helps to circumvent these restrictions.¹

But, as with every other technique available for enhancing privacy or security, steganography can also be utilized by criminals. An employee of a company could

¹See <http://www.cryptolaw.org/cls-sum.htm> for an overview on the different crypto regulations around the world.

try to smuggle out company secrets in innocent cover media or criminals could hide incriminating images within innocent looking ones, leaving the authorities helpless when trying to accuse them of illegal possession.

Furthermore, the use of steganography does not necessarily require sophisticated software. Ker [49, p. 99] proposed the probably shortest steganographic tool consisting of a single line of PERL code:

```
perl -n0777e '$_=unpack"b*",$_;split/(\s+)/,<STDIN>,5;
    @_[8]=~s/{.}{${&&v254|chop()}&v1}ge;
    print@_' <input.pgm >output.pgm secrettextfile
```

This code embeds a message backwards in the least significant bits (LSBs) of the pixels in an image in PGM format. Although it does not produce secure steganography, it exemplifies that, for example, a disgruntled employee who wants to smuggle company secrets with the help of steganography only needs basic knowledge in PERL or simply can scribble these characters on a piece of paper. If the company has no restrictions on sending innocent images from inside the company to the outside and no algorithms aiming to detect the use of steganography check the traffic, this smuggling will most likely pass unnoticed.

Very early in steganography research, the idea occurred that certain parts of a given cover object are better *suitable* for embedding than others. This led to the domain of content-adaptive embedding schemes, i.e., embedding schemes which explicitly take the content of a specific cover object into account. For example, one of the most common assumptions is that in digital images areas with a high local variance are more suitable than flat areas, as slight changes in color will most likely go unnoticed in highly textured areas. Unfortunately, the development of content-adaptive embedding schemes is often based on the respective author's intuition rather than on theoretically well-founded security principles. All content-adaptive embedding schemes have in common that the steganographer uses some kind of *side information* to identify the supposedly better suitable areas, termed *adaptivity criterion*.

The term *side information* in steganography is used inconsistently and only the steganographer is assumed to have access to the side information. In modern steganography it is agreed to follow Kerckhoffs' principle [59] and thus the security of a steganographic system should only rely on the secrecy of the key. As the usage of side information is defined in the embedding scheme, we have to assume a steganalyst to know if side information is used for embedding. It seems only rational for her to make use of this knowledge. Following from this, one of the research questions we tackle is which strategy is the best when facing such an informed steganalyst. Can we do better than using only the best suitable positions for embedding or trying to minimize information-theoretic measures of undetectability?

Motivated by the fact that a steganographer would change her strategy if she knew a steganalyst assumes her to use only the best suitable positions for embedding, we justify a presentation of the steganographic problem by means of *game theory*. We develop a game-theoretic framework, which explicitly uses side information as one of the input parameters.

1.2 Outline and Contribution

This thesis consists of six chapters. After presenting the motivation, basic structure, and notation of the thesis in this chapter, Chapter 2 introduces the preliminaries of the research in the field of digital steganography. It aims at introducing the basic set-up of a steganographic system and the security notions specific to such a system. Furthermore, we compare a steganographic system to its counterpart from cryptography and pay special attention to the different protection goals. Chapter 2 is constructed in a way to familiarize the reader without prior knowledge about the field of steganography. We introduce the common terms and practices which are relevant for the remainder of the thesis. Thus, readers already familiar with research in steganography may skip this part with a clear conscience.

Chapter 3 reveals how side information in a steganographic system can be exploited. To the best of our knowledge we are the first to give a theoretically well-founded definition of side information in steganography, termed *steganographic side information* (SSI) and the term of *uncertainty* on the side of the steganalyst. We explicitly tie the definition of uncertainty to the steganographic protection goal of undetectability. After giving initial evidence that side information is already commonly used in steganography, we formalize adaptive steganography and steganalysis. Then, we introduce a statistically most powerful steganalysis method for the detection of LSB replacement and present a popular approximation of this method, the so-called Weighed-Stego Image (WS) steganalysis. We pay special attention to an extension of WS steganalysis to detect initial sequential embedding and show how this extension can be reformulated to a targeted detector aiming to detect the use of naïve adaptive embedding. We say a steganographer performs naïve adaptive embedding when she deterministically selects those positions from a cover object to embed that seem most suitable, according to some kind of SSI. We scrutinize the assumption that this embedding strategy always improves steganographic security.

Finally, we show the performance of a targeted attack against several widely-used adaptivity criteria, together with a formal analysis of affected adaptivity criteria. All experiments are performed on a large image database and the results are compared to the untargeted versions of WS steganalysis.

Motivated by the results in Chapter 3 we conclude that a rational steganalyst will use the side information available to her. As a rational steganographer will use her knowledge about the possibility that a steganalyst might exploit the side information and thus will adapt her embedding strategy accordingly, we argue why side-informed steganography is best studied using game theory in Chapter 4. Game theory is the mathematical study of two or more rational opponents with different goals. We briefly introduce the concepts of game theory needed in this thesis and give an overview of game-theoretical approaches in the field of steganography. We continue to present our game-theoretical framework which models a steganographic system with side information. As we assume that both steganographer and steganalyst try to maximize their payoff, measured in high and low detection rates, respectively, the classical information-theoretical analysis

cannot cover this contest.

Chapter 5 instantiates the proposed framework for several possible modeling choices for adaptive embedding and the detection thereof. We justify each instantiation and show that game-theoretically optimal embedding and detection strategies coincide with the commonly used strategies only in degenerate cases of unrealistic cover source models. Furthermore, we show that the information-theoretically optimal strategies differ from the game-theoretically optimal ones. In all instantiations we implicitly assume two kinds of SSI present in the steganographic system, one determining the order of the positions and one determining the suitability of single positions. We differentiate assumptions about the power of the steganalyst. Initially, we grant the steganalyst exact knowledge of the order of suitability, but relax this assumption later.

We argue that game-theoretical solutions should be considered for practical embedding strategies. All findings are illustrated numerically. Furthermore, we identify a very promising new paradigm for finding game-theoretical optimal embedding strategies. We show that the concept of the so-called *equalizer strategies* results in more secure embedding functions when the steganalyst anticipates adaptive embedding. We give an outlook on how to utilize this paradigm for empirical cover sources, even when the cover distribution is unknown, and point out the limitations of our approaches.

The final Chapter 6 summarizes the findings of this thesis, opens a discussion of the results and identifies areas for further research.

Note that large parts of this thesis build upon our conference publications and journal submissions. Sections 3.1.3, 3.2.2.2 and 3.3 reuse parts of [79], Sections 4.4, 5.2.2 and 5.2.3.2 are adapted from [78], Section 5.1.1 is a revised versions of [45], Section 5.1.2 of [46], Section 5.1.3 of [80], and Section 5.2.1 extends [77]. Then, Sections 5.2.1.5, 5.2.2, and 5.2.3 were rewritten after the initial submission of this thesis, according to the helpful comments of the anonymous reviewers of our journal submission [78]. The discussion with these reviewers also pointed out the need and gave the basic idea for Section 5.2.5, which was also added after the initial submission of this thesis.

With regard to the conference publications, we want to remark that several of them originated from joint work with other researchers. So, [45] was written in cooperation with Benjamin Johnson from University of California, Berkeley, USA and [46, 80] are joint work with Aron Laszka from Budapest University of Technology and Economics, Hungary, Benjamin Johnson from University of California, Berkeley, USA, and Jens Grossklags from Pennsylvania State University, USA.

Finally, note that Section 5.3.2 was added after the evaluation of the initial submission of this thesis according to the comments of the reviewers.

1.3 Notation

Random variables are denoted as upper-case letters, their realizations (and constants) in lower case. Vectors and matrices, shorthand for one- and two-dimensional arrays, respectively, are typeset boldface $\mathbf{x} = (x_0, \dots, x_{n-1})$ or $\mathbf{a} = (x_0, \dots, x_{nm})$ with n and $n \times m$ implicit. Following the convention for real numbers \mathbb{R} and natural numbers \mathbb{N} ,

sets are written in double-line notation.

Following the notation in [8], superscript (0) in $x_i^{(0)}$ denotes a symbol before embedding and superscript (1) in $x_i^{(1)}$ denotes a symbol after embedding, $\mathbf{x}_{(i)}^{(1)}$ the stego object with position i changed, and $\mathbf{x}^{(0)}$ denotes a cover object. By extension, superscript (\bar{a}) in $x_i^{(\bar{a})}$ means that the symbol has been changed by embedding with probability \bar{a} and $\mathbf{x}^{(p)}$ denotes a stego object with payload p . \mathcal{P}_0 is the probability distribution of the cover source $\mathbf{X}^{(0)}$. \mathcal{P}_1 is the probability distribution of stego objects $\mathbf{X}^{(1)}$. $\mathcal{P}_{(x_i)}$ is the probability distribution after embedding only in the i -th element and $\mathcal{P}_{(\bar{a})}$ the stego distribution when following the embedding strategy \bar{a} .

A function $\zeta : \mathbb{Z}^n \times \{0, \dots, n-1\} \rightarrow \mathbb{R}$ calculates a local criterion for the i -th element (position) of the n -dimensional input vector (cover). To simplify the notation, we may skip the explicit argument of the entire vector and write $\zeta(x_i^{(0)}) = \zeta(\mathbf{x}^{(0)}, i)$ and $\zeta(x_i^{(p)}) = \zeta(\mathbf{x}^{(p)}, i)$ to denote the criterion calculated for the i -th pixel of the cover and stego object, respectively. We write $\zeta(\mathbf{x}^{(0)}, \boldsymbol{\theta})$ if $\zeta(\cdot)$ is based on some kind of side information $\boldsymbol{\theta}$. Usually, $\zeta(\cdot)$ measures the suitability of locations for embedding and establishes an order within a cover $\mathbf{x}^{(0)}$. We write $\mathbf{y}^{(0)}$ for a cover $\mathbf{x}^{(0)}$ with elements ordered by *decreasing* suitability for embedding, i. e., $\zeta_{i-1}(\mathbf{y}^{(0)}, \boldsymbol{\theta}) \geq \zeta_i(\mathbf{y}^{(0)}, \boldsymbol{\theta})$ for $1 \leq i < n-1$. We use the hat notation to express the estimation of values or vectors. So, $\hat{\mathbf{x}}^{(0)}$ is an estimation of the cover $\mathbf{x}^{(0)}$, $\hat{\zeta}$ is the steganalyst's estimation of the values of ζ and $\hat{\mathbf{y}}^{(1)}$ is the stego object ordered by recoverable suitability.

We use the standard notation for Binomial coefficients, i. e., $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Chapter 2

Preliminaries

The purpose of a steganographic communication system is to hide the mere existence of a secret communication. An attacker, termed warden or steganalyst in the steganographic jargon, should not be able to detect the presence of a secret message, neither by visually examining possible stego objects nor by applying statistical or machine learning-based methods.

A minimal steganographic embedding function takes a key and a message as input. But to communicate the message to the recipient, we would need a channel where such messages seem plausible, i.e., where they do not cause suspicion. In [8, p. 104] it is argued that if we had a channel where completely random looking cover objects were plausible, the problem of secure steganography would be reduced to cryptography with the protection goal of indistinguishability of ciphertexts from random sequences.

As these channels are uncommon in practice and their existence might be suspicious by itself, a handy convention in steganography is to select a plausible communication channel and tweak some of the objects sent through it so that they contain the secret message. By this, we get an additional input parameter into a steganographic embedding function, the *cover object*. This cover object, together with the message and the secret key, is processed into a *stego object*, which is sent over the communication channel to the intended recipient. To achieve undetectability of the communication, the stego object has to look like a plausible cover object when inspected by a warden monitoring the channel. Thus, from now on, we assume the cover object, or the cover source, i.e., the source producing regular covers, to be an essential part of a steganographic communication system.

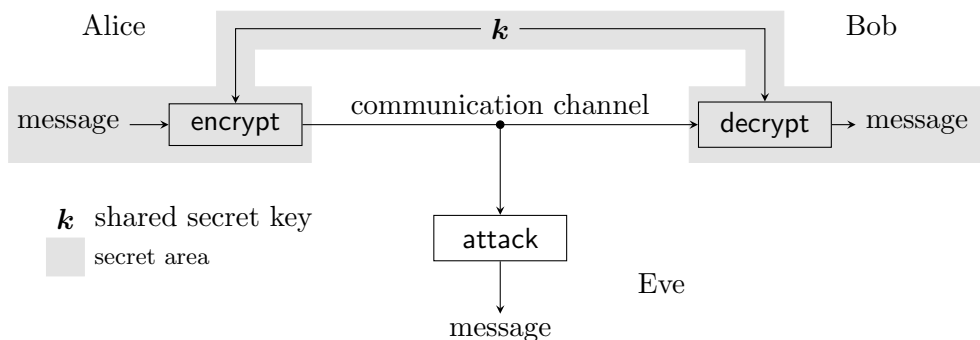
In contrast to cryptography, where the protection goal is confidentiality, steganography goes one step further. In cryptography everyone who is interested knows that there is a communication between sender and recipient. The protection goal of *undetectability* is to disguise the fact that a secret communication takes place.

For a better comparison of symmetric cryptography and steganography Figure 2.1 shows the basic set-ups as block diagrams. We use the common names Alice for the sender, Bob for the recipient, and Eve for the passive attacker.

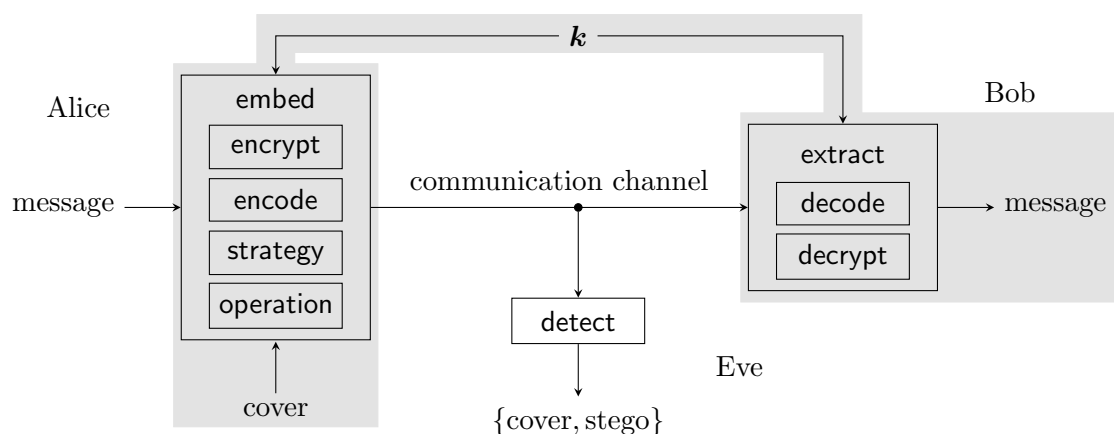
Both scenarios last on the assumption that a secret key \mathbf{k} , from the key space \mathcal{K} , is shared between Alice and Bob beforehand and the purpose of both systems is to communicate a message from the message space \mathcal{M} , usually $\{0, 1\}^*$, from Alice to Bob.

Figure 2.1(a) shows the schematic building blocks of a symmetric crypto system, namely the *encryption function* `encrypt` on the side of Alice, the *decryption function* `decrypt` on the side of Bob and the *attacking function* `attack` utilized by Eve.

Figure 2.1(b) illustrates a steganographic communication system.



(a) Block diagram of a symmetric cryptographic communication system



(b) Block diagram of a steganographic communication system

Figure 2.1: Comparison of symmetric cryptography and steganography

Here, the fundamental building blocks are Alice's *embedding function* `embed`, the *extraction function* `extract` used by Bob, and Eve's *detector* `detect` that has to decide if the object under observation is a cover or a stego object.

In a steganographic communication system, `embed` combines four different functions, namely `encrypt`, the function that encrypts a message, `encode`, the function that encodes an encrypted message, `strategy`, the function that selects the embedding positions, and `operation`, the function that changes the selected embedding positions so that they contain the encoded message. The function `extract` combines two functions that ensure the successful receipt of the original message, namely `decode`, which decodes the encrypted message from the stego object, and `decrypt`, which decrypts the original message.

So, the cryptographic functions `encrypt` and `decrypt` belong to a steganographic system and we assume them to be secure. The functions `encode` and `decode` ensure that Bob can retrieve the original message from the stego object. Here, we assume perfect encoding, as this problem is mainly solved with the with the recent introduction of *Syndrome-Trellis-Codes* (STC) [22]. This encoding scheme allows message extraction without the need to share the embedding positions with Bob and is asymptotically perfect.² Clearly, the encryption and encoding function depend on each other and so do the embedding strategy and operation.

Assuming the availability of secure cryptography and perfect encoding schemes, we restrict ourselves to the embedding strategy in combination with the embedding operation in the remainder of this thesis.

In both block diagrams, the gray shaded area symbolizes the secret area of the communication systems, i.e., the part that the attacker has no access to. Although Eve may know the encryption or embedding function, she does not know the instantiation of the function with the specific key.

A cryptographic scheme is considered broken if an attacker is able to read the encrypted messages, a steganographic system is already considered broken if an attacker can detect the mere circumstance of hidden communication, i.e., if the attacker can distinguish cover objects from stego objects with more than 50% accuracy.

As indicated, the message is not part of the secret area in a steganographic system, which seems counterintuitive at a first glance. But highlighting the protection goal of undetectability, in secure steganography we may even give Eve knowledge about the message, she still should not be able to distinguish if this specific message is hidden in a potential stego object.

Similar to the evolution of cryptanalysis as the counterpart of cryptography, in steganography, the field of steganalysis emerged. In accordance with the different protection goals, steganalysis does not aim at reading the hidden messages, but simply at detecting the use of steganography. If the steganalyst detects the use of steganography, she may become an active attacker and simply block the communication channel to prevent further communication.

Steganography borrows Kerckhoffs' principle [59] from cryptography, which states that the embedding function is known to the attacker and the security of a scheme should be guaranteed solely relying on the secrecy of the shared key. The exact interpretation of the principle for steganography is heavily discussed in the research community, as it is not entirely clear which of the additional parameters in a steganographic system should count as common knowledge. Can the steganalyst know the bit-length of the hidden message? Is she allowed to know the cover source? The only thing that, by now, is agreed upon is that she should not be able to get hold of the cover object itself, as she could simply compare the cover object with a potential stego object and, although she might not be able to read the hidden message, she still can see that the cover object was altered and thus conclude that it must contain a message.

²STC are basically an extension of Wet Paper Codes, which will be introduced in Section 2.1.3.3.4.

2.1 Principles of Steganography

We limit our explanations and examples to the area of digital images but still refer to positions instead of pixels. Most of the theoretical findings can easily be transformed to, for example, audio files, where positions would refer to samples or digital videos, where positions could be frames or pixels within frames.

2.1.1 Set-Up of a Steganographic System

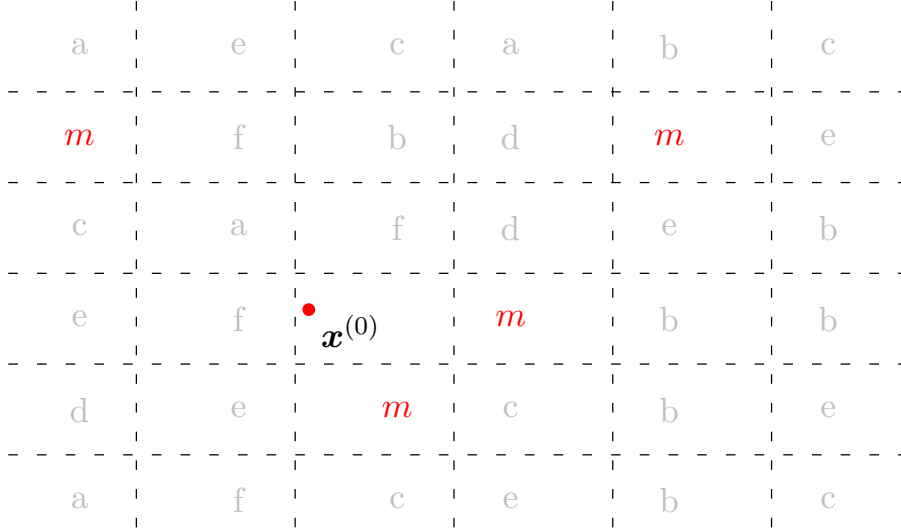
Steganography is often called *the art and science of hidden communication*. While the first use of the term steganography (*Steganographia*, then) indeed dates back to the year 1499 and other methods to enable hidden communication are known from ancient Greece [27, p. 3], the art developed into the scientific discipline of *digital steganography* with the upcoming and common usage of digital media. As, e.g., digital images are sent everyday over the Internet they are by definition plausible. Furthermore, empirical images are hard to model, thus slight changes might be assumed to go unnoticed.

Another perspective of the impact of the steganographic embedding function on a cover object is to think of cover objects as points in a high-dimensional space. Then, we can assume this space to be partitioned, often key-dependent, into disjoint regions corresponding to the elements of the set of all hidden messages. A steganographic embedding function outputs a point within the region associated with the given cover object and message. The available coding schemes allow the steganographer to partition the high-dimensional space over the message space such that embedding a given message has many possible solutions [3, 21, 33].

This is exemplarily depicted in Figure 2.2, where we reduce the high-dimensional space for the sake of clarity to two dimensions. In Figure 2.2(a) we see the cover space partitioned into regions corresponding to different messages. The regions with a highlighted m portray the regions that would correspond to the desired message m , depending on the coding scheme.

In practice, the high-dimensional space is sparsely populated with empirical covers and the message space is large. Thus, the idea to draw covers until one is found that falls in the desired region, i.e., already contains the message to hide, a method called “rejection sampling” [42], is unfeasible. Therefore, the standard approach in practical steganography is to take a given cover and move it into the region corresponding to the desired hidden message by slightly modifying its positions.

Figure 2.2(b) shows the location of all the possible stego objects that do contain m . Now, Alice chooses, which of the possible stego objects she uses. This decision is often influenced by her belief about which region is hardest to model by Eve.



(a) Cover object and coding-dependent partitioning of the high-dimensional space

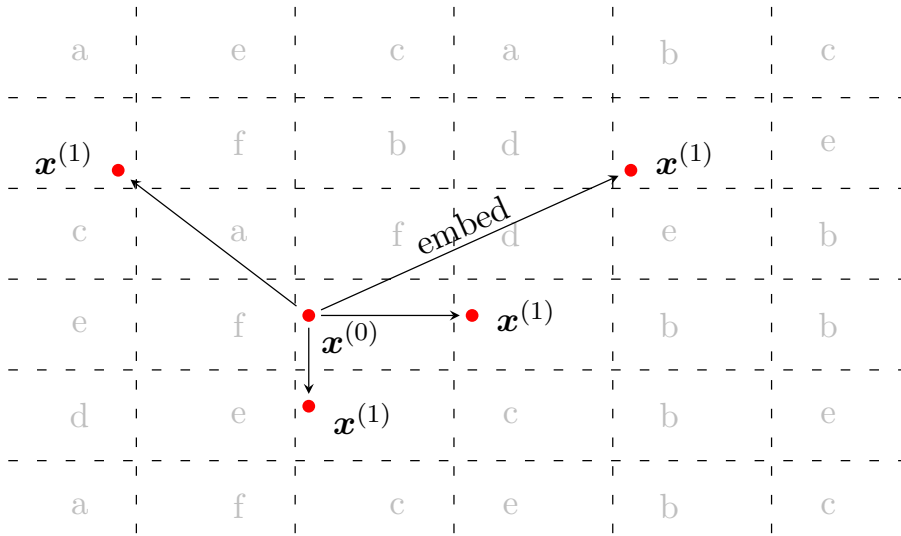

 (b) Different possibilities to embed m in $\mathbf{x}^{(0)}$

Figure 2.2: Cover and possible stego objects in a high-dimensional space

It is useful to introduce some formalism about the basic components of a steganographic communication system.

Definition 2.1 (Cover). A vector $\mathbf{x}^{(0)} = (x_0^{(0)}, \dots, x_{n-1}^{(0)})$ of n discrete symbols is called cover, if it is a realization of the cover source $\mathbf{X}^{(0)}$ drawn from \mathcal{P}_0 . Specifically, every symbol $x_i^{(0)}$ of the cover can take values in $\mathbb{X} := \{0, \dots, 2^\ell - 1\}$.

The embedding function is a key-dependent mapping of cover $\mathbf{x}^{(0)}$ and message to

a stego object $\mathbf{x}^{(1)}$. To study steganography, we decompose the embedding function into atomic operations that modify or select individual cover symbols.

Without loss of generality, we assume for simplicity a one-to-one mapping between cover symbols and bits carrying steganographic semantic, typically the encrypted and encoded representation of the message m .

In general, the steganographer has to decide *how* to change single cover positions and *which* of the positions she changes. This leads to the partitioning of the *embedding function* `embed` into the *embedding strategy*, i.e., the method on how the embedding positions within the cover are chosen and the *embedding operation*, i.e., the method how single positions within a cover are altered to embed the steganographic payload.

Definition 2.2 (Embedding Operation). *A function `emb(·)` that takes a cover symbol $x_i^{(0)}$ as input and outputs the corresponding symbol $x_i^{(1)}$ with the opposite steganographic semantic is called embedding operation.*

Definition 2.3 (Embedding Strategy). *A function is called embedding strategy if it takes as input an encrypted and encoded message m and a cover object $\mathbf{x}^{(0)}$ and outputs positions $\{i\}$ of $\mathbf{x}^{(0)}$ such that m can be hidden in these positions using the embedding operation `emb(·)`.*

The output of the embedding function `embed` is called stego object.

Definition 2.4 (Stego Object). *A vector $\mathbf{x}^{(1)} = (x_0^{(1)}, \dots, x_{n-1}^{(1)})$ of n discrete symbols is called stego object, if it stems from a cover object $\mathbf{x}^{(0)}$ that was processed by the embedding function `embed` and contains a hidden message m . In particular, every symbol $x_i^{(1)}$ of the stego object takes values in the same domain as the values of the cover object, i.e., $x_i^{(1)} \in \mathbb{X}$.*

We denote the probability distribution of stego objects by \mathcal{P}_1 .

2.1.2 Embedding Operations

The first parameter of choice in a steganographic communication system is the way the chosen embedding positions are altered such that the secret message can be communicated. As we assume each possible embedding position to be represented with a fixed number ℓ of bits, e.g., $\ell = 8$ for grayscale images, changes in the *least significant bit* (LSB) are supposed to introduce the least detectable artifacts. Also, the resemblance of the LSB plane of empirical images to white noise is assumed to make changes within this layer harder to recognize. Consequently, almost all of the early embedding functions assumed the secret message to be a binary sequence and the embedding operation ensured that the LSBs along a chosen path coincided with the message bits.

2.1.2.1 LSB Replacement

The simplest form of an embedding operation is to replace the LSBs with the respective bit of the message. This embedding operation is called *LSB replacement* (LSBR) (or

sometimes simply *LSB embedding* [27]). The simplicity of this embedding operation turns out to be also its most severe weakness. LSBR simply swaps the values $2j$ by $2j + 1$, and vice versa, for $j \in \{0, \dots, 2^{\ell-1}\}$. This can be expressed by

$$\text{LSBR}(x) := \begin{cases} x + 1 & : \text{if } x \text{ is even, and } x < 2^{\ell-1} \\ x - 1 & : \text{if } x \text{ is odd, and } x > 0. \end{cases} \quad (2.1)$$

It can be seen from Equation (2.1) that even values will never be decreased and odd values will never be increased. Thus, if x is within the LSB pair $\{2j, 2j + 1\}$ it will stay in this pair after embedding. This leads to powerful, so-called structural detection methods for LSBR, e.g., [16, 32].

2.1.2.2 LSB Matching

To overcome the creation of LSB pairs, in [82] the LSBs of the embedding positions were randomly increased or decreased. This little modification of the embedding operation, now known as *LSB Matching* (LSBM), was shown to be much harder to detect than the original LSBR [48].

$$\text{LSBM}(x) := \begin{cases} x + 1 & : \text{with probability } \frac{1}{2}, \text{ if } x < 2^{\ell-1} \\ x - 1 & : \text{with probability } \frac{1}{2}, \text{ if } x > 0. \end{cases} \quad (2.2)$$

Several other embedding operations use more than only the LSB, namely *Mod-k Replacement* and *Mod-k Matching*, but as they are not covered within this thesis, we omit the description here and refer the interested reader for example to [8, pp. 39].

2.1.3 Embedding Strategies

The second parameter of choice of a steganographer is to decide in which positions of the cover she will hide her message, the *embedding strategy*.

2.1.3.1 (Initial) Sequential Embedding

The simplest form of an embedding strategy is to place the secret message at the beginning (usually starting with the top left corner) of a cover object, a method known as *initial sequential embedding*. The recipient simply reads the same positions out and can retrieve the secret message. But, a potential attacker also knows in which positions to look for traces of an embedded message. Together with the embedding operation of LSBR, one of the leading researchers in the field of steganography stated:

“Sequential replacement of LSBs was one of the first steganographic embedding methods described, and is perhaps the worst.” [51, p. 456]

Sequential embedding which does not start in the top left corner does not improve the security.

2.1.3.2 Random Uniform Embedding

Inspired by the observation that the LSB plane of natural images roughly resembles white noise, the idea emerged that randomly flipping a subset of bits from the LSB plane will go undetected [27, p. 61]. Then, the idea of choosing a (pseudo-)random path through the image and embed the secret message along this path emerged. This embedding strategy is one of the most prevalent in steganography and is known as *random uniform embedding*. To circumvent a shared secret key of the same length as the message, the path could, for example, be created using a pseudorandom number generator (PRNG) and the secret key would be the seed value for the PRNG.

2.1.3.3 Side-Informed Embedding

In steganography, we have one input parameter more than in cryptography, namely the cover object. As it can be seen in Figure 2.1(b) this is contained in the secret area, and thus, the attacker has no access to it. Side-informed embedding uses (side) information from the cover object to gain an advantage over the attacker. Here, we present three of the most commonly used embedding paradigms.

2.1.3.3.1 Content-Adaptive Embedding

Already in the earliest days of steganographic research the idea occurred that some parts of a cover object might be more *suitable* for embedding than other parts [2]. More suitable in this case means that, specific to the embedding operation, changes within certain areas are harder to detect, i.e., the modified values are more plausible to stem from a cover object. This led to *content-adaptive* embedding strategies. As the name suggests, a strategy is called content-adaptive if it takes the content of the cover realization into account explicitly. Almost all recently developed embedding schemes are content-adaptive in some way.

All content-adaptive embedding schemes have in common that they define a so-called *adaptivity criterion* $\zeta(\cdot)$, which identifies more suitable embedding positions. The adaptivity criteria can be roughly divided into locally calculated criteria and distortion minimizing criteria. An example for the first category is the assumption that areas with a high local variance are more suitable. The second category assumes that embedding positions introducing less distortion are preferable. The claimed purpose of all adaptivity criteria is to identify a (partial) ordering of all available embedding positions according to their suitability for embedding.

2.1.3.3.2 Naïve Adaptive Embedding

The first idea that comes to mind after defining an adaptivity criterion and measuring the suitability of all positions in a given cover is to use only the best suitable positions for embedding. We call this embedding strategy *naïve adaptive embedding*. The formal definition is:

Definition 2.5 (Naïve Adaptive Embedding). *A steganographer uses naïve adaptive embedding when she defines an adaptivity criterion ζ that measures the suitability of all embedding positions and solely embeds in the $p \cdot n \leq n$ most suitable symbols, where $p \leq 1$ is the embedding rate.*

Figure 2.3 visualizes the course of action for naïve adaptive embedding. The suitability of the positions of the cover in Figure 2.3(a) are measured with the adaptivity criterion ζ in Figure 2.3(b). Under the assumption that a higher value for $\zeta(x_i)$ expresses better suitability for embedding, the best suitable positions are highlighted in Figure 2.3(c). Finally, Figure 2.3(d) shows the cover with highlighted embedding positions, ordered by decreasing suitability. Of course, the stego object is transmitted with its positions in the original order.

x_0	x_1	x_2	x_3	x_4
x_5	x_6	x_7	x_8	x_9
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}

(a) Original cover

0.23	2.53	4.39	0.96	2.76
6.12	5.40	0.58	3.90	4.72
1.43	3.66	6.91	1.22	3.48
2.29	6.74	0.57	4.92	1.35
6.11	6.26	1.54	2.08	0.78

(b) Values of the adaptivity criterion ζ

x_0	x_1	x_2	x_3	x_4
x_5	x_6	x_7	x_8	x_9
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}

(c) Best embedding positions highlighted

x_{12}	x_{16}	x_{21}	x_5	x_{20}
x_6	x_{18}	x_9	x_2	x_8
x_{11}	x_{14}	x_4	x_1	x_{15}
x_{23}	x_{22}	x_{10}	x_{19}	x_{13}
x_3	x_{24}	x_7	x_{17}	x_0

(d) Cover object ordered by suitability

Figure 2.3: Visualization of naïve adaptive embedding

Figure 2.4 compares random uniform embedding (dashed blue line) with naïve adaptive embedding (dotted blue line) in a cover object with positions sorted according to their suitability (red line), showing that random uniform embedding potentially uses

all available embedding positions, whereas naïve adaptive embedding solely concentrates on the most suitable embedding positions.

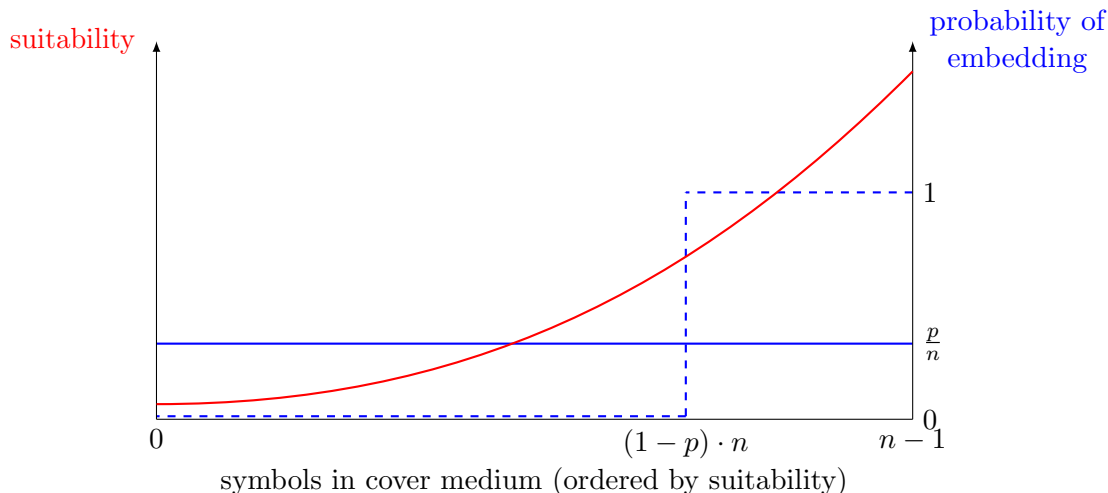


Figure 2.4: Comparison of random uniform (solid blue line) and naïve adaptive embedding (dashed blue line).

2.1.3.3.3 Perturbed Quantization Embedding

Perturbed Quantization (PQ) steganography [33], assumes that Alice obtains the cover image through some preprocessing that ends with quantization (or any other information-reducing process), for example lossy compression. The sender uses the so-called *pre-cover* before processing to identify the elements in the cover that yield the highest uncertainty to an attacker about their values prior to processing. The authors state that the side information which is used by the steganographer is *in principle unavailable* to the recipient and thus also to an attacker.

The basic concept is best shown in a case where Alice has a raw grayscale image $\mathbf{z}^{(0)}$ that has never been processed before of which she wants to send a downgraded version over the communication channel. This setting is realistic, as it can be assumed that Alice has the device to capture the cover image and thus has access to the raw image. But sending raw images over a communication channel might be suspicious by itself, so she downgrades it, for example by lossy compression (e.g., JPEG compression).

Alice then performs a transformation F of the following form:

$$F = Q \circ T : \mathcal{Z}^N \rightarrow \mathcal{X}^n, \quad (2.3)$$

where \mathcal{X}^n is the range of the cover (and stego) object, \mathcal{Z}^N is the range of the elements of the pre-cover and $T : \mathcal{Z}^N \rightarrow \mathbb{R}^n$ is some form of processing. $Q \circ T(\mathbf{z}^{(0)})$ stands for $Q(T(\mathbf{z}^{(0)}))$ and $T(\mathbf{z}^{(0)})$ is the intermediate image or the cover object in the basic sense. The mapping Q is an integer scalar quantizer that quantizes the values of $T(\mathbf{z}^{(0)})$ to its

nearest integer value $x_i^{(0)}$. So, it holds that $x_i^{(0)} \leq z_i^{(0)} < x_i^{(0)} + 1$. Now, Alice can identify the positions i in $T(\mathbf{z}^{(0)})$ that have the largest quantization error $\varepsilon_i = |z_i^{(0)} - x_i^{(0)}|$ and choose these positions for embedding. Consequently, she minimizes the errors introduced while embedding the message. To give a small example: If $z_i^{(0)} = 42.47$, the quantized value would be $x_i^{(0)} = 42$ and thus $\varepsilon_i = 0.47$. If the message bit m_i at position i is 1, Alice would flip the bit, thus getting $x_i^{(1)} = 43$. By this, she introduces an error of 0.53 that is only 0.06 larger than ε_i . If she chooses a position with an original value of, e.g., $z_i^{(0)} = 42.04$ before quantization and had to change it to $x_i^{(1)} = 43$ for embedding, she would introduce an error of 0.96 which is 0.92 larger than the ε_i of 0.04 in that case.

The authors of [33] propose that Alice can choose a small ε (e.g. $\varepsilon = 0.1$) and use only the positions i for which it holds that $\varepsilon_i \in [0.5 - \varepsilon, 0.5 + \varepsilon]$. By this, the difference between the average rounding distortion of the regular quantizer and the perturbed version is ε^2 instead of 0.25 which would be the average rounding distortion.

Furthermore, Alice could make additional use of her knowledge about the image content and confine her changes to regions in the image she thinks are best suitable for embedding, as described in the previous section.

2.1.3.3.4 Non-Shared Selection Channel

Both embedding strategies presented in the previous sections share a common problem. The steganographer uses information in her embedding strategy that might not be available to the recipient. With content-adaptive embedding, the changes in the positions or those around it might alter the value $\zeta(x_i^{(1)})$ for some positions i , so that the recipient is not able to establish the same order as the steganographer and thus will not receive the original message. With PQ embedding, the information used by the steganographer is even assumed to be completely removed from the stego object.

To ensure a successful extraction of the message on the side of the recipient, there has to be a way to communicate the so-called *selection channel* [3] without revealing it to the steganalyst.

A solution to this is presented (together with the PQ embedding approach) in [33] with the so-called *Wet Paper Codes* (WPC). The name of WPC is a metaphor for a cover image that was exposed to rain before embedding. The steganographer is only able to use the (random) positions that were not hit by the rain, referred to as *dry positions*, to embed her secret message. During transmission of the stego object the *wet positions* dry out and thus the recipient is not able to identify which positions were used for embedding. If the sender uses WPC to encode her message, the recipient is nonetheless able to extract the message.

The idea for WPC originates from n -bit memory channels with up to $n - k$ defective cells. The capacity of such a channel is k bit, the bit-length of our message m .

First, it is assumed that the recipient knows the length of the message.³ The sender wants to communicate the message $\mathbf{m} = \{m_0, \dots, m_{k-1}\}^T$. The sender first uses the

³This assumption is removed later.

shared secret key to generate a pseudorandom binary matrix \mathbf{D} of dimension $k \times n$. To use the example of the PQ method from above, the sender will round $z_j, j \in C$ to the column vector \mathbf{z} , so that the modified binary column vector \mathbf{b} with $b_i = y_i \bmod 2$ satisfies

$$\mathbf{D}\mathbf{b} = \mathbf{m}. \quad (2.4)$$

To achieve this, the sender has to solve a system of linear equations in \mathbb{Z}_2 .

Then, the sender sends the stego object to the recipient, who forms the vector \mathbf{b} with $b_i = y_i \bmod 2$ and then multiplies it with the shared secret matrix \mathbf{D} :

$$\mathbf{m} = \mathbf{D}\mathbf{b}. \quad (2.5)$$

It is shown in [33] that the expected number of bits that can be communicated is likely close to k and that the message length does not have to be communicated beforehand, as the matrix \mathbf{D} can be generated row by row and the sender can reserve the first $\lceil \log_2(n) \rceil$ bits of the message for a header containing the overall message length and thus the number of rows in \mathbf{D} . So, the recipient first has to generate $\lceil \log_2(n) \rceil$ rows⁴ to read the message length and then create the remaining rows of \mathbf{D} needed for the extraction of the message. Using this method, the expected number of bits that can be communicated is reduced to $k - \log_2(n)$ [33].

2.2 Security in Steganographic Systems

In Section 2.1 we used the protection goal of undetectability informally. In this section we formalize security in a steganographic communication system. We start with the original proposal of the protection goal of undetectability and then move on to a theoretical and an empirical point of view.

2.2.1 The Prisoners' Problem

In 1983 the cryptographer Gustavus Simmons [83] formulated the *prisoners' problem* as a scenario for secure steganographic communication. In this scenario, two prisoners, Alice and Bob, who are held captive in different cells, want to scheme an escape plan but are not allowed to communicate directly. The warden Eve allows them to communicate but inspects every message they send to each other before delivering it. If Eve gets the slightest hint that the two plan to escape, she will throw both of them into solitary confinement and thereby cutting every possibility of further communication. The solution for Alice and Bob is to use steganography.

It has to be noted that Eve is not allowed to falsely accuse the two of communicating covertly. The consequences of this cannot really be framed in the set-up of the prisoners' problem but one can think about it in such a way that a false accusation would create costs on the side of Eve and thus should be avoided.

⁴Note that n is common knowledge, as it is the length of the stego object.

2.2.2 Theoretical Security Notions

Similar to cryptography, several theoretical security notions exist in steganography. This section is not meant to give an exhaustive overview, but to highlight the security notions from different domains. We provide two theoretical security notions from information theory and from complexity theory.

2.2.2.1 Information-Theoretic Approach

Cachin [11] studies steganography from an information-theoretic perspective and defines the security of a steganographic communication system using the notions of entropy, mutual information and Kullback-Leibler divergence.

Definition 2.6 (Entropy). *The entropy $H(X)$ of a probability distribution P_X over an alphabet \mathcal{X} is defined as:*

$$H(X) = - \sum_{x \in \mathcal{X}} P_X(x) \log P_X(x). \quad (2.6)$$

Definition 2.7 (Conditional Entropy). *The conditional entropy of X given Y is given by:*

$$H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y). \quad (2.7)$$

Definition 2.8 (Mutual Information). *The mutual information between X and Y is defined as:*

$$I(X; Y) = H(X) - H(X|Y). \quad (2.8)$$

Definition 2.9 (Kullback-Leibler Divergence). *The Kullback-Leibler divergence KLD (also known as the relative entropy) between two probability distributions P_X and P_Y is defined as:*

$$\text{KLD}(P_X || P_Y) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_Y(x)}. \quad (2.9)$$

Cachin considers the steganalyst's capability of detecting an embedded message using statistical hypothesis testing.

Let $\mathcal{P}_0, \mathcal{P}_1$ be two probability distributions. H_0 and H_1 are two hypotheses for an observed measurement Q . If Q was generated according to \mathcal{P}_0 , H_0 is true. If Q was generated according to \mathcal{P}_1 , H_1 is true. Formally:

$$\begin{aligned} H_0 &: Q \text{ was generated by } \mathcal{P}_0 \\ H_1 &: Q \text{ was generated by } \mathcal{P}_1 \end{aligned} \quad (2.10)$$

In steganography, \mathcal{P}_0 is the cover distribution and \mathcal{P}_1 the stego distribution. Cachin assumes a computationally unbounded steganalyst who fully knows \mathcal{P}_0 and \mathcal{P}_1 . He

defines a steganographic system as consisting of the cover distribution \mathcal{P}_0 , the message space \mathcal{M} and the algorithms $\langle \text{embed}, \text{extract} \rangle$. Additionally it must hold that the mutual information between the embedded messages M and the messages after extraction M' is greater than zero, $I(M, M') > 0$. This condition implies that the steganographic system is useful in that the recipient learns at least some information about the original messages m . Based on these notions he defines the security of a steganographic communication system as follows.

Definition 2.10 (Perfect Steganographic Security). *A steganographic communication system is called perfectly secure if*

$$\text{KLD}(\mathcal{P}_0 || \mathcal{P}_1) = 0, \quad (2.11)$$

or is called ϵ -secure if

$$\text{KLD}(\mathcal{P}_0 || \mathcal{P}_1) \leq \epsilon. \quad (2.12)$$

Hypothesis testing incorporates two error types. Type I error (false positive) is denoted by α and means that hypothesis H_1 is accepted although H_0 is actually true. Vice versa, a type II error (false negative) denoted by β means accepting H_0 when H_1 is true.

The binary relative entropy $d(\alpha, \beta)$ of two distributions with parameters $(\alpha, 1 - \alpha)$ and $(1 - \beta, \beta)$ is given by:

$$d(\alpha, \beta) = \alpha \log \frac{\alpha}{1 - \beta} + (1 - \alpha) \log \frac{1 - \alpha}{\beta}. \quad (2.13)$$

Because one of the basic properties in hypothesis testing is that deterministic processing cannot increase the relative entropy, it follows that $d(\alpha, \beta) \leq \text{KLD}(\mathcal{P}_0 || \mathcal{P}_1)$. This can be used if an upper bound α^* for type I errors is given, to show that there exists a lower bound for type II errors.

Translated to the steganographic system, this means that if the warden fails to detect a stego object, she makes a type II error (with probability β) and if she decides that a stego object was sent although it was a cover she makes a type I error (with probability α). It follows that in an ϵ -secure stegosystem α and β satisfy $d(\alpha, \beta) \leq \epsilon$. In particular, if $\alpha = 0$ then $\beta \geq 2^{-\epsilon}$. See Table 2.1 for an overview of all possible decisions and the corresponding probabilities of occurrence.

2.2.2.2 Complexity-Theoretic Approach

Katzenbeisser and Petitcolas [47] model the steganographic system as a probabilistic challenge-response protocol, called “game” by the conventions in cryptology, to formalize the advantage of computationally bounded steganalysts. The steganographic system is defined as the triple $\langle \text{gen}, \text{embed}, \text{extract} \rangle$, where gen is the key generation function, embed is the embedding function and extract the extraction function.

The authors formally define the *Steganographic Decision Problem* (SDP) by:

Table 2.1: Error probabilities for different detector outputs

Steganalyst's Decision	Reality	
	cover object	stego object
cover object	correct rejection $1 - \alpha$	false negative β
stego object	false positive α	correct detection $1 - \beta$

Definition 2.11 (Steganographic Decision Problem). *Given $s \in \mathbb{X}^n$, determine if there exists a $\mathbf{k} \in \mathcal{K}$ in the range of \mathbf{gen} and a message $m \in \mathcal{M}$ such that $\mathbf{extract}(s, \mathbf{k}) = m$.*

Here, \mathbb{X}^n is the set of possible covers, k is the shared secret key and \mathcal{M} the set of all possible messages.

The authors introduce an additional player, the Judge. The Judge generates a key and provides an oracle to the steganalyst, who can perform polynomial (in $|\mathbf{k}|$) many queries to the oracle with chosen covers and messages and gets the corresponding stego objects in return. Additionally, she can query the oracle for clean covers. All her queries can be interwoven and depend on all the earlier queries. After finishing her queries, the Judge randomly selects two cover objects and a message and produces a stego object from one of the cover objects and the message. Then, he flips a coin and either forwards the clean cover object or the stego object with equal probability of $1/2$ to the steganalyst. Now, the steganalyst has to solve the SDP for the given object. A steganographic system is called secure if the success probability of the steganalyst is only negligibly⁵ better than random guessing.

This is known as *conditionally secure* as the authors assume the steganalyst to be computationally bounded, i.e., she can only perform polynomial many queries.⁶

2.2.3 Empirical Security Notions

For practical embedding schemes, neither the information-theoretic nor the complexity-theoretic approach are feasible. Thus, to test implementations of different steganographic implementations, several empirical measures have been introduced.

⁵The authors define the success probability to be negligible if it is a negligible sequence n_i , i.e., if for all polynomials p there exists an integer i_0 such that $\forall i \geq i_0 : n_i < 1/p(i)$.

⁶This security notion is inspired and very similar to the notion of adaptive *Chosen Plaintext Attack* (CPA) security in cryptography.

2.2.3.1 Receiver Operating Characteristic

Almost all steganalysis methods base their binary output (cover or stego object) on an internal state of higher precision [8], for example a continuous threshold τ . By adjusting τ , the error rates α and β (as defined in Section 2.2.2.1) can be traded off. When τ varies, *receiver operating characteristic* (ROC) curves, as commonly used in signal detection theory, illustrate the performance of such a binary classifier. A ROC curve plots the false positive rate against the correct detection rate (α vs. $1 - \beta$). As steganalysis can be seen as a binary classification problem, the utilization of these curves is commonly accepted in steganography. Although, it is argued that it might not be possible to compare different steganalysis methods according to their respective ROC curve. For example, if the performance of two steganalysis methods is plotted within the same ROC curve and the lines intercept each other, there is no way to say which of the two methods outperforms the other.

2.2.3.2 Single Number Measures

To circumvent the problem of comparing different performances as ROC curves, several single figures of performance have been introduced. We only show the notions we use in later parts of the thesis and refer the reader to [8, p. 19] for a more detailed list.

- ▶ The *equal error rate* (EER) can simply be read off from the above mentioned ROC curve. It is the one point on the curve where false positive and false negative rate coincide, e.g., $\alpha = \beta$. A lower EER means that less classification errors occur and thus indicates a superior performance of the detector, or conversely, a less secure steganographic function.
- ▶ Another popular choice on how to present steganographic security within a single figure is the *average error rate* (under equal priors⁷) (AER), which is simply calculated from the false positive and the false negative rate divided by 2: $\frac{\alpha + \beta}{2}$.
- ▶ The *false positive rate at 50% detection rate* (FP_{50}) gives the value of α at the fixed point where $1 - \beta$ equals $1/2$.
- ▶ So-called *quantitative* steganalysis methods do not only aim at differentiating stego from cover objects but additionally want to estimate the hidden payload p . The adequate measurement for the performance of such a method is the difference between the actual embedding rate p and the estimation \hat{p} , $|\hat{p} - p|$. As this figure can strongly vary between different images, the common approach in steganography is to report the Mean Absolute Error (MAE) of the estimation over many images. The MAE is the average of all absolute errors. To furthermore give a statement about the robustness of the estimation, most often the MAE is reported together with the Interquartile Range (IQR).

⁷The *equal-prior-assumption* states that cover and stego objects are equally likely on the communication channel. Although this will probably not hold for a real-world scenario, this assumption is tied to the fairness of the Judge from Section 2.2.2.2.

2.3 Summary

In this chapter we have

- ▶ introduced the basic components of a steganographic communication system and, in particular, gave formal definitions of
 - ▷ cover objects,
 - ▷ embedding operation,
 - ▷ embedding strategy, and
 - ▷ stego objects;
- ▶ explained the most common embedding operations and strategies,
- ▶ outlined the different approaches to measure the security of a steganographic system, with means of information theory, complexity theory, and empirical methods.

This chapter is written mainly from the perspective of the steganographer. We now turn to the side of the attacker.

Chapter 3

Exploiting Side Information in Steganalysis

In this chapter we are mainly concerned with steganalysis, the research that aims at detecting the usage of steganography. First, we formally introduce and define the terms of *steganographic side information* and *uncertainty*. Then, we provide initial evidence that side information can be utilized in steganalysis and formalize the area of content-adaptive steganography from the viewpoint of a targeted attack.

Next, we present a method that is statistically almost optimal at detecting random uniform LSB replacement and several enhancements of an approximated version of this method. Finally, we show that generally naïve adaptive embedding is a bad choice when faced with a steganalyst who anticipates it and makes use of her knowledge about the adaptivity criterion.

3.1 Side Information in Steganalysis

In cryptography, the term *side information* is mainly connected with the existence of side channels that can be used to attack a specific implementation of a cryptographic scheme. Side channel attacks are based on (side) information which is gained from the physical devices that run the cryptographic algorithms. Examples of such information are power consumption [60] or latency [61] of the algorithm under observation. Thus, these attacks do not fall in the classical field of cryptanalysis, as they do not search for theoretical weaknesses or perform brute force attacks. In steganography, we have an additional input parameter, the empirical cover object, that has no direct counterpart in cryptography. As empirical media allow for different interpretations of side information, the notion of side information in steganography is inconsistent.

3.1.1 Steganographic Side Information and Uncertainty

There are notions of side information in steganography that resemble the information leakage similar to cryptography. For example, the attacker may get access to the device used to capture the cover objects and thus the sensor noise pattern [35]. Then, sometimes the term side information is used as something that is emitted from the (pre-)cover object, as in the example of PQ steganography from Section 2.1.3.3.3. Sometimes, authors utilize side information, but do not explicitly mention the term. For example, in [24] the authors explicitly assume that Eve has access to (a version of) the cover object and thus can compare the potential stego object with this cover, but this information

flow is not called “side information” by the authors.

We set out to remove inconsistencies concerning *side information* and the term of *uncertainty*. We provide a rigorous definition and differentiate between unconditionally and conditionally perfect side information, a concept we borrow from cryptography, where unconditional secure encryption is provably secure even against a computationally unbounded attacker.

3.1.1.1 Steganographic Side Information

Definition 3.1 (Steganographic Side Information). *A source of steganographic side information (SSI) Θ is an information source that is fully available to the steganographer and the use of which is defined in the embedding function. When the side information is exclusively available to the steganographer and not to the steganalyst, we speak of perfect steganographic side information; if the steganalyst is able to (partially) reconstruct it, we speak of (partially) reconstructible steganographic side information.*

We specify two concrete types of SSI. The SSI is called ...

1. *unconditionally perfect: if there exists no mutual information between the stego objects and the source of side information, i.e., $I(\mathbf{X}^{(1)}, \Theta) = 0$, and*
2. *conditionally perfect: if the availability of the side information relies on an assumption about the computational power or the model of the steganalyst.*

This definition requires some reflection:

Remark 3.1. *A similar approach to measure leakage about the shared secret key with mutual information in a general data hiding framework is presented in [70].*

Remark 3.2. *Steganographic side information is no secret in the sense of a cryptographic secret, e.g., stemming from a shared secret key. It is available to Alice, but not (fully) under her control.*

The connection of SSI and practical adaptivity criteria is straightforward.

Remark 3.3. *All practical adaptivity criteria are instances of SSI.*

Remark 3.4. *The way steganographic side information is interpreted for a given cover realization is part of the embedding function and has to be assumed to be known to Eve. However, how much of the steganographic side information used for embedding is reconstructible from the stego object also depends on the embedding function.*

In Section 2.1.3.3 we have seen several examples of how Alice can utilize SSI in her embedding function. While the SSI, together with the processing of it, in content-adaptive embedding belongs mainly to the class of (partially) reconstructible SSI, PQ steganography takes one step forward to perfect side information as the SSI is (mostly) lost during the information-reducing process.

But, the authors of PQ steganography already suggest that the embedding positions with the lowest distortion should additionally be confined to areas that are more suitable, as defined by a content-adaptive criterion.

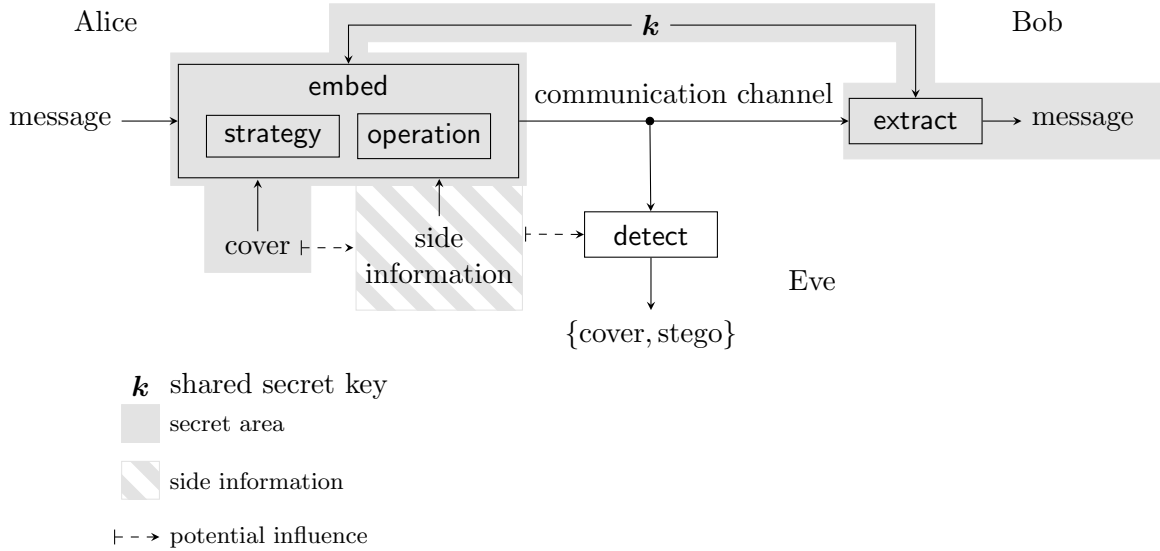


Figure 3.1: Block diagram of a steganographic system with side information

Remark 3.5. *Steganographic side information which has no relation to the content of a given cover object leads to embedding positions randomly spread across the cover.*

To stay with the example of PQ steganography, if we assume that the quantization levels are spread uniformly across the cover object, the selection of positions with values close to $1/2$ before embedding will be spread uniformly over the whole cover object, similar to a pseudo-random path.

Remark 3.6. *Until now, the common use of side information in steganography research is as if it was perfect SSI, because it is assumed to be only utilized by Alice. Thereby the fact that it might be, at least partially, reconstructible by Eve, as it often originates from either the cover objects' content or some publicly available parameters, is disregarded.*

For example, the leading textbook on steganography states that “Often, the values [...] are computed from [...] side-information that is not available to Bob.” [27, p. 167] and thus not available to Eve, either. Even the most recent publication on the topic of side information by the same author, entitled *On the Role of Side Information in Steganography in Empirical Covers*, does not consider the possibility of the side information being reconstructed by Eve. But the author acknowledges that: “The failure of current steganalysis to reliably detect side-informed schemes should, however, be taken with a grain of salt because it could simply mean that current steganalysis lacks the right models (feature spaces).” [29, p. 86650I-1] The author of this statement believes in machine learning-based steganalysis, where the model or feature space refers to the collections of features that are used to characterize the data.

Figure 3.1 shows the basic steganographic communication system augmented with side information. As indicated by the striped area, side information itself may not

fully belong to the secret area and the dashed arrows indicate that the cover object potentially influences the side information, which in turn might influence the detection decision of Eve. For perfect SSI, the striped area would completely belong to the trusted area and the influence of the side information on the detection decision would disappear. For the sake of clarity we do not include the possible impact of the secret key on the usage of the side information, but stress that it may exist.

3.1.1.2 Uncertainty

Steganographic security is strongly tied to ensuring uncertainty on Eve's side. Perfect steganographic security, in the sense of Definition 2.10, can be described such that Eve is completely uncertain if an analyzed object originates from the cover distribution \mathcal{P}_0 or the stego distribution \mathcal{P}_1 . Uncertainty also ensures the protection goal of undetectability, as it can be seen as the absence of detectable artifacts in stego objects.

Definition 3.2 (Uncertainty). *Uncertainty in a steganographic system quantifies the steganalyst's lack of knowledge about the type of object, cover or stego, she faces. With perfect uncertainty, the steganalyst's only option is to guess. With no uncertainty, i.e., perfect information, the steganalyst always knows the type of the object exactly.*

Formally, for a given object \mathbf{x} we have ...

- ▶ perfect uncertainty, if

$$\Pr(\mathbf{x}|\mathbf{x} \sim \mathbf{X}^{(0)}) = \Pr(\mathbf{x}|\mathbf{x} \sim \mathbf{X}^{(1)}) > 0 \quad (3.1)$$

- ▶ perfect information, if

$$\Pr(\mathbf{x}|\mathbf{x} \sim \mathbf{X}^{(q)}) = 0 \wedge \Pr(\mathbf{x}|\mathbf{x} \sim \mathbf{X}^{(1-q)}) > 0 \text{ for } q \in \{0, 1\}. \quad (3.2)$$

For a derivation of the above definition with means of information theory see Appendix A.1. The general definition of uncertainty can be broken down to single positions.

Definition 3.3 (Uncertainty with Regard to Positions). *We denote the uncertainty with regard to positions as the steganalyst's lack of knowledge about the likelihood of values at single positions x_i in an object \mathbf{x} .*

The position x_i is called ...

- ▶ a perfectly uncertain position, if

$$\Pr(x_i = u|\mathbf{x} \sim \mathbf{X}^{(0)}) = \Pr(x_i = u|\mathbf{x} \sim \mathbf{X}^{(1)}) > 0, \quad (3.3)$$

for at least two values $u \in \{0, \dots, 2^\ell - 1\}$, and

- ▶ a perfectly informative position, if

$$\Pr(x_i = u|\mathbf{x} \sim \mathbf{X}^{(q)}) = 0 \wedge \Pr(x_i = u|\mathbf{x} \sim \mathbf{X}^{(1-q)}) > 0 \text{ for } q \in \{0, 1\}, \quad (3.4)$$

and $\forall u \in \{0, \dots, 2^\ell - 1\}$.

Remark 3.7. *The condition in Equation (3.3) includes the realization of all positions in \mathbf{x} , not only the specific value of x_i .*

Remark 3.8. *We need at least two values u_1, u_2 for which Equation (3.3) holds. With only one value u_1 , the position would no longer be perfect for Alice if this value occurred at position i in a given cover realization, as she would have to change this value to one for which Equation (3.3) does not hold.*

Definitions 3.2 and 3.3 are directly tied to information-theoretically secure steganography.

Remark 3.9. *Only if perfect uncertainty holds, the embedding is information-theoretically secure, in the sense of Definition 2.10.*

See Appendix A.2 for a proof of this remark. The result illustrates how hard it is to achieve perfect steganographic security in the sense of Definition 2.10 for practical embedding functions.

Remark 3.10. *If the steganographer embeds only in perfectly uncertain positions with an embedding operation for which Equation (3.3) holds, the embedding is information-theoretically secure.*

This follows because if Equation (3.3) holds for all positions, Equation (3.1) holds for the whole object and thus, the embedding is information-theoretically secure.

Remark 3.11. *Ideally, SSI should help Alice to find the most uncertain positions in a given cover object.*

The separation between perfectly uncertain and perfectly informative positions is similar to the *cover composition approach* used in [8, p. 104], which argues that all cover objects are composed of an indeterministic part, necessary for steganographic security, and a deterministic part, necessary for the plausibility of the cover. Furthermore, this decomposition of cover objects is one of the underlying building blocks of model-based steganographic methods, as introduced for example in [76].

Remark 3.12. *As adaptivity criteria are instances of SSI and uncertainty is tied to the protection goal of undetectability, we can directly translate the notion of uncertainty with regard to positions to the notion of suitability as defined by empirical adaptivity criteria., i.e., more uncertain positions should induce more suitable positions.*

3.1.1.3 Common Use of Side Information in Steganography

Most side-informed embedding functions use the side information to find the best embedding strategy. It is also commonly accepted that granting Eve knowledge about the embedding positions will result in less secure steganography, as she might be able to compare statistical properties of the positions used for embedding with those excluded [34].

Occasionally, there are embedding functions that use side information to implement an adaptive embedding operation. For example, in the HUGO algorithm [71] the embedding operation measures the distortion induced by changing the LSB by either $+1$ or -1 and chooses the direction which introduces less distortion.⁸ Or, in [91] the embedding operation uses more than the LSB to hide payload per position if the position is in a supposedly noisier area.

In the remaining part of the thesis we restrict ourself to side-informed embedding strategies, but the extension to the embedding direction or depth is an interesting avenue to pursue.

3.1.2 Initial Evidence

This section shows that side information and uncertainty are already present in steganographic literature and gives two examples that motivate further research on both properties of a steganographic scheme. The first example is a steganographic scheme with reconstructible steganographic side information and a targeted attack on it. The second approach is closely related to our notion of uncertainty.

3.1.2.1 Targeted Attack on PSP Steganography

In 2002, Franz published a steganographic algorithm called “Preserving Statistical Properties” (PSP) [23] which was designed to withstand the chi-square attack of [89]. For this, the PSP scheme introduces two modifications in comparison to LSB replacement. First, the image to embed is divided in sets of pixels S_k with the same shade k , i.e., the same grayscale value. Then, sets that only differ in their LSB are summarized into groups $G_k := S_{2k} \cup S_{2k+1}$. These groups are further divided into *good* groups \mathcal{G}^+ and *bad* groups \mathcal{G}^- , and only the good groups are used for embedding. To identify the good groups, a within-group dependency test is run on the co-occurrence matrices⁹ C . It is specified in the PSP algorithm that 4 co-occurrence matrices are tested and, if one of the test fails, i.e., detects within-group dependencies, the whole group is classified as bad. The second modification the PSP algorithm introduces is that it overwrites the LSBs with exactly the same distribution that is found in the cover object. By this, the first order statistics will be exactly preserved.

As Böhme and Westfeld state in [10] both methods reduce the capacity of the cover image significantly, but make it indeed secure against chi-square attacks. To attack the PSP algorithm nonetheless, the idea of the authors is to evaluate the between-group dependencies instead of the preserved within-group dependencies. For this to be possible, it is necessary that an attacker can exactly reconstruct which of the groups are good

⁸To the best of our knowledge this is the only attempt to use adaptive LSB matching, one reason might be that already in [35] it is shown that the direction can be estimated better than random guessing, another that the usage of ternary embedding yields a higher capacity and is impossible with adaptive LSB matching.

⁹A co-occurrence matrix is a transition histogram between adjacent pixels for a defined relation in the spatial domain and contains the frequency of a certain shade depending on the shade of a defined neighbor.

and which of them are bad. But, as the second modification introduced by the PSP algorithm exactly preserves the relevant statistics, an attacker can recalculate the initial classification of good and bad groups. We can see this as fully reconstructible SSI. Based on the assumptions that adjacent pixels correlate strongly, the authors are able to perform a between-group dependency test for all good groups and compare it to a certain threshold.

Empirical tests show that this attack can perfectly detect the use of PSP steganography. Even more, when LSB embedding is used with the reduced capacity of the PSP algorithm (on average 77% of the original capacity), LSB embedding is more secure against the chi-square attack the PSP was developed to withstand. As far as we know, this was the first targeted attack that made use of side information not only available to the steganographer.

3.1.2.2 The Detectability Profile

In [26], Fridrich sets out to study the trade-off between the number of embedding changes and their amplitude for secure steganographic systems. By this, she creates a *detectability profile* that is closely connected to our notion of (position-wise) uncertainty from above. Fridrich argues that every practical embedding operation introduces distortion to the cover object and examines strategies to minimize the overall distortion. She assigns a scalar value, the *detectability measure* ρ_i to every position i . By sorting the values from the smallest to the largest the non-decreasing detectability profile $\boldsymbol{\rho} = (\rho_1, \dots, \rho_n)$ is created. Furthermore, she assumes that the distortion introduced by changing several positions of the cover object is additive, so the overall distortion introduced by changing the set $\{j_1, \dots, j_q\}$ is $\sum_{j=1}^q \rho_{i_j}$. An appropriate measure for ρ_i always depends on the context, for the example in Section 2.1.3.3.3, $\rho_i = 1 - 2\varepsilon_i$ would be good if the steganographer wanted to minimize the additional embedding distortion.

Fridrich mentions that the intuitive approach for a steganographer who wants to minimize the overall embedding impact is to use only the positions that will introduce the least detectable artifacts when changed. She approaches the selection of embedding positions from the viewpoint of the absolute number necessary to change versus the introduced distortion. Her main argument against this naïve approach is that allowing the embedding function to use more positions implies the possibility of syndrome encoding and thus decreases the number of embedding changes. Furthermore, Fridrich states that if the ρ_i are uniform in the cover object, the best strategy is to use all positions because this would allow us to minimize the total number of necessary embedding changes.

Her numerical analysis under the assumption that optimal coding schemes exist reveals that it is never optimal to choose the q positions with the smallest detectability measure. Furthermore, for any detectability profile $\boldsymbol{\rho}$ it is optimal to use all positions for embedding, given a message of a certain length.

3.1.3 Formalizing Adaptive Steganography and Steganalysis

Remark 3.3 states that all practical adaptivity criteria are instances of steganographic side information. An adaptivity criterion ranks embedding positions in the cover by the risk of being detected. Thus, it tries to identify areas where embedding changes result in small distortion or similarly, the most uncertain embedding positions compared to the rest of the cover. However, most of the adaptivity criteria used in the literature lack a sound justification that the embedding positions they select indeed reduce detectability.

Furthermore, the implicit assumption that steganalysts are unaware of the fact that the embedding function is content-adaptive is overly optimistic and violates Kerckhoffs' principle, as the usage of the SSI is described in the embedding function and thus must be assumed to be known to the steganalyst.

With this knowledge, the steganalyst might try to recalculate the values of the adaptivity criterion from the stego image (leading to a rough approximation), or even find ways to estimate the values more precisely than mere recalculation using knowledge about the impact of the embedding operation. Formally this means that the steganalyst uses a function $\zeta'(\cdot) \neq \zeta(\cdot)$ which might give him a more accurate estimation of the original values of the positions than $\zeta(\cdot)$ itself. This leads to the following remark.

Remark 3.13. *To err on the side of caution, one should consider to base security analyses on the premise that the steganalyst has knowledge about the exact values of the adaptivity criterion for all pixels.*

It is useful to introduce some formalism to state such rationales more precisely. Let \mathbf{x} denote a cover or stego object in its natural order, then \mathbf{y} denotes the same symbols sorted by the adaptivity criterion $\zeta(\cdot)$ with the convention that $\zeta(y_j) \geq \zeta(y_{j+1})$ ¹⁰ for $j \in \{0, \dots, n-2\}$. So, in \mathbf{y} the *most suitable* symbols are at the beginning and the *least suitable* symbols at the end. Naïve adaptive embedding would use symbols $(y_0^{(0)}, \dots, y_{p \cdot n-1}^{(0)})$ to embed a payload of length $p \cdot n$ ($0 \leq p \leq 1$) into the cover object. Note that this does not imply that the embedding path is fully deterministic. The embedding function would still distribute the actual embedding changes according to a key-dependent pseudo-random path through the leading symbols in \mathbf{y} .

In content-adaptive embedding, we assume that the SSI is used to identify the most suitable embedding positions, which can be tied to the notion of uncertainty via Remark 3.12. With this assumption, we can specify the notion of reconstructible SSI into a notion tailored to the area of content-adaptive embedding.

As the stated goal is to *recover* the embedding positions with the help of some kind of *reconstructible* SSI, we call embedding positions *recoverable* instead of reconstructible.

Definition 3.4 (Perfect Recoverability). *An adaptivity criterion $\zeta(\cdot)$ is perfectly recoverable (for embedding rate p) if it is invariant to the embedding function, i. e., if it holds that*

$$\zeta(x_i^{(0)}) = \zeta(x_i^{(p)}), \forall i \in \{0, \dots, n-1\}. \quad (3.5)$$

¹⁰Note, to simplify the notation, we skip the explicit argument of the entire vector and write $\zeta(y_j) = \zeta(\mathbf{y}, j)$.

An adaptivity criterion that has perfect recoverability is fully reconstructible in the sense of Definition 3.1.

Definition 3.5 (Order Recoverability). *An adaptivity criterion has a recoverable order (for embedding rate p), if $\forall(i, j) \in \{0, \dots, n-1\}^2$ it holds that:*

$$\zeta(x_i^{(0)}) < \zeta(x_j^{(0)}) \Rightarrow \zeta(x_i^{(p)}) < \zeta(x_j^{(p)}). \quad (3.6)$$

It is questionable if order recoverability falls into the category of fully or partially reconstructible SSI. In the context of content-adaptive embedding, we argue that, the order of the embedding positions is most important and thus, order recoverability also belongs to fully recoverable SSI.

Obviously, perfect recoverability implies order recoverability. However, already the weaker condition may be sufficient for the steganalyst to substantially gain detection performance.

Proposition 3.1 (Reduction to sequential embedding). *If the steganalyst can recover the order of $\zeta(x_i^{(0)})$ from $\mathbf{x}^{(p)}$, then the detection problem for naïve adaptive embedding reduces to the problem of detecting initial sequential embedding.*

The proposition follows readily from Definitions 2.5 and 3.5 and is supported by Figure 2.3(d) (on page 15).

3.2 Powerful Steganalysis of LSB Replacement

The LSB replacement embedding operation, as introduced in Section 2.1.2.1, is probably the oldest and best-understood embedding operation in digital steganography. Although all modern embedding functions implement other embedding operations, insights drawn from the study of LSBR are helpful to better understand the interplay between the changing of symbols in the cover object and detectability. Thus, we restrict our following analysis mainly to LSBR.

3.2.1 Asymptotically Uniformly Most Powerful Test

In 2012 Fillatre [18] presented an almost optimal statistical test to detect LSB replacement. Utilizing the hypotheses test of [11] (cf. Section 2.2.2.1) he designs an *Asymptotically Uniform Most Powerful* (AUMP) test for the detection of random uniform LSB replacement.

One of his motivations is to create a detection method that warrants a prescribed probability of false alarm α .

To reformulate Equation (2.10) with $\mathbf{x}^{(p)}$ being a potential stego object, the two composite hypotheses H_0 and H_1 are:

$$\begin{aligned} H_0 : \mathbf{x}^{(p)} &\sim \mathcal{P}_0 \\ H_1 : \mathbf{x}^{(p)} &\sim \mathcal{P}_1, \forall p \in (0; 1] \end{aligned} \quad (3.7)$$

The goal is to find a test $\varphi\{0, \dots, 2^\ell - 1\}^n \rightarrow \{H_0, H_1\}$ such that hypothesis H_i is accepted, if $\varphi(\mathbf{x}^{(p)}) = H_i$. Fillatre bases his test on the assumption that the image under investigation can be modeled with the help of a zero-mean independent Gaussian noise variable ξ_i . He assumes that for every pixel intensity x_i it holds that:

$$x_i = l_i + \xi_i, \quad (3.8)$$

where l_i is the mathematical expectation of the pixel x_i . By this, the random variable x_i follows a Gaussian distribution with probability density function (PDF)

$$f_{\Xi_i}(x_i) = \frac{1}{\sqrt{2\pi\epsilon_i^2}} e^{-\frac{(x_i - l_i)^2}{2\epsilon_i^2}}, \quad (3.9)$$

and is entirely characterized by $\Xi_i = (l_i, \epsilon_i)^T$. To simplify notation, $\boldsymbol{\omega}$ is defined as the vector holding the parameters of all n pixels and Ω_n denotes the set of all possible parameters $\boldsymbol{\omega}$. Furthermore, Δ_n denotes the quantization step, $\Phi(\cdot)$ the Gaussian cumulative distribution function (CDF) and $\Phi^{-1}(\cdot)$ its inverse.

Let

$$\mathcal{K}_\alpha = \{\varphi : \sup_{\boldsymbol{\omega} \in \Omega_n} \mathcal{P}_0(\varphi(\mathbf{x}^{(p)}) = H_1) \leq \alpha\} \quad (3.10)$$

be the class of tests with an upper-bound false alarm probability α . Then, the corresponding power function $\beta_\varphi(\boldsymbol{\omega}, p)$ is defined by the probability for correct detection:

$$\beta_\varphi(\boldsymbol{\omega}, p) = \mathcal{P}_1(\varphi(\mathbf{x}^{(p)}) = H_1). \quad (3.11)$$

The hypotheses H_0 and H_1 are composite. This means that one of the hypotheses depends on an unknown parameter, the embedding rate p in our case. In general there is no way to design an optimal test for composite hypotheses. The solution is to design an Uniformly Most Powerful (UMP) test which uniformly maximizes the power function with respect to $\boldsymbol{\omega}$ and p .

This is a test $\varphi^* \in \mathcal{K}_\alpha$ whose power function β_{φ^*} satisfies

$$\beta_{\varphi^*}(\boldsymbol{\omega}, p) = \sup_{\varphi \in \mathcal{K}_\alpha} \beta_\varphi(\boldsymbol{\omega}, p), \forall \boldsymbol{\omega} \in \Omega_n, \forall p \in (0, 1]. \quad (3.12)$$

Unfortunately, these tests rarely exist in practice. Fillatre follows the way to design an asymptotically UMP test, as the quantization step Δ_n vanishes, when n tends to infinity. The AUMP test is defined as follows (Definition 1 in [18]):

Definition 3.6 (AUMP Test). *Let $0 < \alpha < 1$. The test $\varphi^*(\mathbf{x}^{(p)})$ is AUMP in the class \mathcal{D}_α , given as*

$$\mathcal{D}_\alpha = \{\varphi : \lim_{n \rightarrow \infty} \sup_{\boldsymbol{\omega} \in \Omega_n} \Pr_{\mathcal{P}_0}(\varphi(\mathbf{x}^{(p)}) = H_1) \leq \alpha\},$$

to decide between H_0 and H_1 if the following two requirements are satisfied:

1. $\varphi^* \in \mathcal{D}_\alpha$;

2. $\limsup_{n \rightarrow \infty} (\beta_\varphi(\boldsymbol{\omega}, p) - \beta_{\varphi^*}(\boldsymbol{\omega}, p)) \leq 0$ for any $\boldsymbol{\omega} \in \Omega_n$ and $p \in (0, 1]$, for all other tests $\varphi \in \mathcal{D}_\alpha$.

Theorem 1 in [18] shows that under some assumptions about the image model, the following test is AUMP:

$$\varphi^*(\mathbf{x}^{(p)}) = \begin{cases} H_0 & \text{if } \Lambda^*(\mathbf{x}^{(p)}) \leq \lambda^* \\ H_1 & \text{else,} \end{cases} \quad (3.13)$$

where

$$\Lambda^*(\mathbf{x}^{(p)}) = \sum_{i=0}^{n-1} w_i (x_i^{(p)} - l_i)(x_i^{(p)} - \bar{x}_i^{(p)}) \text{ with } w_i = \frac{\bar{\sigma}_n}{\epsilon_i^2 \sqrt{n}}, \quad (3.14)$$

and

$$\lambda^* = \Phi^{-1}(1 - \alpha). \quad (3.15)$$

Here, $\bar{x}_i^{(p)}$ denotes the position i with LSB flipped and $\bar{\sigma}_n$ denotes the square root of $\bar{\sigma}_n^2$, the mean variance of the image, defined by:

$$\frac{1}{\bar{\sigma}_n^2} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\epsilon_i^2}. \quad (3.16)$$

If the cover parameters, namely $\boldsymbol{\omega}$, are unknown, the parameters w_i, l_i, ϵ_i and $\bar{\sigma}_n$ have to be estimated and will be replaced by their estimates $\hat{w}_i, \hat{l}_i, \hat{\epsilon}_i$ and $\hat{\sigma}_n$ in Equation (3.14). With these estimates, Fillatre proposes the following adaptive AUMP test in Theorem 2 of [18]:

$$\hat{\varphi}^*(\mathbf{x}^{(p)}) = \begin{cases} H_0 & \text{if } \hat{\Lambda}^*(\mathbf{x}^{(p)}) \leq \hat{\lambda}^* \\ H_1 & \text{else,} \end{cases} \quad (3.17)$$

where

$$\hat{\Lambda}^*(\mathbf{x}^{(p)}) = \sum_{i=0}^{n-1} \hat{w}_i (\tilde{x}_i^{(p)} - \hat{l}_i)(x_i^{(p)} - \bar{x}_i^{(p)}) \text{ with } \hat{w}_i = \frac{\hat{\sigma}_n}{\hat{\sigma}_k^2 \sqrt{K_n(m-q)}}, \quad (3.18)$$

and

$$\hat{\lambda}^* = \Phi^{-1}(1 - \alpha). \quad (3.19)$$

σ_k^2, K_n, m and q are parameters chosen per image, whose meaning is not relevant for the overall result.

By this, Fillatre *a posteriori* justifies the good performance of a class of LSB replacement detectors much older, namely the Weighted Stego-Image (WS) detectors, introduced by Fridrich and Goljan in [31] and improved by Ker and Böhme in [56].

We present the original WS approach and some of its important modifications in the following sections.

3.2.2 Weighted Stego-Image Steganalysis

In 2004 Fridrich and Goljan presented the original form of WS, from now on called standard WS, as a mathematically well-founded minimization problem [31]. WS is a quantitative steganalysis method, i.e., aiming to estimate the length of the hidden message. The authors generate a so-called *weighted stego image* $\mathbf{x}^{(p,\lambda)}$, which then allows for the estimation of the length of the hidden message. The basic idea is, similar to Fillatre's approach above, to generate an object with all LSBs flipped (thus simulating an embedding in all locations) and then finding the nearest distance between this WS image and the object under suspicion.

Following [56], we define $\mathbf{x}^{(p,\lambda)}$ as:

$$\mathbf{x}^{(p,\lambda)} = \lambda \bar{\mathbf{x}}^{(p)} + (1 - \lambda) \mathbf{x}^{(p)}, \quad (3.20)$$

where λ describes the weighting and $\bar{\mathbf{x}}^{(p)}$ denotes the stego image with every element's LSB flipped.

Theorem 1 in [31] states that the Euclidean distance between $\mathbf{x}^{(p,\lambda)}$ and $\mathbf{x}^{(0)}$ is minimized for $\lambda = q/(2n)$. As the cover is unknown to an attacker, she has to estimate it from the stego image. This can be achieved by using a linear filter, i.e., a weighted average of the local neighborhood. An example for such a filter is presented by Ker and Böhme [56] using a filter of the form (3.21).

$$\begin{array}{ccc} -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{array} \quad (3.21)$$

By differentiating the Euclidean distance for λ , we can estimate p using Equation (3.22).

$$\hat{p} = \frac{2}{n} \sum_{i=1}^n \left(x_i^{(p)} - \hat{x}_i^{(0)} \right) \left(x_i^{(p)} - \bar{x}_i^{(p)} \right) \quad (3.22)$$

Several improvements of this method have been proposed. Most notably is Weighted WS steganalysis which adds element weights to the stego image. This second weighting takes differences in local predictability into account. Elements which can be estimated with high confidence contribute more to the estimation than elements where errors are expected:

$$\hat{p} = 2 \sum_{i=1}^n w_i \left(x_i^{(p)} - \hat{x}_i^{(0)} \right) \left(x_i^{(p)} - \bar{x}_i^{(p)} \right), \quad (3.23)$$

using $\sum_{i=1}^n w_i = 1$.

Fridrich and Goljan [31] propose weights of the form $w_i^{-1} \propto 1 + \sigma_i^2$; experimental results in [56] suggest that $w_i^{-1} \propto 5 + \sigma_i^2$ provide more accurate estimates for two large

image databases. In both cases, σ_i^2 denotes the local variance in the neighborhood of pixel i (but excluding the center pixel). Although some rationales for the choice of weights are given in [8], it is fair to say that the optimal choice of WS weights for real images is not sufficiently understood, even in the case of a uniform random distribution of embedding positions. Both basic methods assume that changes in the cover are spread uniformly [56].

For a WS variant tailored to JPEG covers see [6] and for the extension of it to mod- k replacement instead of LSB replacement, we refer the reader to [93].

3.2.2.1 WS Steganalysis for Sequential Embedding

Ker [51] proposes a WS steganalysis variant tailored to initial sequential embedding. He decomposes Equation (3.22) into two parts, reflecting that embedding changes only occur in the first elements. The first elements are therefore weighted using $\lambda = 1/2$ while the remaining elements are weighted with $\lambda = 0$. Note that in this scenario any change to the weighting, e.g., based on local predictability, will degrade the estimation. This is the case because it is certain that the first elements contain the hidden message. The resulting estimator (Eq. (3.24)) is minimized for the point where the embedding ends, i.e., $k = q$ [51].

$$E(k) = \sum_{i=1}^k \left(\frac{1}{2} \left(x_i^{(p)} + \bar{x}_i^{(p)} \right) - \hat{x}_i^{(0)} \right)^2 + \sum_{i=k+1}^n \left(x_i^{(p)} - \hat{x}_i^{(0)} \right)^2 \quad (3.24)$$

This approach outperforms the previously introduced detectors when applied to initial sequential embedding [51]. However, Equation (3.24) cannot simply be differentiated because its derivative has no closed form and can have multiple local minima. To solve this, Ker [51] proposes the following recurrence:

$$e_0 = 0 \\ e_k = e_{k-1} + \left(\frac{1}{2} \left(x_{k-1}^{(p)} + \bar{x}_{k-1}^{(p)} \right) - \hat{x}_{k-1}^{(0)} \right)^2 - \left(x_{k-1}^{(p)} - \hat{x}_{k-1}^{(0)} \right)^2, \quad (3.25)$$

which generates $e_k = E(k) - \sum_{i=0}^{n-1} \left(x_i^{(1)} - \hat{x}_i^{(0)} \right)^2$. Because the last term is constant, the minimum term of e_k coincides with the minimum of $E(k)$, and so, only linear time is required to generate and examine the sequence e_k .

3.2.2.2 WS Steganalysis for Naïve Adaptive Embedding

Two possible approaches come to mind for tailoring the WS method to detect adaptive embedding. First, we could modify the local weights w_i and insert the inverse suitability

for embedding, based on an estimate of the adaptivity criterion $\hat{\zeta}(x_i^{(0)})$. If this estimation is good enough, positions which are more likely to be changed during embedding will get higher local weight and influence the overall decision more, and vice versa. As mentioned above, the choice of optimal weights is still not perfectly understood and all practical proposals are validated only experimentally. In addition, a good adaptivity criterion is supposed to identify regions which are hard to predict, so the positions preferred for embedding coincide with those where estimates $\hat{x}_i^{(0)}$ are poor. In fact, the superiority of weighted WS over unweighted WS stems from assigning more weight to better predictable positions. This advantage collapses as soon as the embedding function hides exclusively in the less predictable positions. For these reasons, we do not pursue this approach.

Our specialized WS variant leverages Proposition 3.1 (on page 33) and, as we shall see, enables almost perfect detection of naïve adaptive embedding in cases where the known WS methods fail.

Our approach is an elegant modification of the WS method to initial sequential embedding from the last section. For naïve adaptive embedding, Proposition 3.1 suggests a clear procedure. Upon receiving or intercepting a potential stego object $\mathbf{x}^{(p)} = (x_0^{(p)}, \dots, x_{n-1}^{(p)})$ the steganalyst tries to recover the order of $\zeta(x_i^{(0)})$ and obtains $\mathbf{y}^{(p)} = (y_0^{(1)}, \dots, y_{p \cdot n - 1}^{(1)}, y_{p \cdot n}^{(0)}, \dots, y_{n-1}^{(0)})$. Small errors in the recovery of the order can be tolerated.

With this notation, Equation (3.24) translates to

$$E(k) = \sum_{i=0}^{k-1} \left(\frac{1}{2} (y_i^{(p)} + \bar{y}_i^{(p)}) - \hat{y}_i^{(0)} \right)^2 + \sum_{i=k}^{n-1} (y_i^{(p)} - \hat{y}_i^{(0)})^2 \quad (3.26)$$

and is minimized at $k = p \cdot n$.

3.3 A Targeted Attack on Naïve Adaptive Embedding

In this section we leverage Proposition 3.1 and the WS steganalysis tailored to naïve adaptive embedding from Section 3.2.2.2 to attack several practical embedding algorithms. First, we give an overview of widely used adaptivity criteria, then we outline our set-up, before we define the evaluation strategy and finally the results of our attack.

3.3.1 Overview of Adaptivity Criteria

In this section we provide an overview of the adaptivity criteria most commonly used in practice. As there is a real abundance of content-adaptive embedding schemes, we point out that this overview is not exhaustive, but captures the most prevalent approaches. Furthermore, we restrict our overview on algorithms designed for digital images, as these are the main topic of this thesis.

Table 3.1: Overview of different adaptivity criteria for content-adaptive embedding and their respective embedding operations.

Adaptivity Criterion	Name (Source)	Embedding Operation	Embedding Strategy	Recovery Intended [‡]	Tested Attacks
Edges	Singh [84]	LSBR	random ad.	Yes	SP
	Hempstalk [40]	LSBR	random ad.	Yes	ML
	Hussain [43]	LSBR	naïve ad.	Yes	-
PVD	Wu [90]	k-Bit	naïve ad.	Yes	RS
	Wang [88]	k-Bit	naïve ad.	Yes	RS
	Yang [92]	k-Bit	naïve ad.	Yes	RS,SP
Variance	Fridrich [30]	Block LSBR	r-random	Yes	-
	Luo [65]	k-Bit	r-random	Yes	RS
Texture	Fridrich [30]	Block LSBR	r-random	Yes	-
	Franz [25]	LSBR	r-random	Yes	χ^2 ,RS
	Pramitha [73]	2-LSBR	naïve ad.	Yes	χ^2
Distortion-Minimizing	HUGO [71]	LSBM	STC	No	ML
	WoW [41]	ternary	STC	No	ML
	S-Uniward [15]	ternary	STC	No	ML
	UED [38]	ternary	STC	No	ML

Block LSBR: blockwise LSBR; naïve : r-random: restricted random embedding; STC: Syndrome Trellis Codes; ML: machine learning-based; RS: regular/singular analysis; SP: sample pair analysis;

[‡] Yes, if the recipient needs to recover the exact order

Table 3.1 is divided into five different types of adaptivity criteria, but we point out that the schemes in the category titled *PVD* (Pixel Value Differencing) are strongly related to the category of edge sensitive embedding algorithms. We list the original publications and the respective embedding operations and strategies in the table, together with the property if the recipient has to recover the exact order of the adaptivity criterion, i.e., if the scheme has to ensure fully reconstructible SSI.

Finally, the rightmost column shows the attacks the authors considered before publishing their scheme. As can be seen, most of the schemes are evaluated with benchmark steganalysis methods that do not anticipate the adaptivity. Excluding the machine-learning based case, none of the tested attacks utilizes knowledge about the embedding strategy.¹¹ Furthermore, RS analysis [32] proposed in 2001, SP analysis [16] proposed in 2003, and the χ^2 test [89] proposed in 2000 are explicitly developed to detect the use of random uniform LSBR. This bears the risk of substantially overestimating the security.

3.3.2 Data and Set-up

We use the BOSSBase [5] for the empirical evaluation of our specialized WS method. This image collection consists of 10 000 grayscale images, each downscaled to 512×512

¹¹There is an ongoing discussion in the research community if machine learning-based steganalysis captures features connected to the adaptivity criteria by definition or not.

pixels. For robustness, we also performed all experiments on 700 raw images of the Dresden Image Database (DIB) [36], cropped to 512×512 pixels. As the results are very similar, qualitatively and quantitatively, our findings seem unlikely to be an artifact of the homogeneous pre-processing (scaling) applied to all images in the BOSSBase.

As the WS method is designed to detect LSB replacement only, our approach is to borrow four different adaptivity criteria from Table 3.1 and use them in combination with LSB replacement. We apply naïve adaptive embedding, i. e., we use the first $p \cdot n$ pixels to embed a payload of random bits. After calculating the values of $\zeta(\cdot)$ for every pixel, we order the pixels according to it, most suitable first, to obtain $\mathbf{y}^{(0)}$ and simulate embedding by flipping 50% of the leading $p \cdot n$ positions in \mathbf{y} . The permutation of $\mathbf{y}^{(p)}$ is inverted to obtain $\mathbf{x}^{(p)}$. To eliminate boundary conditions, we exclude the image borders from embedding and detection attempts.

If a perfect order cannot be established in the neighborhood of $y_{p \cdot n}$, which may happen if $\zeta(\cdot)$ is discretized to the same value for a sequence of elements in $\mathbf{y}^{(0)}$, we distribute the embedding positions uniformly over all elements sharing the same value.

Furthermore, we do not implement methods to preserve the exact order of the adaptivity criterion in the stego object, although this might be specified in the original algorithm. As all recently proposed embedding schemes utilize non-shared selection channels the recovery of the order is not needed at the side of the recipient. By this, we make our analysis more meaningful for modern steganographic schemes.

3.3.3 Evaluation Strategy

Although practical detectors may tolerate some errors, the recoverability of the order (Definition 3.5) is a necessary and critical condition for our detection strategy. Therefore, we need to quantify the degree of order recovery. For naïve adaptive embedding, the cover consists of solely two distinct areas of interest: the actual embedding positions and the unused positions. These areas are fixed by the payload size, which gives an empirical threshold τ in the value range of $\zeta(\cdot)$, until where the embedding can take place. For a fixed payload size, the metric of interest with regard to recovery is related to the transitions of pixels from *suitable* to *unsuitable* and vice versa, between the steganographer’s and the steganalyst’s knowledge about the values of $\zeta(\mathbf{x}^{(0)})$.

Definition 3.7 (Recovery Rate). *For a fixed adaptivity criterion $\zeta(\cdot)$, a fixed cover $\mathbf{x}^{(0)}$ of size n , and a fixed embedding rate p , let $\tau^{(0)} = \zeta(y_{p \cdot n}^{(0)})$ be the value of the adaptivity criterion of the most suitable among the pixels not used for naïve adaptive embedding. And $\tau^{(p)} = \hat{\zeta}(\hat{y}_{p \cdot n}^{(0)})$ is the corresponding threshold if the adaptivity criterion is estimated from the stego object. The recovery rate r is defined as*

$$r = \frac{1}{p \cdot n} \left| \left\{ i \mid \zeta(x_i^{(0)}) > \tau^{(0)} \wedge \hat{\zeta}(x_i^{(p)}) > \tau^{(p)} \right\} \right|. \quad (3.27)$$

Note that both “>” operators change to “<” for adaptivity criteria where lower values indicate better suitability.

In plain English, r measures the fraction of embedding positions correctly identified by the steganalyst. Furthermore, we can see the recovery rate as an empirical measure that shows us, what amount of the chosen embedding positions can be recovered.

Among the different ways to recover the order, we restrict our experiments to the simple method of applying $\zeta(\cdot)$ directly to the object under investigation. There might be better ways, e. g., by carefully studying and inverting the effect of the embedding operation on the adaptivity criterion. Since it is hard to tell how good the best possible estimator can be, we produce benchmark measurements with the best conceivable estimator, i. e., giving the steganalyst side information about the values of $\zeta(\cdot)$ applied to the cover, but not allowing her to simply compare $\zeta(x_i^{(1)}) \stackrel{?}{=} \zeta(x_i^{(0)})$ as detection strategy.

For relatively high payloads, the expected security gain from choosing the embedding positions adaptively is limited. Therefore, we hypothesize that our specialized WS method to detect naïve adaptive embedding performs best for low (and thus relevant) payload lengths. Note that the sequential WS method is pretty vulnerable to “gaps” in the payload stream [56]. Our specialized WS method inherits this problem and its severity is related to the (local) discreteness of adaptivity criteria, which makes it difficult to establish an order of embedding positions if the last bin is incompletely used.

3.3.4 Attacked Adaptivity Criteria

We decided to include the following four criteria in our study because they cover the space of proposed criteria pretty well. Each is from a different type, as seen in the first column of Table 3.1. The criteria differ in how they identify the more suitable embedding positions, i. e., the exact formulation of the adaptivity criterion $\zeta(\cdot)$. Unless otherwise mentioned, x_i is more suitable for embedding than x_j if it holds that: $\zeta(x_i) > \zeta(x_j)$.

3.3.4.1 Local Variance

Local variance is the most popular criterion, possibly because it can be tied to the detectability of embedding in a Gaussian cover, e. g., in [30, 85]. It can be calculated from the cover by,

$$\zeta(\mathbf{x}^{(0)}) = \left(\mathbf{x}^{(0)} - \mathbf{x}^{(0)} * \mathbf{a} \right)^2 * \mathbf{a}, \quad (3.28)$$

where \mathbf{a} is a 3×3 mean filter, $*$ is the 2-dimensional convolution operator, and the square operation is element-wise in Equation (3.28). As pixels with higher local variance are believed to be less predictable, content-adaptive embedding should concentrate embedding changes in positions with relatively high local variance. Note that local variance differs from the term σ_i^2 in the calculation of WS weights (see Section 3.2.2) in that the center pixel may be included in the variance estimation, whereas it must not be included for WS weights. Figure 3.2(a) shows an example of the embedding positions chosen by this adaptivity criterion, when embedding with embedding rate $p = 0.3$.

3.3.4.2 Edges

Here, we use the Laplacian edge detector to identify edges in an image and select them for embedding, as suggested, for instance, in [84]. This adaptivity criterion can be calculated by a linear filter:

$$\zeta(\mathbf{x}^{(0)}) = \mathbf{x}^{(0)} * \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}. \quad (3.29)$$

Figure 3.2(b) shows an example of the embedding positions chosen by this adaptivity criterion, when embedding with embedding rate $p = 0.3$.

3.3.4.3 Texture

In [30], the authors suggest to prefer more textured areas of a cover image to hide the steganographic payload. To identify textured areas, they propose four steps:

1. Divide the image into blocks of size 3×3 .
2. Divide each of these blocks into four 2×2 sub-blocks, with the center pixel being contained in every block.
3. Each sub-block is ‘good’, if there are at least three different grayscale levels.
4. The entire block is ‘good’, if all sub-blocks are ‘good’.

We extend their measure by defining $\zeta(x_i^{(0)})$ as the sum of the different grayscale levels in the four sub-blocks. Figure 3.2(c) shows an example of the embedding positions, when embedding with embedding rate $p = 0.3$.

3.3.4.4 NUGO (Not so Undetectable steGO)

Pevný et al.’s HUGO (Highly Undetectable steGO) algorithm [71] is inspired by machine learning-based steganalysis. The algorithm aims to minimize the distance between stego and cover image in a very high-dimensional feature space composed of quantized co-occurrence tables. This is done by measuring the distance in feature space for hypothetical embedding operations applied to each position in the cover independently, thereby obtaining an adaptivity criterion. The algorithm not only chooses the embedding positions adaptively, but also the embedding direction for an LSB matching operation. As the WS method is bound to LSB replacement, considering this additional level of content-adaptivity is beyond the scope of our analysis¹².

Therefore, we use a modified version of HUGO’s adaptivity criterion for LSB replacement, stripping the change direction component. Our modification, called ‘NUGO’, is most likely much inferior than the original criterion combined with the

¹²But we deem it possible that the embedding direction is reconstructible itself and thus might present another weakness of the real HUGO algorithm.

original embedding operation, but it adds some diversity compared to the other three criteria. Because the criterion measures potential distortion in the feature space, lower values of $\zeta(x_i^{(0)})$ indicate more suitable embedding positions. Figure 3.2(d) shows an example of the embedding positions chosen by this adaptivity criterion, when embedding with embedding rate $p = 0.3$.

Comparing Figures 3.2(a) and 3.2(d), we see a very similar selection of embedding positions. This highlights again Remark 3.13 that another function $\zeta'(\cdot) \neq \zeta(\cdot)$ might give a very accurate estimation of the embedding positions. Furthermore, a very complex adaptivity criterion like NUGO might be estimated with means of an adaptivity criterion with very low complexity like local variance.

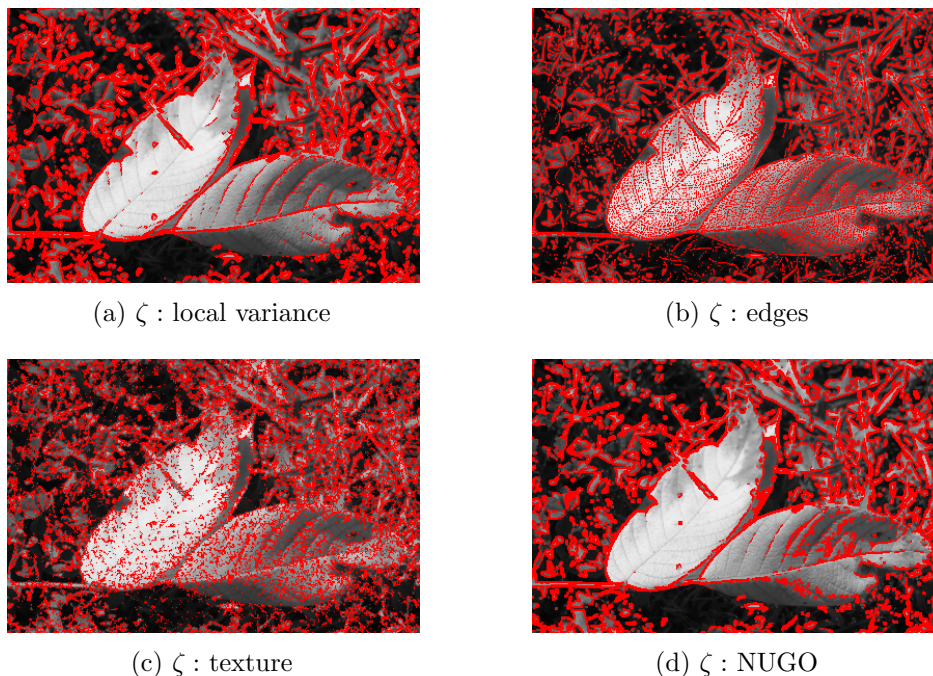



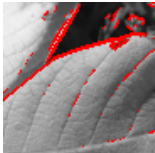
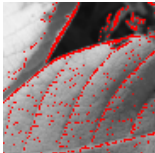
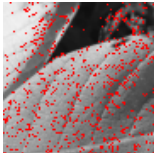
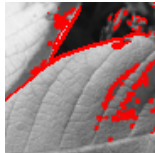
Figure 3.2: Examples of naïve adaptive embedding and payload $p = 0.3$

3.3.5 Recoverability of the Adaptivity Criteria

Table 3.2 shows the mean recovery rates for the four different adaptivity criteria introduced in Section 3.3.4. Observe the differences in the recoverability between the different adaptivity criteria. The most popular criterion, local variance, is almost perfectly recoverable at all embedding rates. We can see a significant decrease in recoverability for our only discretized adaptivity criterion, texture. This is expected, as even with perfect recoverability, a steganalyst cannot reconstruct the order in the last bin used. Interestingly, the least recoverable of the here examined adaptivity criteria is NUGO. Overall, the mean recovery rate seems to be roughly constant and independent

of the payload. For a fixed embedding operation, recoverability can be seen as a property of the adaptivity criterion. Different criteria can be ranked by recoverability.

Table 3.2: Recovery rate, calculated according to Definition 3.7.

Images	p	local variance	edges	textured	NUGO
					
BOSSBase	0.01	0.995	0.971	0.718	0.411
	0.05	0.992	0.956	0.732	0.318
	0.10	0.990	0.943	0.749	0.395
	0.20	0.985	0.925	0.787	0.415
DIB	0.01	0.991	0.934	0.690	0.252
	0.05	0.987	0.917	0.697	0.435
	0.10	0.982	0.906	0.714	0.378
	0.20	0.973	0.898	0.756	0.531

3.3.6 Empirical Results – Detecting Naïve Adaptive Embedding

Tables 3.3 and 3.4 summarize the results of our tests. It shows that our specialized WS with estimation from the stego object outperforms both standard WS methods for very small payloads, independent of the adaptivity criterion among the ones tested. It is on par for higher payloads, here with some variation between adaptivity criteria. The third column shows the specialized WS method with perfect recovery of the adaptivity criterion. This is not achievable in practice, but serves as a benchmark.

Figures 3.3 to 3.7 (on pages 48 to 52) display detection performance measured by the mean absolute error (MAE), $|p - \hat{p}|$, as a function of the embedding rate p . Lower values indicate better performance. Figure 3.3 compares the performance of the weighted and unweighted WS for the embedding methods random uniform embedding and naïve adaptive embedding, with adaptivity criteria local variance and NUGO. Surprisingly, both methods are similarly accurate at detecting NUGO as they are at detecting random uniform embedding. For adaptive embedding into the areas with higher local variance, the unweighted WS clearly outperforms the weighted version. This confirms that the local weights, calculated as suggested in [56], are counter-productive in this case of adaptive embedding.

Figures 3.4 to 3.7 show the performance of the different WS methods, including our specialized one, for naïve adaptive embedding using the four selected adaptivity criteria, for both image databases tested. In Figure 3.4 the adaptivity criterion is local variance, in Figure 3.5 it is the edge criterion, in Figure 3.6 the more textured areas are preferred and finally, Figure 3.7 shows the detection accuracy for the NUGO criterion.

Here and from the Tables 3.2, 3.3 and 3.4 , it can be seen that, as the recovery rate declines, the accuracy of the specialized WS with estimation from the stego object also decreases.

Furthermore, Tables 3.3 and 3.4 show the superiority of adaptivity criteria which are less recoverable. The recovery rates (Table 3.2) for NUGO are much lower than those for the other adaptivity criteria and so is the performance of our specialized method (fourth column). Observe that for all adaptivity criteria, our specialized WS still outperforms the known WS methods for small payloads.

Another conclusion drawn from Tables 3.3 and 3.4 is that the estimation error of our specialized WS increases with increasing payload, even if the recoverability rate stays roughly constant. This is a consequence of the absolute length of the “gaps” introduced by the estimation error, which increases for higher payloads. This confirms our conjecture from Section 3.3.3 that our specialized version has best performance for very small payloads.

3.4 Summary

In this chapter, we were mainly concerned with the side of steganalysis. From this perspective, we

- ▶ introduced the formal concepts of steganographic side information and uncertainty, and
- ▶ established a connection between the definitions of these two terms and information-theoretic security.

Our goal is to remove the inconsistent usage of these terms with these definitions. With the help of the definitions, we formalized content-adaptive embedding, and

- ▶ proposed that the problem of detecting naïve adaptive embedding can be reduced to the problem of detecting initial sequential embedding, and
- ▶ presented an extension of WS steganalysis specifically tailored to detect naïve adaptive embedding.

Finally, we developed a targeted attack on four widely used adaptivity criteria. This attack detects the embedded message length very accurately, depending on the recovery rate of the respective adaptivity criterion. So, we confirmed that any rational steganalyst would gain from recalculating or estimating the most likely embedding positions.

Taking this result into account, we believe that no rational steganographer would use naïve adaptive embedding. To frame the contest of an anticipating steganographer and a counter-anticipating steganalyst, we turn to the area of game theory and give a formal game-theoretical framework modeling this situation.

Table 3.3: Summary of detection results for 10 000 images of the BOSSBase

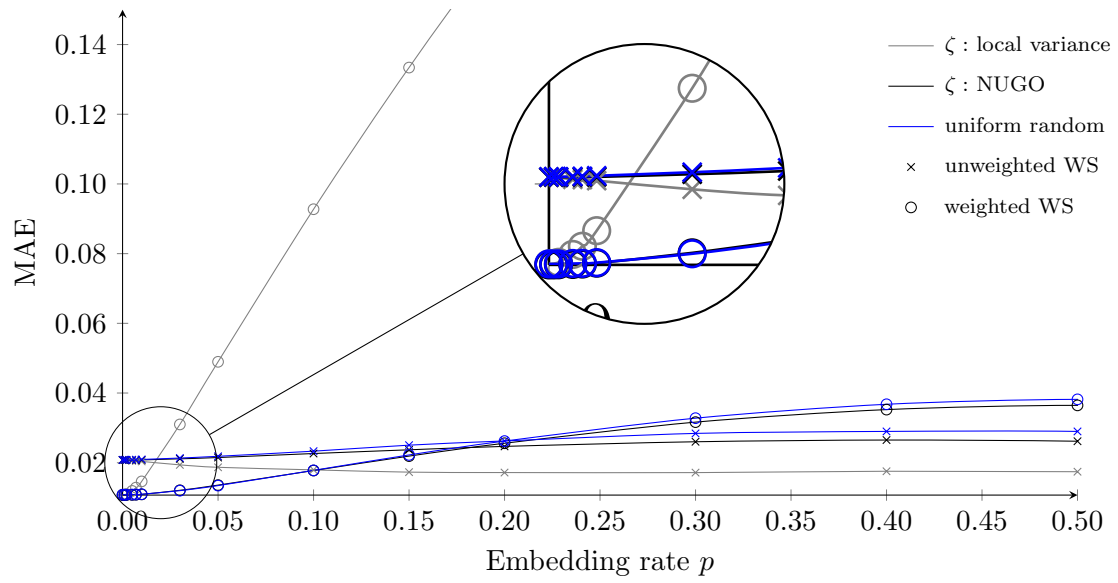
Embedding rate p	Adaptivity criterion	Detector														
		Specialized WS (proposed)						Standard WS (benchmark)								
		perfect recovery		estimated from stego		unweighted		weighted		unweighted		weighted				
	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	
0.01	Local variance	0.008,	(0.005),	0.20	0.008,	(0.005),	0.20	0.020,	(0.023),	0.31	0.015,	(0.013),	0.495	0.015,	(0.013),	0.495
	Edge	0.005,	(0.005),	0.15	0.004,	(0.002),	0.17	0.019,	(0.023),	0.32	0.014,	(0.013),	0.47	0.014,	(0.013),	0.47
	Texture	0.004,	(0.002),	0.07	0.006,	(0.002),	0.10	0.021,	(0.024),	0.31	0.013,	(0.013),	0.40	0.013,	(0.013),	0.40
	NUGO	0.002,	(0.001),	0.02	0.007,	(0.008),	0.08	0.021,	(0.024),	0.29	0.011,	(0.013),	0.14	0.011,	(0.013),	0.14
0.05	Local variance	0.007,	(0.002),	0.04	0.007,	(0.003),	0.04	0.019,	(0.022),	0.07	0.049,	(0.013),	0.41	0.049,	(0.013),	0.41
	Edge	0.005,	(0.001),	0.01	0.011,	(0.007),	0.02	0.018,	(0.023),	0.07	0.042,	(0.018),	0.20	0.042,	(0.018),	0.20
	Texture	0.005,	(0.002),	0.01	0.016,	(0.006),	0.01	0.020,	(0.024),	0.07	0.040,	(0.015),	0.15	0.040,	(0.015),	0.15
	NUGO	0.002,	(0.001),	0.00	0.038,	(0.020),	0.02	0.021,	(0.025),	0.06	0.014,	(0.016),	0.01	0.014,	(0.016),	0.01
0.1	Local variance	0.005,	(0.001),	0.01	0.006,	(0.003),	0.01	0.018,	(0.022),	0.02	0.093,	(0.017),	0.25	0.093,	(0.017),	0.25
	Edge	0.003,	(0.001),	0.00	0.024,	(0.018),	0.01	0.018,	(0.024),	0.02	0.072,	(0.033),	0.00	0.072,	(0.033),	0.00
	Texture	0.005,	(0.002),	0.00	0.029,	(0.010),	0.01	0.021,	(0.025),	0.02	0.076,	(0.020),	0.05	0.076,	(0.020),	0.05
	NUGO	0.003,	(0.001),	0.00	0.068,	(0.046),	0.02	0.023,	(0.027),	0.06	0.018,	(0.022),	0.01	0.018,	(0.022),	0.01
0.2	Local variance	0.004,	(0.001),	0.00	0.005,	(0.004),	0.00	0.017,	(0.022),	0.01	0.170,	(0.033),	0.05	0.170,	(0.033),	0.05
	Edge	0.002,	(0.000),	0.00	0.056,	(0.047),	0.00	0.021,	(0.028),	0.01	0.108,	(0.068),	0.00	0.108,	(0.068),	0.00
	Texture	0.005,	(0.001),	0.00	0.049,	(0.019),	0.00	0.021,	(0.026),	0.01	0.142,	(0.039),	0.01	0.142,	(0.039),	0.01
	NUGO	0.003,	(0.001),	0.00	0.133,	(0.079),	0.00	0.025,	(0.032),	0.00	0.026,	(0.032),	0.00	0.026,	(0.032),	0.00

Performance is measured by the mean absolute error (MAE), the inter-quartile range (IQR) of the estimation error, and the false positive rate at 50% detection (FP₅₀). Lower values indicate better performance for all three metrics.

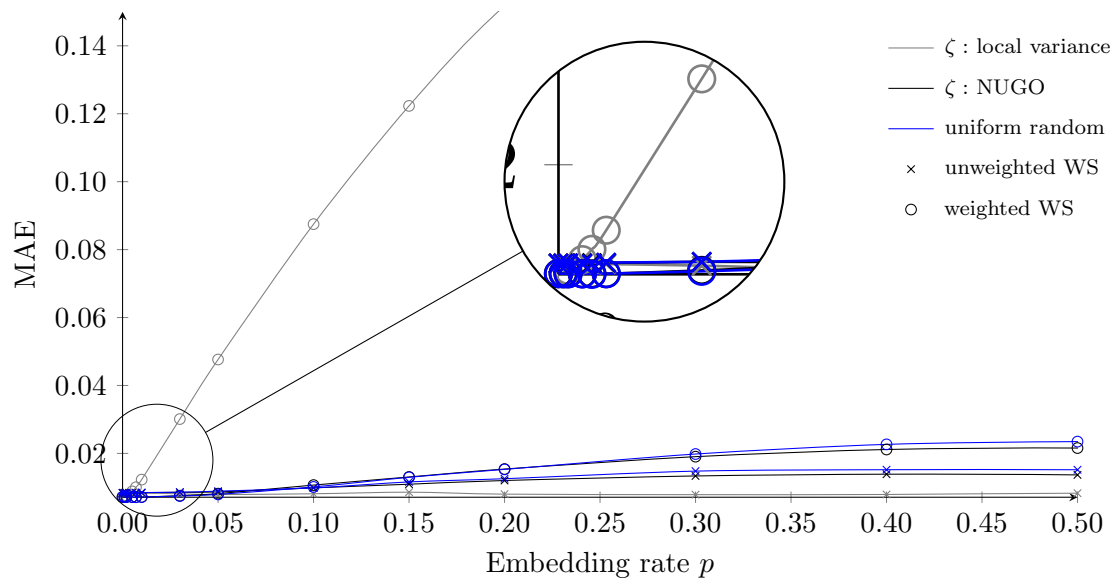
Table 3.4: Summary of detection results for 700 images of the Dresden Image Base

Embedding rate p	Adaptivity criterion	Detector											
		Specialized WS (proposed)						Standard WS (benchmark)					
		perfect recovery			estimated from stego			unweighted			weighted		
		MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀	MAE	IQR	FP ₅₀
0.01	Local variance	0.002, (0.030), 0.001	0.002, (0.030), 0.001	0.002, (0.030), 0.001	0.002, (0.030), 0.001	0.008, (0.140), 0.011	0.008, (0.140), 0.011	0.012, (0.480), 0.009	0.011, (0.420), 0.010	0.011, (0.360), 0.009	0.007, (0.060), 0.009	0.048, (0.320), 0.011	0.036, (0.030), 0.021
	Edge	0.002, (0.030), 0.001	0.003, (0.040), 0.002	0.003, (0.040), 0.002	0.003, (0.040), 0.002	0.008, (0.140), 0.010	0.008, (0.140), 0.010	0.011, (0.420), 0.010	0.011, (0.360), 0.009	0.007, (0.060), 0.009	0.048, (0.320), 0.011	0.036, (0.030), 0.021	
	Texture	0.001, (0.010), 0.000	0.004, (0.020), 0.001	0.004, (0.020), 0.001	0.004, (0.020), 0.001	0.008, (0.170), 0.010	0.008, (0.170), 0.010	0.011, (0.360), 0.009	0.011, (0.360), 0.009	0.007, (0.060), 0.009	0.048, (0.320), 0.011	0.036, (0.030), 0.021	0.036, (0.030), 0.021
	NUGO	0.000, (0.000), 0.000	0.009, (0.110), 0.002	0.009, (0.110), 0.002	0.009, (0.110), 0.002	0.008, (0.103), 0.010	0.008, (0.103), 0.010	0.007, (0.060), 0.009	0.007, (0.060), 0.009	0.007, (0.060), 0.009	0.048, (0.320), 0.011	0.036, (0.030), 0.021	0.036, (0.030), 0.021
0.05	Local variance	0.001, (0.000), 0.000	0.002, (0.000), 0.001	0.002, (0.000), 0.001	0.002, (0.000), 0.001	0.008, (0.010), 0.010	0.008, (0.010), 0.010	0.048, (0.320), 0.011	0.036, (0.030), 0.021	0.036, (0.030), 0.021	0.048, (0.320), 0.011	0.036, (0.030), 0.021	0.036, (0.030), 0.021
	Edge	0.001, (0.000), 0.000	0.017, (0.000), 0.013	0.017, (0.000), 0.013	0.017, (0.000), 0.013	0.010, (0.010), 0.011	0.010, (0.010), 0.011	0.040, (0.060), 0.012	0.040, (0.060), 0.012	0.040, (0.060), 0.012	0.008, (0.000), 0.009	0.008, (0.000), 0.009	0.008, (0.000), 0.009
	Texture	0.001, (0.000), 0.000	0.016, (0.000), 0.004	0.016, (0.000), 0.004	0.016, (0.000), 0.004	0.009, (0.010), 0.010	0.009, (0.010), 0.010	0.008, (0.000), 0.009	0.008, (0.000), 0.009	0.008, (0.000), 0.009	0.048, (0.320), 0.011	0.036, (0.030), 0.021	0.036, (0.030), 0.021
	NUGO	0.001, (0.000), 0.000	0.033, (0.000), 0.029	0.033, (0.000), 0.029	0.033, (0.000), 0.029	0.008, (0.000), 0.002	0.008, (0.000), 0.002	0.087, (0.080), 0.018	0.056, (0.000), 0.043	0.073, (0.010), 0.021	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010
0.1	Local variance	0.002, (0.000), 0.000	0.002, (0.000), 0.002	0.002, (0.000), 0.002	0.002, (0.000), 0.002	0.008, (0.000), 0.010	0.008, (0.000), 0.010	0.087, (0.080), 0.018	0.056, (0.000), 0.043	0.073, (0.010), 0.021	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010
	Edge	0.001, (0.000), 0.000	0.036, (0.000), 0.028	0.036, (0.000), 0.028	0.036, (0.000), 0.028	0.013, (0.000), 0.013	0.013, (0.000), 0.013	0.073, (0.010), 0.021	0.073, (0.010), 0.021	0.073, (0.010), 0.021	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010
	Texture	0.001, (0.000), 0.000	0.031, (0.000), 0.007	0.031, (0.000), 0.007	0.031, (0.000), 0.007	0.010, (0.000), 0.010	0.010, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010	0.011, (0.000), 0.010
	NUGO	0.001, (0.000), 0.000	0.078, (0.000), 0.036	0.078, (0.000), 0.036	0.078, (0.000), 0.036	0.008, (0.000), 0.004	0.008, (0.000), 0.004	0.152, (0.010), 0.052	0.075, (0.000), 0.063	0.127, (0.000), 0.045	0.015, (0.000), 0.012	0.015, (0.000), 0.012	0.015, (0.000), 0.012
0.2	Local variance	0.002, (0.000), 0.000	0.004, (0.000), 0.004	0.004, (0.000), 0.004	0.004, (0.000), 0.004	0.008, (0.000), 0.011	0.008, (0.000), 0.011	0.152, (0.010), 0.052	0.075, (0.000), 0.063	0.127, (0.000), 0.045	0.015, (0.000), 0.012	0.015, (0.000), 0.012	0.015, (0.000), 0.012
	Edge	0.002, (0.000), 0.000	0.077, (0.000), 0.050	0.077, (0.000), 0.050	0.077, (0.000), 0.050	0.010, (0.000), 0.013	0.010, (0.000), 0.013	0.075, (0.000), 0.063	0.075, (0.000), 0.063	0.127, (0.000), 0.045	0.015, (0.000), 0.012	0.015, (0.000), 0.012	0.015, (0.000), 0.012
	Texture	0.002, (0.000), 0.000	0.053, (0.000), 0.012	0.053, (0.000), 0.012	0.053, (0.000), 0.012	0.017, (0.000), 0.012	0.017, (0.000), 0.012	0.075, (0.000), 0.063	0.075, (0.000), 0.063	0.127, (0.000), 0.045	0.015, (0.000), 0.012	0.015, (0.000), 0.012	0.015, (0.000), 0.012
	NUGO	0.002, (0.000), 0.000	0.097, (0.000), 0.153	0.097, (0.000), 0.153	0.097, (0.000), 0.153	0.012, (0.000), 0.012	0.012, (0.000), 0.012	0.075, (0.000), 0.063	0.075, (0.000), 0.063	0.127, (0.000), 0.045	0.015, (0.000), 0.012	0.015, (0.000), 0.012	0.015, (0.000), 0.012

Performance is measured by the mean absolute error (MAE), the inter-quartile range (IQR) of the estimation error, and the false positive rate at 50% detection (FP₅₀). Lower values indicate better performance for all three metrics.

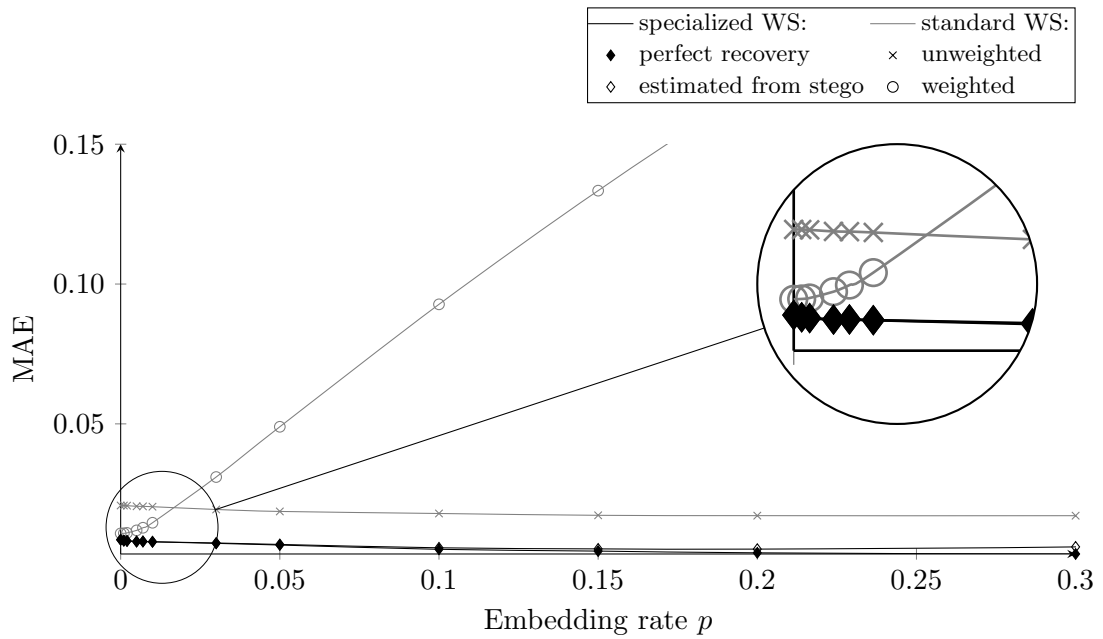


(a) BOSSBase

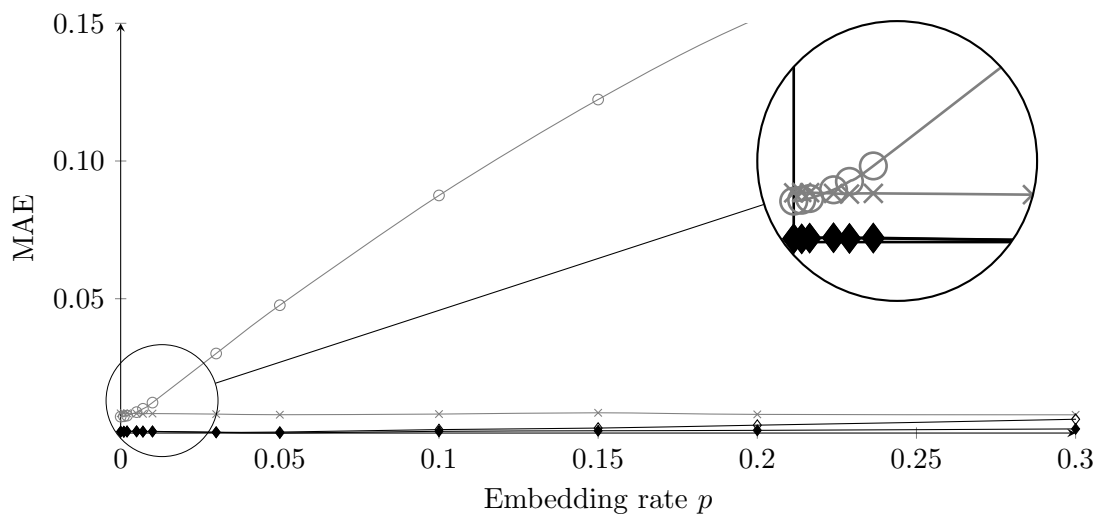


(b) Dresden Image Database

Figure 3.3: Mean absolute error (MAE) of the standard WS variants as function of the embedding rate p for different embedding schemes.

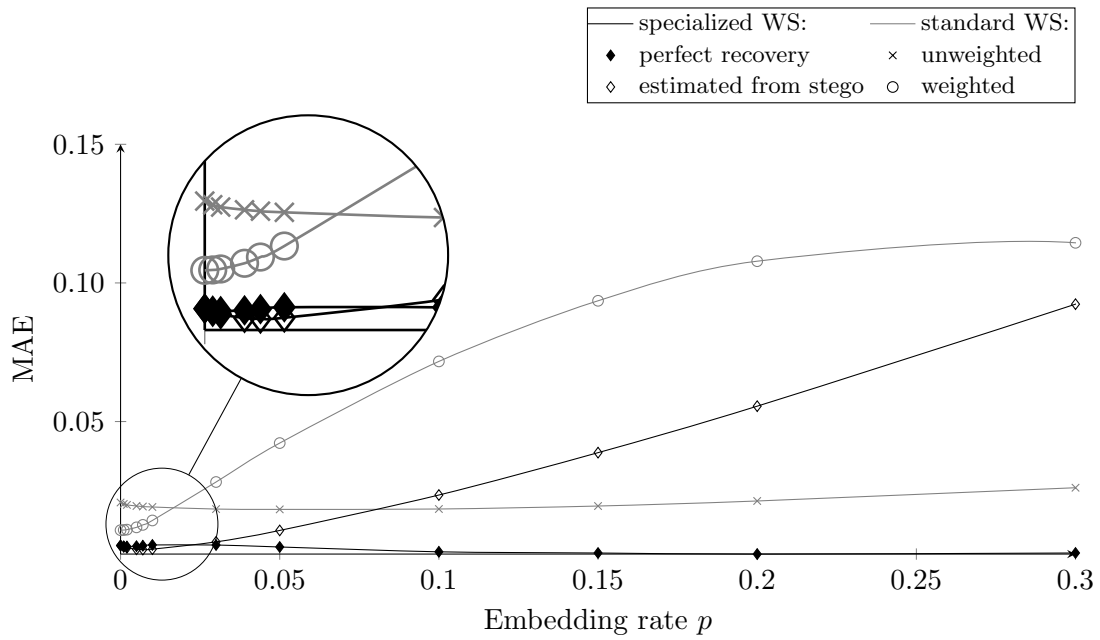


(a) BOSSBase

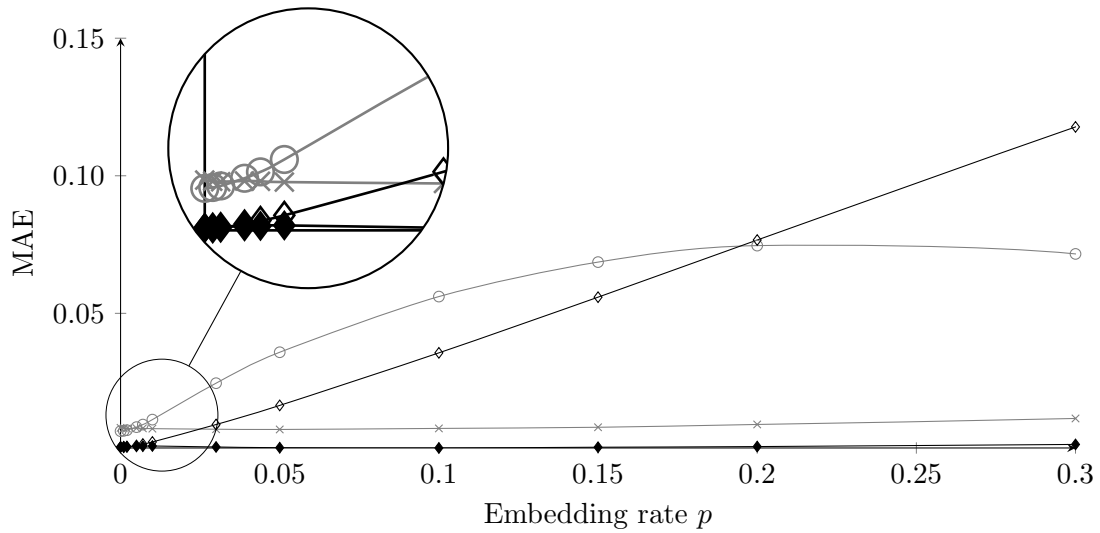


(b) Dresden Image Database

Figure 3.4: Adaptivity criterion: local variance. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p .

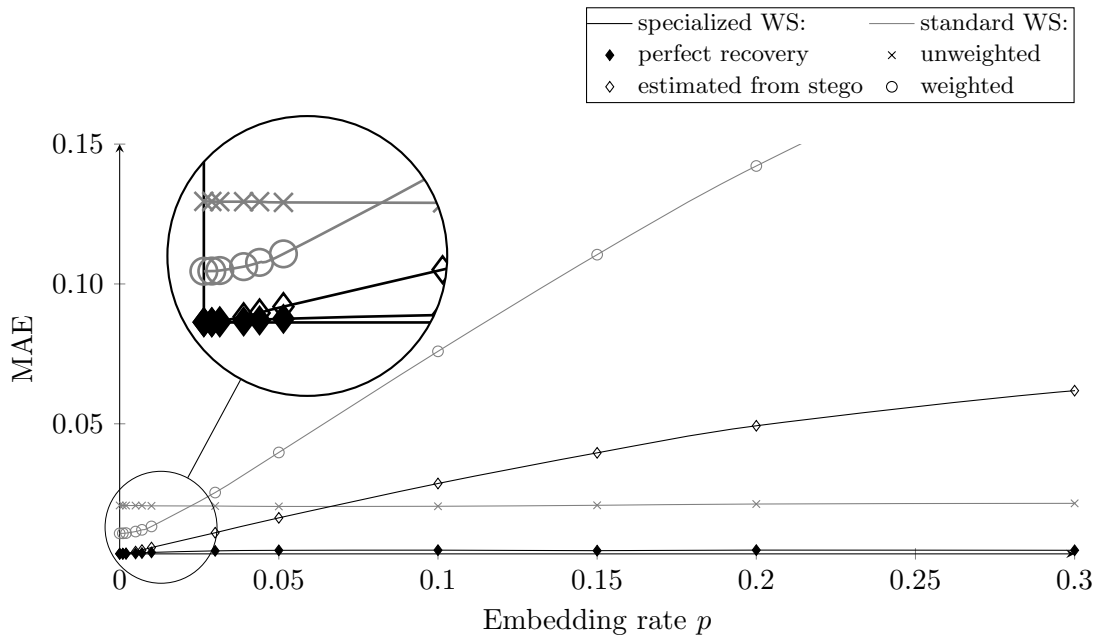


(a) BOSSBase

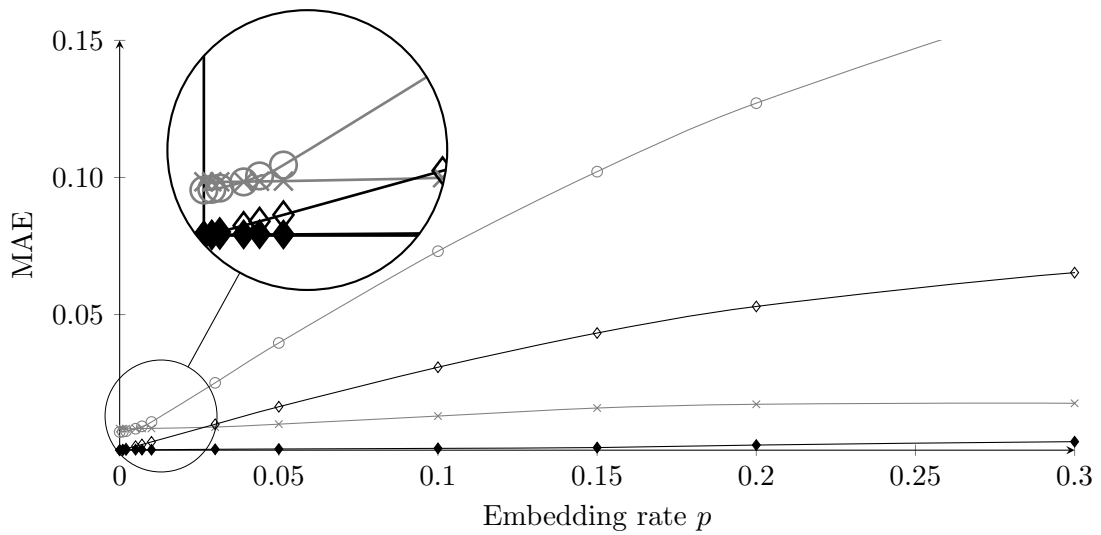


(b) Dresden Image Database

Figure 3.5: Adaptivity criterion: edges. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p .

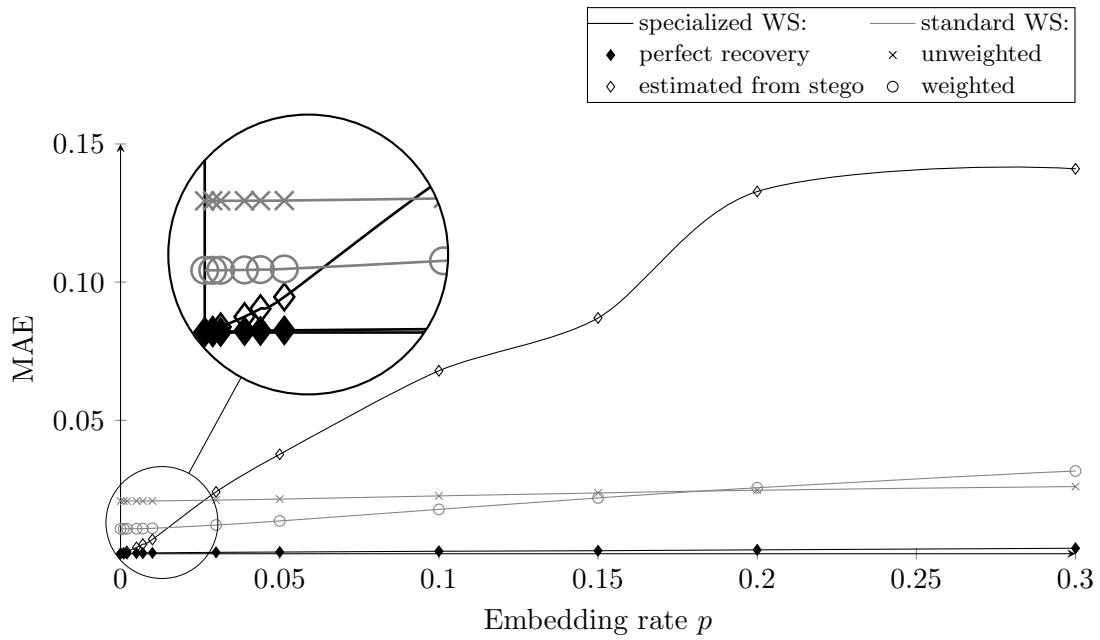


(a) BOSSBase

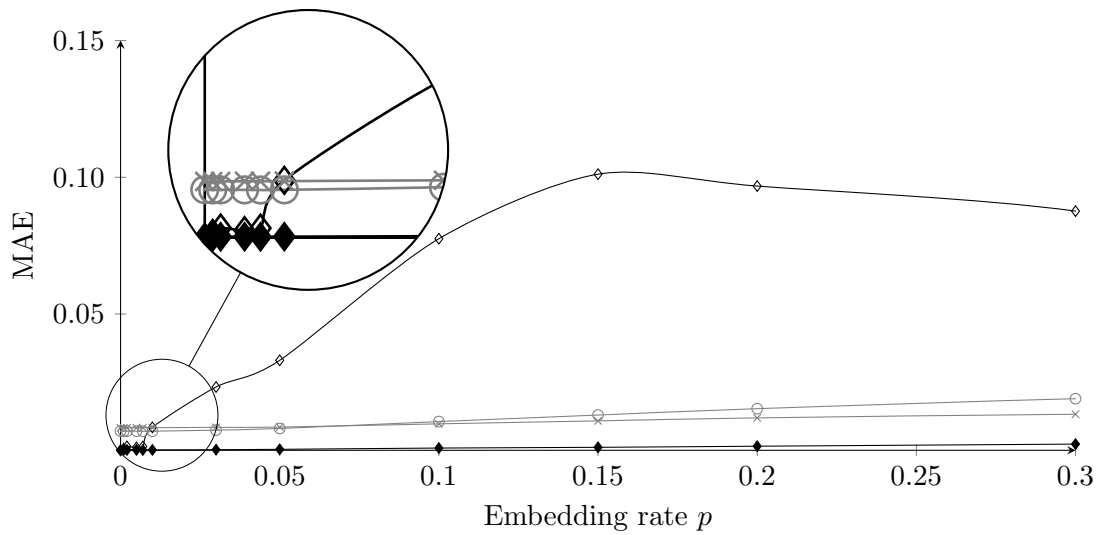


(b) Dresden Image Database

Figure 3.6: Adaptivity criterion: texture. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p .



(a) BOSSBase



(b) Dresden Image Database

Figure 3.7: Adaptivity criterion: NUGO. Mean absolute error (MAE) of specialized and standard WS methods as function of the embedding rate p .

Chapter 4

Game Theory and Steganography

Considering the empirical evidence that a steganalyst can utilize steganographic side information used in content-adaptive embedding to increase her detection performance, it seems plausible that any rational steganalyst will do so. Anticipating this, any rational steganographer would adapt her embedding strategy to this assumption. Game theory is the study of strategic choices and their consequences for two or more rational players with conflicting goals [87]. In this chapter we are going to argue why steganography utilizing steganographic side information must be studied with game theory. Then, we will introduce the basic concepts of game theory necessary to lay the foundation for the rest of the thesis. After briefly reviewing existing approaches to model the contest between steganographer and steganalyst with game theory, we present our framework that captures all relevant properties of a steganographic communication system with side information.

4.1 Motivation

In Section 3.1 we introduced the term steganographic side information (SSI) and motivated its use in steganography to gain an advantage in knowledge over the steganalyst. Most often, the use of SSI to select more uncertain positions for embedding relies on the authors' judgment or heuristics inspired by known steganalysis methods (cf. Table 3.1), as in the examples of the adaptivity criteria presented in Section 3.3. When reporting security gains over random uniform embedding, the authors often disobey Kerckhoffs' principle by not considering that the steganalyst knows how the SSI is used in the embedding function and might be able to reconstruct the SSI from the stego object. As a result, the security of many side-informed embedding schemes against a so-informed attacker remains an open research question.

Before modeling the situation faced, we have to identify the different parties involved within a steganographic system and make a realistic assumption about the different levels of information each of them has.

It is reasonable to assume that the steganographer does not know the global cover distribution \mathcal{P}_0 , because with that knowledge she could perform perfect steganography [88]. Granting the steganalyst access to both global distributions \mathcal{P}_0 and \mathcal{P}_1 , as suggested by the strictest interpretation of Kerckhoffs' principle for steganography [20], would enable her to attack with the best-possible detector. This is unrealistic for practical settings and studied sufficiently. Instead, we follow Böhme and Ker, who argue that a realistic set-up is characterized by incomplete information and bounded computational resources for all actors [8, 53, 54]. This means that both actors, unaware of the global

distributions, must resort to local models based on public knowledge.

The steganographer chooses along which dimensions the cover should be moved to the message region (as depicted in Figure 2.2 on page 11), possibly based on some kind of SSI. The steganalyst chooses element weights to aggregate local evidence into a global decision, being aware that the steganographer might have used SSI to identify her embedding positions. Both choices are clearly interdependent and jointly affect the security of the steganographic communication. Therefore, both choices have to be strategic, i. e., anticipating the opponent's choice. This suggests that adaptive steganography and optimal adaptive steganalysis is best studied in the context of game theory, which is mathematically well-established to model situations of two (or more) parties who act strategically [87].

Game theory is the adequate option to evaluate situations with imperfect knowledge and strategic behavior of the players. With perfect knowledge on both sides, the problem would reduce to an optimization problem. Furthermore, game theory finds stable situations, so-called equilibria, where no rational player would deviate from her strategy.

4.2 Principles of Game Theory

First, we give the basic notations used in game theoretic set-ups and then list the different solution concepts used in the following sections. This overview is by no means exhaustive but provides all the necessary components for the following analyses. We base most of the definitions on the textbook [64] by Leyton-Brown and Shoham, but refer to the original sources where we deem it helpful.

4.2.1 Basic Definitions of Game Theory

A game consists of at least two rational players $v_i, i \in \{1, \dots, k\}$, their respective set of strategies s_i and a payoff (or utility) function u_i for each player and each possible outcome of the game [87]. It is a common and handy convention to denote the strategy of all players except player i with s_{-i} .

Our steganographic setting entails two players, the steganographer and the steganalyst. We thus restrict all definitions to the case of two-player games. For a more common definition, see the original textbook [pp. 3][64]. We keep the s_{-i} notation even in the two-player case, for consistency with the standard game theory jargon.

Definition 4.1 (Two-Player Game). *A two-player game is a tuple (\mathbb{S}, u) where*

- ▶ $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2$, where \mathbb{S}_i is a finite set of strategies available to player $i \in \{1, 2\}$. Each vector $\mathbf{s} = (s_1, s_2) \in \mathbb{S}$ is called a strategy profile;
- ▶ $\mathbf{u} = (u_1, u_2)$, where $u_i : \mathbb{S} \rightarrow \mathbb{R}$ is a real-valued utility (or payoff) function for player i .

A strategy profile in any game can be expressed as $\mathbf{s} = (s_i, s_{-i})$.

The most common representation of games is the one of a *normal form*, or *matrix game*. Here, every player's utility for every state of the world is represented in the special case, where the states of the world depend only on the player's combined actions.

Definition 4.2 (Zero-Sum Game). *A two-player game is called zero-sum game, if for each strategy profile $\mathbf{s} \in \mathbb{S}$ it holds that $u_1(\mathbf{s}) + u_2(\mathbf{s}) = 0$.*

Zero-sum games are strictly competitive, as one player's gain always is the other player's loss.

When identifying the players' strategies, game theory allows to study randomized strategies, so-called *mixed strategies*, in comparison to deterministic *pure strategies*.

Definition 4.3 (Pure Strategy). *A pure strategy a_i is a deterministic choice of player i in every possible state of the world.*

Definition 4.4 (Mixed Strategy). *A mixed strategy s_i for player i is a probability distribution over her pure strategies.*

Definition 4.5 (Support). *The support of a mixed strategy s_i for a player i is the set of pure strategies $\{a_i | s_i(a_i) > 0\}$.*

A pure strategy is a special case of a mixed strategy where the support consists of only one strategy. A *fully mixed strategy* is a strategy that assigns positive probability on every pure strategy, i.e., a strategy that has full support.

Game theory tries to identify strategies that are optimal in the sense that they guarantee each player a certain payoff. As all players are assumed to act rationally, all of them would use strategies that maximize their payoff or minimize their loss. This leads to the notion of *best response strategies*.

Definition 4.6 (Best Response Strategy). *Player i 's best response s_i^* to the strategy s_{-i} is a mixed strategy such that $u_i(s_i^*, s_{-i}) \geq u_i(s_i, s_{-i})$ for all strategies $s_i \in \mathbb{S}_i$.*

Generally, best response strategies are not unique and most often there are infinitely many such strategies. Furthermore, a given best response strategy assumes knowledge about the exact strategy of the opponent.

At first glance it might be difficult to identify single optimal strategies, but it might be easier to identify strategies the players would never use.

Definition 4.7 (Dominant and Dominated Strategy). *A strategy s_i dominates another strategy s'_i if for all $s_{-i} \in S_{-i}$ it holds that $u_i(s_i, s_{-i}) > u_i(s'_i, s_{-i})$. This implies that a strategy s'_i is dominated if some other strategy s_i dominates s'_i .*

Game theory differentiates not only several types of strategies but also assumptions about the knowledge of the players. The basic model, where each player has full information about the possible strategies of her opponent and her respective payoffs is called a situation of *complete information*. There are two ways to loosen this assumption.

Definition 4.8 (Incomplete Information). *When some, or all, players are uncertain about the payoffs, the available strategies or the amount of information of their opponents, the game is said to have incomplete information¹³. The players' uncertainties are represented as a probability distribution over all possibilities.*

Definition 4.9 (Imperfect Information). *In an imperfect information game, each player's choice situations are partitioned into so-called information sets. All situations within an information set look the same to the player and she cannot distinguish between them.*

There exists a way to transform any game with incomplete information into a game with imperfect information by introducing a probabilistic player called *Nature* [39]. Nature does not act strategically in the way that it has no payoff which it wants to maximize. The actions of Nature are known to all players but the instantiations are not predictable.

4.2.2 Solution Concepts

To identify optimal strategy profiles that each player will follow, we have to find stable situations, such that no player has incentives to deviate from her strategy. Such stable situations are called *equilibria* in game theory.

4.2.2.1 Dominant Strategy Equilibrium

If we have single dominant strategies for both players in a zero-sum game, we can assume that both would follow them, leading to the notion of a *dominant strategy equilibrium* (DSE).

Definition 4.10 (Dominant Strategy Equilibrium). *A strategy profile $\mathbf{s} = (s_1, s_2)$ is called a dominant strategy equilibrium in a two-player game, if for both players s_i is a dominant strategy for player i .*

Unfortunately, these equilibria seldomly exist in practice.

4.2.2.2 Nash Equilibrium

The most popular solution concept in game theory is the Nash equilibrium [67] and makes use of the best response strategies.

Definition 4.11 (Nash Equilibrium). *A strategy profile $\mathbf{s}^* = (s_1^*, s_2^*)$ is called a Nash equilibrium in a two-player game, if for both players s_i^* is a best response to s_{-i}^* .*

This definition implies that no player would unilaterally change her strategy, as she could do no better by doing so, assuming that the other player plays her equilibrium strategy.

In comparison to DSE we can be sure that such an equilibrium exists, although it might be hard to compute its exact strategies.

¹³Sometimes these game are called *Bayesian games*.

Theorem 4.1 (Nash [67]). *Every game with a finite number of players and strategy profiles has at least one Nash equilibrium.*

4.2.2.3 Maxmin and Minmax Strategy

Another solution concept that is especially well-motivated in two-player zero-sum games are the *maxmin* and *minmax strategies*. The intuition behind these strategies is that a player following her maxmin strategy wants to maximize her worst case payoff, i.e., in a situation where the other player wants to minimize it. The minmax strategy is the dual in that it minimizes the maximum payoff of the other player.

Definition 4.12 (Maxmin and Minmax Strategy). *The maxmin strategy of player i is $\operatorname{argmax}_{s_i} \min_{s_{-i}} u_i(s_i, s_{-i})$ and the minmax strategy is $\operatorname{argmin}_{s_i} \max_{s_{-i}} u_i(s_i, s_{-i})$.*

The minimum amount of payoff that is guaranteed to player i by playing her maxmin strategy is called *maxmin value* and the minimum of the maximum value that player i can ensure by playing a minmax strategy is the *minmax value*.

In our case of two-player zero-sum games, one of the fundamental proofs in game theory states:

Theorem 4.2 (von Neumann [86]). *In any two-player, zero-sum game with a finite number of strategies, in any Nash equilibrium each player receives a payoff that is equal to both her maxmin and her minmax value.*

This theorem has three implications for two-player, zero-sum games:

1. Both players' maxmin values equal their minmax value.
2. For both players, the maxmin strategies coincide with the minmax strategies.
3. Any maxmin strategy profile is a Nash equilibrium and the payoff in all Nash equilibria is the same.

4.2.2.4 Equalizer Strategies

A more recent solution concept are the so-called *equalizer strategies*. These are strategies that yield the same expected payoff for each player, regardless of the (pure or mixed) strategy chosen by the other player [74].

Definition 4.13 (Equalizer Strategies). *A strategy s_i is called an equalizer strategy for player i , if, for some $v \in \mathbb{R}$ it holds that $u_i(s_i, s_{-i}) = v$ for any $s_{-i} \in S_{-i}$.*

This definition implies that playing an equalizer strategy makes the opponent indifferent about which strategy to choose, as she cannot influence the payoff when her opponent plays an equalizer strategy.

A necessary and sufficient condition for the existence of equalizer strategies is that no pure or mixed strategy of any player is dominated by a convex combination of her other strategies. Thus, the existence of equalizer strategies and dominant strategies are mutually exclusive.

Theorem 4.3 (Pruzhansky [74]). *Let $\mathbf{s}^* = (s_1^*, s_2^*)$ be a Nash equilibrium in completely mixed strategies. If there exists an equalizer strategy for player i , then such an equalizer strategy guarantees the equilibrium payoff for player i against any strategy of the opponent.*

An equilibrium in equalizer strategies plays a special role. In such an equilibrium, the payoff is mutually independent of the opponent's choice.

4.3 Game-Theoretical Approaches in Steganography

Game theory gains more and more importance in practically all areas concerned with security. Examples are real-world security like the patrols at airports [72], the modeling of phishing strategies [12], network defense [66], and team building in the face of a possible insider threat [63].

We restrict our presentation to the area of steganography here, but list three more examples from areas connected to steganography in Appendix B.

We are aware of three other independent publications using game theory in the context of steganography that precede our initial publication, none of which uses side information, and one recent publication building on our set-up.

4.3.1 Game Theory and Capacity

Back in 1998, Ettinger [17] proposed a two-player, zero-sum game between a steganographer and an active steganalyst whose purpose it is to interrupt the steganographic communication. Both players are subject to a distortion constraint. The steganographer chooses a distribution of locations to hide her message, which is assumed to resemble pseudorandom noise. The steganalyst also chooses a distribution over positions she can overwrite automatically in every sequence. The distortion constraint d is the same for both players. The payoff measures the amount of data that is communicated, so the steganographer wants to maximize the payoff and the steganalyst wants to minimize it. Ettinger uses the most simple distortion measure: changing the LSB of a position introduces 1 unit of distortion, the next-to-LSB 2 units and so forth. Here, the steganographer has an advantage over the steganalyst, as in half of the cases, the bit she wants to change already has the right semantic, so she only needs to change $q/2$ bits to communicate q bits, while the steganalyst always flips a bit, should she choose to distort that specific position. The payoff function is constructed using coding theory and assuming that every bit is an own channel C_I with a specific channel capacity. The payoff function is:

$$P(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{l-1} x_i \left(1 - H\left(\frac{y_i}{n}\right)\right), \quad (4.1)$$

where $\mathbf{x} = (x_0, \dots, x_{l-1})$ and $\mathbf{y} = (y_0, \dots, y_{l-1})$ are the steganographer's, respectively steganalyst's strategy and H is the binary entropy. Ettinger derives analytically that for the equilibrium strategy for the steganalyst it must hold that for $0 < j < k < l - 1$:

$$2^{k-j} \left(1 - H\left(\frac{y_j}{n}\right)\right) = \left(\frac{y_k}{n}\right), \quad (4.2)$$

and for the steganographer:

$$\frac{x_j}{x_k} = \frac{2^{j-k} \log \frac{p_k^*}{1-p_k^*}}{\log \frac{p_k^*}{1-p_k^*}}, \quad (4.3)$$

where $p_k^* = y_i^*/n$. The steganalyst's equilibrium strategy is to equalize as many of the lowest order channel capacities as allowed by the distortion constraint.

This set-up differs from the conventional steganographic model as the protection goal is availability, not undetectability.

4.3.2 Game Theory and Batch Steganography

Ker [50] uses game theory to find strategies in the special case of batch steganography, where the payload can be spread over many cover objects. The steganalyst anticipates this and tries to detect the existence of any secret message (so-called pooled steganalysis). For this Ker defines the so-called *Threshold Game* in which the warden chooses a quantitative steganalysis method and a threshold t and then counts how many of the objects have an estimate exceeding t . The steganographer on the other hand chooses her strategy for distributing the secret message over the objects. For this, she can vary the parameter p , which expresses the amount of hidden context per cover object. ($p = 1$ means to spread the secret message as thinly as possible, $p = B$ ($0 < B \ll 1$) means to spread it in as few covers as possible.) With various assumptions about the steganalysis method Ker formulates a zero-sum game and finds *minmax* and *maxmin* solutions in pure strategies for both parties (under the assumption that one of the parties has to move first). Furthermore, when both parties move simultaneously (or do not know what strategy the other party follows), there is a unique Nash equilibrium in mixed strategies. Ker's conclusion is that the steganographer either spreads the secret message as thinly as possible over all covers or uses as few covers as possible with maximum embedding capacity.

4.3.3 Game Theory and Detection Performance

Orsdemir et al. [69] frame the competition between steganographer and steganalyst with the help of *set theory*. The steganographer has the possibility to use either a naïve or a sophisticated strategy, where in the sophisticated strategy she incorporates statistical indistinguishability constraints. The passive steganalyst can either assume a naïve steganographer or a sophisticated one and train a machine-learning based classifier on the respective assumption. This results in a matrix game with detection performance as payoffs. Unsurprisingly, a sophisticated steganographer performs better against a naïve steganalyst, but in their set-up, a sophisticated steganalyst performs worse against a naïve steganographer than against a sophisticated one. Thus, there is no equilibrium in pure strategies. The authors numerically calculate mixed strategy equilibria for specific embedding rates but no generally valid strategies are presented. As the embedding functions are black boxes, the resulting equilibria do not directly inform about the design of secure embedding functions or optimal detectors.

4.3.4 Game Theory and Adaptive LSB Matching

Following our first game-theoretic approach [77], recently other researchers picked up the topic of game theory and steganography. In [14] the authors examine the embedding operation of LSB matching with a content-adaptive embedding strategy. In their set-up, the cover model follows a simple multivariate Gaussian model, the strategies of both players consist of the probabilistic embedding strategy and the payoff is measured as the steganalyst's detection performance. The steganographer tries to minimize an additive distortion function (as defined in Section 3.1.2.2) and the steganalyst knows the payload p and the embedding costs ρ_i for all positions $i \in \{1, \dots, n\}$ and performs a likelihood-ratio (LR) test. For a comparison with information-theoretic optimal embedding, the authors introduce two models for the steganalyst: first, the *omnipotent* steganalyst who is granted exact knowledge about the steganographer's actions (i.e., the embedding probabilities) and secondly, the *ignorant* steganalyst who does not know the actions. Faced with an omnipotent steganalyst, the steganographer tries to minimize an information-theoretic measure, the KL divergence. The authors assume that the steganographer prefers to embed in positions that have a higher variance and when she chooses one of the positions, she either increases or decreases it by 1, with equal probability. With this embedding operation and the assumption about the image model, it is possible to express the stego distribution as a Gaussian mixture distribution. Then, the steganalyst makes a conjecture about this distribution and performs a LR test for each given object. Due to the fact that the steganalyst's distribution over embedding positions consists of a manifold convolution, the authors state that, in general, their solution has no closed form. So, the authors show numerically that there exists a unique Nash equilibrium and continue their analysis with a two-position cover. Here they show, again numerically, that the information-theoretic optimal and the game-theoretic optimal strategies differ and that the steganographer is better off with introducing slightly more distortion. But, as the heterogeneity in the cover source increases, the optimal strategies become more similar.

4.4 The Game-Theoretical Framework

We build on the protocol by Katzenbeisser and Petitcolas introduced in Section 2.2.2.2 which is already called a “game” by conventions in cryptography. We augment it with both players' strategies to make it a game in the sense of game theory and to obtain a payoff metric under equal priors. As the following results all deal with SSI realized for content-adaptive embedding, we formally define the key components in this settings that exceed the definitions given in Section 2.1. The framework itself is easily transformable to other settings where SSI is used in steganography, for example to a situation with an adaptive choice of the embedding direction.

4.4.1 Basic Definitions

For a given \mathcal{P}_0 and a uniform prior over encrypted messages, \mathcal{P}_1 depends on the embedding operation. The *Kullback–Leibler divergence* (KLD) between \mathcal{P}_0 and \mathcal{P}_1 is an information-theoretic measure of steganographic security with regard to undetectability (cf. Definition 2.10). We leverage this to distinguish between homogeneous and heterogeneous cover sources.

Definition 4.14 (Homogeneous vs Heterogeneous Cover Source). *A cover source $X^{(0)}$ is called homogeneous with regard to a fixed embedding operation, if for every $i, j \in \{0, \dots, n-1\}, i \neq j$, and for any subset of the cover space and the corresponding subsets of the stego spaces, it holds that $\text{KLD}(\mathcal{P}_0, \mathcal{P}_{(x_i)}) = \text{KLD}(\mathcal{P}_0, \mathcal{P}_{(x_j)})$. Otherwise the cover source is called heterogeneous.*

This definition implies that homogeneous cover sources offer the same security regardless of *where* in any given cover the embedding changes are made. For typical embedding operations, all i. i. d. and the common Markov cover models [19] are homogeneous cover sources. Because adaptive steganography exploits variation in uncertainty between embedding positions, we need to model heterogeneous cover sources. In this case, the security impact of changing individual embedding positions may depend on the realization $\mathbf{x}^{(0)}$. Therefore, we define a notion of suitability for embedding per position and per cover that is closely related to the *uncertainty with regards to positions* (cf. Definition 3.3) by decomposing the KLD measure into differences in the likelihood of hypothetical stego objects.

Definition 4.15 (Suitability). *Position i of cover $\mathbf{x}^{(0)}$ is more suitable for embedding than position j , if the stego object $\mathbf{x}_{(i)}^{(1)}$ is a more likely realization of the cover distribution \mathcal{P}_0 than the stego object $\mathbf{x}_{(j)}^{(1)}$, i. e., if $\mathcal{P}_0(\mathbf{x}_{(i)}^{(1)}) > \mathcal{P}_0(\mathbf{x}_{(j)}^{(1)})$.*

Recall that $\mathbf{x}_{(i)}^{(1)}$ is the cover object $\mathbf{x}^{(0)}$ with position i changed and $\mathcal{P}_0(\cdot)$ denotes the probability of occurrence under the cover distribution \mathcal{P}_0 and note that this definition is agnostic about multiple embedding changes appearing together, a common assumption in the literature [26].

Remark 4.1. *The definition of suitability extends the definition of uncertainty with regards to positions in two ways. First, it allows to compare different positions, something that is necessary to establish an order, and second, it is applicable for all positions that are between perfectly uncertain and perfectly informative.*

Since \mathcal{P}_0 is unknown for empirical cover sources, practical adaptive embedding functions use an *adaptivity criterion* to approximate the suitability of individual embedding positions, as shown in Section 3.3. For the use in our theoretical framework we define it as follows:

Definition 4.16 (Adaptivity Criterion). *A family of tractable functions, e. g., $\zeta_i : \{0, \dots, 2^\ell - 1\}^n \times \Theta \rightarrow \mathbb{R}$, is called adaptivity criterion if it establishes an order of all*

n embedding positions in a cover $\mathbf{x}^{(0)}$ by their approximate suitability. More specifically, $\zeta_i(\mathbf{x}^{(0)}, \boldsymbol{\theta}) > \zeta_j(\mathbf{x}^{(0)}, \boldsymbol{\theta})$ implies that, to the best of the steganographer's knowledge, position i appears more suitable for embedding than position j .

Definitions 4.15 and 4.16 require some reflection.

Remark 4.2. *The adaptivity criterion may use steganographic side information $\boldsymbol{\theta} \in \Theta$ to improve the quality of the approximation.*

Remark 4.3. *The mere order relation in Definition 4.16 ignores quantitative differences in the likelihoods of Definition 4.15.*

This is no drawback of the framework, as for example in [52] it is argued that most leading steganalysis methods base their decision on small groups of positions as well and view them as independent. So, we can assume that our framework captures most of the relevant properties in this regard.

Remark 4.4. *The assumption of a complete order is a simplification. Some practical schemes establish partial orders and resolve them with random (key-dependent) tie-breaking rules, as seen in Section 3.3.*

The framework is sufficiently expressive to study canonical embedding and detection strategies. Replacing the order with a quantitative detectability profile (cf. 3.1.2.2) or more realistic non-linear distortion functions is formally straightforward, but depends on detailed knowledge of the specific cover source.

Similar to Section 3.1.3, we write $\mathbf{y}^{(0)}$ for a cover $\mathbf{x}^{(0)}$ with elements ordered by *decreasing* suitability for embedding, i. e., $\zeta_{i-1}(\mathbf{y}^{(0)}, \boldsymbol{\theta}) \geq \zeta_i(\mathbf{y}^{(0)}, \boldsymbol{\theta})$ for $1 \leq i < n - 1$. Of course, the stego object is always transmitted with its symbols in original order. In practice, stego objects $\mathbf{x}^{(1)}$ often leak information about the values of ζ to the steganalyst (as shown in Section 3.3), who can thereby learn about likely embedding positions and thus *recover* the order of $\mathbf{y}^{(0)}$. We say that an adaptivity criterion is *perfectly recoverable* if $\hat{\mathbf{y}}^{(1)} = \mathbf{y}^{(1)}$, i. e., if it has a perfectly recoverable order, as in Definition 3.5. The framework is agnostic about quantifying this information leakage. Deviations from perfect recovery are best specified in the context of specific models.

4.4.2 Set-Up and Knowledge

Let *Alice* be the steganographer and *Eve* be the steganalyst. Eve knows the embedding function including its adaptivity criterion. Alice does not know the global cover distribution \mathcal{P}_0 , to prevent her from performing perfect steganography. Similarly, we require that Eve has no access to both global distributions \mathcal{P}_0 and \mathcal{P}_1 . This means that both actors, unaware of the global distributions, must resort to local models.

The different entities in our game are: *Nature*, *Alice*, the *Judge*, and *Eve*. Nature is the heterogeneous cover source that emits a cover $\mathbf{x}^{(0)}$ with n symbols, according to \mathcal{P}_0 . Upon receiving the cover from Nature, Alice changes exactly k bits. She changes position i of the reordered cover $\mathbf{y}^{(0)}$ with probability \bar{a}_i . The Judge is fair and forwards

to Eve with constant probability $\mu = 1/2$ either the cover or the stego object. In the jargon of game theory, the Judge is part of Nature. When Eve gets either the cover or the stego object, she recovers its order and examines symbol $\hat{y}_i^{(\bar{a}_i)}$ with probability \bar{e}_i . Then she decides about the type of object. The accuracy of her decision materializes in the error rates. These rates quantify steganographic security in our framework and thus the payoff for both players.

4.4.3 Strategies

We want to study both *pure* and *mixed* strategies. Alice's strategy space to change k bits out of n positions leads to $\binom{n}{k}$ pure strategies. We simplify this by assigning probabilities in mixed strategies to single positions and only look at the projection of the probabilities onto the positions. Note that we can identify pure strategies in this setting when we see mixed strategies with support k . We define the random binary vector \mathbf{A} , of which Alice's choice $\mathbf{a} = (a_0, \dots, a_{n-1})$ is a realization, and the random binary vector \mathbf{E} , of which Eve's choice $\mathbf{e} = (e_0, \dots, e_{n-1})$ is a realization. A value of $a_i = 1$ means that Alice changes $y_i^{(0)}$ for embedding, and $a_i = 0$ means she does not. Similarly, Eve examines $\hat{y}_i^{(\bar{\mathbf{a}})}$ only if $e_i = 1$.

Let $\bar{a}_i = \Pr(A_i = 1)$ and $\bar{e}_i = \Pr(E_i = 1)$ be Alice's, respectively Eve's, parameters in mixed strategies. This allows us to characterize six canonical strategies.

Definition 4.17 (Canonical Embedding Strategies).

The steganographer's embedding strategy is called ...

- a) random uniform, if $\forall i : \bar{a}_i = k/n$,
- b) naïve adaptive, if $\bar{a}_i = 1$ for $i \in \{0, \dots, k-1\}$ and $\bar{a}_i = 0$ otherwise, and
- c) optimal adaptive, if $\bar{\mathbf{a}} = \bar{\mathbf{a}}^*$, a unique equilibrium of the adaptive steganography game.

Definition 4.18 (Canonical Detection Strategies).

The steganalyst's detection strategy is called ...

- d) unweighted, if $\forall i : \bar{e}_i = k/n$,
- e) weighted, if $\bar{e}_i = 0$ for $i \in \{0, \dots, n-k-1\}$ and $\bar{e}_i = 1$ otherwise, and
- f) optimal adaptive, if $\bar{\mathbf{e}} = \bar{\mathbf{e}}^*$, a unique equilibrium of the adaptive steganography game.

Most practical embedding functions implement random uniform or naïve adaptive embedding (see Table 3.1 on page 39), as they are easy to implement and follow the first intuition of undetectable embedding.

Most steganalysis methods implement unweighted or weighted detection, again because they are easy to implement and yield good results against random uniform

embedding. Observe that weighted detection is blind to naïve adaptive embedding if $k < \frac{n}{2}$, as it puts all the weight on those positions the steganographer will never use.

Our goal in this thesis is to investigate *optimal adaptive* embedding and detection, as equilibrium strategies in our game-theoretical framework. For a better comparison with information-theoretically optimal strategies, we formally define these as well. For this definition, we grant that both Alice and Eve more knowledge than before. The only thing we fix is the, possibly imperfect, embedding function.

Definition 4.19 (Information-Theoretic Optimal Strategies).

A strategy is called information-theoretically optimal ...

- g) embedding, if $\bar{a} = \bar{a}^+ = \operatorname{argmin}_{\bar{a}} \operatorname{KLD}(\mathcal{P}_0, \mathcal{P}_{(\bar{a})})$, and
- h) detection, if the steganalyst performs a Likelihood Ratio test (LRT) between \mathcal{P}_0 and \mathcal{P}_1 .

Until now, information-theoretic optimal embedding was the “*holy grail*” in steganography. The appropriate information-theoretic measure, most commonly the KLD, is assumed to produce the most similar probability distribution \mathcal{P}_1 in comparison to \mathcal{P}_0 and thus the supposedly most secure steganography.

The implementation of either the information-theoretic optimal strategies (strategies g) and h)) requires knowledge of the cover distribution \mathcal{P}_0 for the steganographer, and additionally knowledge of the stego distribution \mathcal{P}_1 for the steganalyst. If the steganographer has full knowledge about \mathcal{P}_0 and additionally can influence the embedding function however she likes, $\min_{\bar{a}} \operatorname{KLD}(\mathcal{P}_0, \mathcal{P}_{(\bar{a})})$ will always be 0 and thus she could perform perfect steganography [88]. Forcing her to use an imperfect embedding function, she still would have to know the cover distribution to perform information-theoretically optimal embedding, which is assumed to be unfeasible for real-world cover sources [7].

The LRT $\frac{\mathcal{P}_0(\mathbf{y}^{(0)})}{\mathcal{P}_1(\mathbf{y}^{(0)})} \stackrel{?}{>} \gamma$ for a given realization $\mathbf{y}^{(0)}$ and an optimal threshold γ is, following the Neyman-Pearson lemma [68], the information-theoretically optimal test to distinguish between objects from distributions \mathcal{P}_0 and \mathcal{P}_1 . Giving the steganalyst knowledge of both these distributions lets her always detect at the information-theoretic bound. Although this situation follows the strictest interpretation of Kerckhoffs’ principle for steganography [20], it is as much infeasible for real-world cover sources as information-theoretic embedding and will result in an optimization problem, not necessarily in a game-theoretic setting. Still, we will use the information-theoretically optimal strategies for comparison with the optimal adaptive strategies.

4.5 Summary

In this chapter, we argued why side-informed steganography should be studied with game theory and other approaches, e.g. information-theoretical ones, do not capture all the relevant properties of this situation. Then, we introduced the basic game-theoretical notation reduced to the case of two-player zero-sum games and the required solution concepts established in the game theory literature.

After giving an brief overview of the, admittedly sparsely, existing game-theoretic approaches in steganography, we formally defined the key components of our game-theoretical framework. We formally defined

- ▶ heterogeneous cover sources,
- ▶ suitability and adaptivity criteria,
- ▶ the players involved and their respective knowledge,
- ▶ six canonical embedding and detection strategies, and, for benchmarking
- ▶ the information-theoretically optimal counterparts.

In the next chapter we instantiate the framework with an embedding operation and different cover generation models. We derive locally optimal detection rules and characterize the equilibrium strategies. Then, we formulate the insights that can be drawn from our artificial models for real-world scenarios.

Chapter 5

Game-Theoretic Insights

The purpose of this chapter is to instantiate the framework introduced in the previous chapter to gain game-theoretic insights from a thorough analysis of these instantiations. We equip the framework with concrete cover sources, embedding operations and adaptivity criteria.

Overall, we decided to include two different set-ups, namely binary cover objects of length n and cover objects of length 2 over an integer alphabet. All set-ups resemble different important properties of real-world cover sources but are simple enough to find analytical solutions.

Both basic set-ups are further divided into overall five concrete models, as follows:

- ▶ **Binary Alphabet, Arbitrary Positions (Sec. 5.1):** cover objects are binary sequences of length n . The suitability of all positions is measured by an abstract function f that returns the probability of the positions to take their more likely value. The order of the positions is fully known to both Alice and Eve, thus, we have perfectly recoverable order.
 - ▷ **Restricted Steganalyst (Sec. 5.1.1) :** In this instantiation, Eve is restricted to base her decision on the value of only one position. Although she knows the order of the positions perfectly, this second kind of SSI is thereby almost perfect in that Eve cannot recover its values for $n - 1$ positions.
 - ▷ **Powerful Steganalyst:** In this instantiations, Eve is allowed to perform a likelihood-ratio test for each sequence she observes. By this, we ensure that the optimal strategy for Alice holds even against the most powerful steganalyst. In this scenario, we distinguish between different kinds of embedding that influence the shape of the stego distribution \mathcal{P}_1 .
 - **Fixed Net Embedding Rate (Sec. 5.1.2):** In this instantiation, Alice has to embed exactly k bits. This implies that she chooses probabilities for subsets of length k to embed in.
 - **Independent Embedding (Sec. 5.1.3):** In this instantiation, Alice has to embed independently with an expectation of k bit. By this, she chooses embedding probabilities for single positions and the positions in the stego distribution remain independent.
- ▶ **Integer Alphabet, Two Positions (Sec. 5.2):** cover objects have two positions from an integer alphabet. The suitability of both positions is described by a probability mass function (PMF) which states the probability of occurrence for all values. We first assume the order in both instantiations to be fully recoverable, but relax this condition to an arbitrary recovery rate in Section 5.2.3. Both Alice and Eve know the cover generating PMF, the embedding function and the recovery rate.

- ▷ Linear Probability Mass Function (Sec. 5.2.1): In this instantiation, the PMF of the cover source is linearly increasing. With this PMF we have a strict order of occurrence of the different values and can model homogeneous and heterogeneous cover sources by adjusting the slope of the PMF.
- ▷ Constant Ratio Probability Mass Function (Sec. 5.2.2): In this instantiation, the PMF of the cover source increases with a constant ratio. This PMF converges asymptotically to a discretized Laplace distribution which is known to model the marginal distribution of real transform-coded covers reasonably well. Furthermore, we can easily calculate the KLD in this model.

We justify every instantiation and show the game-theoretically optimal strategies and illustrate them numerically.

Finally, we leverage the insights from all instantiations combined to give directions for more secure adaptive embedding for real-world cover sources (Section 5.3.1) and highlight the limitations of our approach (Section 5.3.2). We find that the concept of the equalizer strategies (cf. Definition 4.13) can be extended to an embedding strategy, thus creating an *equalizer embedding strategy*.

5.1 Cover Models with Binary Embedding Positions

In this section, we assume the cover source emits binary objects of length n .

To formalize the role of side information, here the adaptivity criterion, let the embedding domain of a cover be a random sequence of n symbols with varying uncertainty, derived from some kind of side information. This side information is fully available to Alice and partly available to Eve. For a better readability, we sort the cover and stego sequences according to their *suitability*, as in Definition 4.15. We assume that both players can exactly reconstruct the order of the symbols by decreasing suitability.

Formally, we consider a vector $\mathbf{Y} = (Y_0, \dots, Y_{n-1})$ of independent random variables drawn from a binary alphabet $\mathbb{X} = \{0, 1\}$, with realizations $\mathbf{y} = (y_0, \dots, y_{n-1})$. Note that real covers may have a larger alphabet, but we settle on bits for a clearer notion of suitability. Moreover, practical embedding functions often work on a vector of binary residuals, such as the sequence of all least significant bits, as already shown in Section 2.1.2. Similarly, popular detectors leverage the concept of residuals as prediction errors from a local image model [31].

The monotonically increasing function $f(i) : \{0, \dots, n-1\} \rightarrow [\frac{1}{2}, 1]$ defines the probability of Y_i taking its most likely value. Without loss of generality, let $f(i) = \Pr(Y_i = 1)$ for the analysis.

To anchor the two ends of the suitability range, we require $f(0) = \frac{1}{2} + \varepsilon$ and $f(n-1) = 1 - \varepsilon$. We need a strictly positive ε to ensure that we have neither perfect uncertainty nor perfect information for any of the positions. If ε was zero, i.e., perfect uncertainty, Alice could embed at least one bit into y_0 without risk of detection. Similarly, if $P(Y_{n-1} = 1) = 1$, i.e., perfect information, embedding into y_{n-1} would allow detection with certainty. For $\varepsilon \rightarrow 0$, we can also write $f(i) : \{0, \dots, n-1\} \rightarrow (\frac{1}{2}, 1)$.

This model still holds if there are some positions that are either perfectly uncertain or perfectly informative. As long as there are less than the k bits that are perfectly uncertain, we can think of our set-up in such a way that we reduce the sequences to positions in the open interval and have to embed less into these sequences.

To simplify the exposition of our results, we introduce the notation

$$\tilde{f}(i) = f(i) - \frac{1}{2}. \quad (5.1)$$

The function $f(i)$ was introduced as the probability of seeing 1 at position i , and it measures the suitability of position i . The function $\tilde{f}(i)$ can be interpreted as measuring the bias of position i , that is the deviation from perfect uncertainty.

We assume throughout the whole section that the positions in the cover distribution \mathcal{P}_0 are independently distributed so that:

$$\mathcal{P}_0[\mathbf{Y} = \mathbf{y}] = \prod_{i=0}^{n-1} \mathcal{P}_0[Y_i = y_i] \quad (5.2)$$

$$= \prod_{y_i=1} f(i) \cdot \prod_{y_i=0} (1 - f(i)) \quad (5.3)$$

$$= \prod_{i=0}^{n-1} (1 - f(i) + 2y_i \tilde{f}(i)). \quad (5.4)$$

We will examine three different specifications with this set-up. First, we model a restricted steganalyst who knows the order of the symbols but can query its exact value only for one position. Then, we remove this restriction of the steganalyst and allow her to base her decision on all n positions, first with a fixed net embedding rate of k bit on the steganographer's side, and then with a situation where the steganographer performs independent embedding of an expected length of k .

5.1.1 Restricted Steganalyst Model

The set-up of our first model is depicted in Figure 5.1. It differs from the standard model (cf. Figure 2.1(b) on page 8) by allowing Eve to query the side information for one position in the cover directly. Recall from Section 3.3 that practical steganalysis can often estimate such side information from the observed object. Therefore, we will elaborate below why we require this explicit interaction in this game.

We begin by formulating Eve's local decision rule. Eve observes the probability $f(i)$ that bit i is 1. Since $f(i)$ is greater than $\frac{1}{2}$, the object is more likely to be a cover if the observed bit is 1, and more likely to be stego if the observed bit is 0. This constrains Eve's decision rule based on her observation at position i .

$$DR(i) = \begin{cases} \text{cover} & \text{if } x_i = 1 \\ \text{stego} & \text{if } x_i = 0 \end{cases}. \quad (5.5)$$

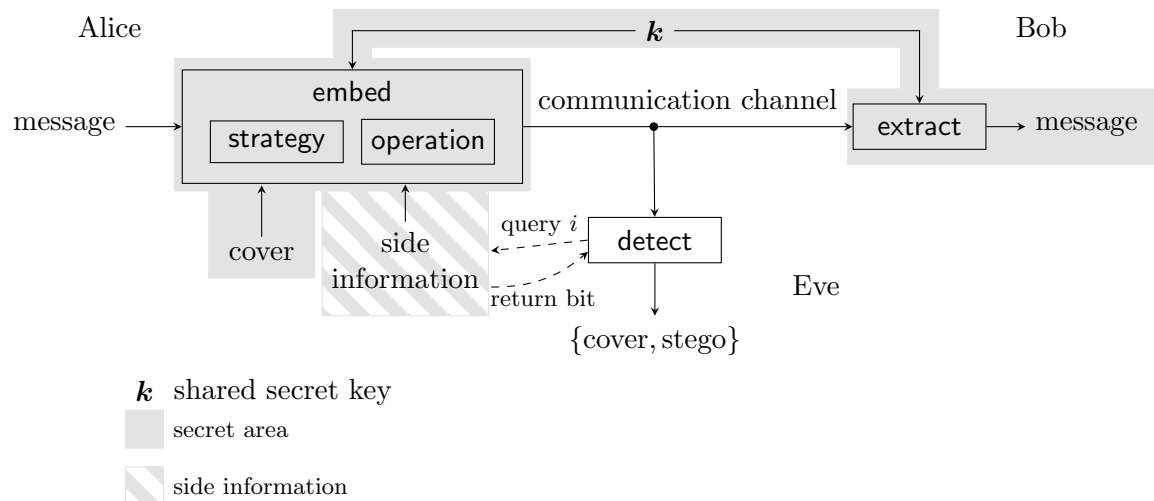


Figure 5.1: Block diagram of steganographic communication system with side information and the possibility for Eve to query one position.

5.1.1.1 Strategies

Alice's action space is to flip k bits of a given cover realization $\mathbf{y}(0)$ to embed a hidden message. Alice chooses a k -sized subset of $\{0, \dots, n-1\}$ indicating the embedding positions. Her mixed strategy action space is a probability distribution \bar{a} over all k -sized subsets of $\{0, \dots, n-1\}$.

Eve tries to decide whether an observed bit vector is a cover or a stego object. We model equal priors over cover and stego objects by assuming that the Judge flips an unbiased coin to decide if Eve sees Alice's stego object or a cover drawn from the cover source \mathcal{P}_0 . Eve's optimal decision rule would be a likelihood ratio test using the joint distributions over all cover and stego objects, \mathcal{P}_0 and \mathcal{P}_1 . In practice, however, \mathcal{P}_0 and \mathcal{P}_1 are unknown and Eve can only make local decisions for individual symbols using a local predictor. We stipulate that Eve can use her knowledge about the marginal distributions of \mathcal{P}_0 and \mathcal{P}_1 to make optimal local decisions, although this is not always the case for practical steganalysis. While this game might be too optimistic for Eve in this respect, we contrast this by requiring that Eve only looks at one position. To justify this constraint in this model, we assume two different kinds of SSI. The SSI that is responsible for the order of the suitability is fully reconstructible, and thus perfectly known to Eve. The second kind of SSI, which measures the deviation from perfect uncertainty, i.e., the bias, is almost perfect and can be queried for only one position. For this, we assume an interactive query mechanism (see Fig. 5.1). As a result, Eve's mixed strategy space is a probability distribution \bar{e} over all n positions for which she can query the most likely value. For all other positions, she cannot tell if $P(Y_i = 0) > P(Y_i = 1)$ or $P(Y_i = 0) < P(Y_i = 1)$ in covers. Therefore, she does not gain any information from

Table 5.1: Payoff for (Eve, Alice)

Eve's decision	Reality	
	cover	stego
cover	(1, -1)	(-1, 1)
stego	(-1, 1)	(1, -1)

including the values at these positions in her decision.

It is obvious that Eve's task is very hard in this set-up, because if $k = 1$, her advantage over random guessing is not better than ε even if Alice deterministically embeds in the first symbol, i.e., if she performs naïve adaptive embedding. If Alice randomizes her strategy, then Eve's advantage shrinks. If Alice embeds more bits, Eve's advantage increases because Alice has to use less suitable positions. Our objective is to quantify by how much, and if there is an equilibrium.

The following objective function defines a zero-sum game: Alice tries to increase her security by maximizing Eve's decision error, whereas Eve tries to minimize it. We map this to the payoff structure given in Table 5.7. Note that this payoff matrix induces an objective function based on the equal error rate (EER), as defined in Section 2.2.3. For practical applications, the payoff matrix might need adjustment to account for the harm caused by false positive and false negatives, respectively.

Figure 5.2 on page 74 summarizes the game for $k = 1$ in an extensive form graph. From left to right, first, nature draws a cover from \mathcal{P}_0 , then Alice chooses her single (because $k = 1$) embedding position, creating a stego object (black nodes). A coin flip, invisible to Eve, decides whether she sees the stego or cover object. Then Eve chooses the position she wants to compare with a prediction to make her decision, and outputs the decision result (c for cover or s for stego). Shaded nodes indicate the cases where Eve wins, i.e., she receives positive payoff.

Recall that Alice's mixed strategy space is a probability distribution over size- k subsets of $\{0, \dots, n-1\}$. For a subset S of k positions, \bar{a}_S is the probability that Alice embeds her bits in these k positions; and we have $\sum_S \bar{a}_S = 1$. Overloading notation, we define the projection of Alice's mixed strategy onto positions to be the total probability that Alice embeds in position i . Formally, we define \bar{a}_i for $i \in \{0, \dots, n-1\}$ as

$$\bar{a}_i = \sum_{\{S:i \in S\}} \bar{a}_S. \quad (5.6)$$

If Alice embeds in just one position, then $\bar{a}_i = \bar{a}(\{i\})$ and $\sum_{i=0}^{n-1} \bar{a}_i = 1$. If Alice embeds k bits, then

$$\sum_{i=0}^{n-1} \bar{a}_i = k. \quad (5.7)$$

Eve's mixed strategy action space is a distribution over positions. Eve queries the most likely value of position i with probability \bar{e}_i and decides stego or cover based only on her observation at position i .

5.1.1.2 Payoff

We quantify the payoff of Eve and Alice as a function of the bias $\tilde{f}(i)$ at each position, Eve's mixed strategy \bar{e}_i , and Alice's mixed strategy \bar{a}_i .

Theorem 5.1 (Game Outcome). *If \tilde{f} is the bias function, \bar{e} is Eve's mixed strategy, and \bar{a} is Alice's mixed strategy, then the total expected payoff for (Eve, Alice) is*

$$\left(2 \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i), -2 \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i) \right). \quad (5.8)$$

Proof. First assume that Eve looks only at position i . Under this assumption, we may determine the probability she wins the game by enumerating all possible ways the world could be, and adding up the respective probabilities. We may think of the process as an orderly sequence of events. First, the Judge chooses whether Eve sees a cover object or a stego object by flipping an unbiased coin. The cover object is then instantiated with a realization x_i of position i , with $P(X_i = 1) = f(i)$. If the Judge chose stego, then Alice flips bit i with probability \bar{a}_i . Finally, Eve decides whether the object is cover or stego by looking at her observed bit. She decides cover if the bit is 1 and stego if the bit is 0. Table 5.2 records the events, probabilities, and decision outcomes for each possible case.

Table 5.2: Game outcome in different states of the world

Reality	Value of $x_i^{(0)}$		Probability	Eve's decision	Winner
	Cover	Observed			
c	1	1	$\frac{1}{2} \cdot f(i)$	c	Eve
c	0	0	$\frac{1}{2} \cdot (1 - f(i))$	s	Alice
s	1	0	$\frac{1}{2} \cdot f(i) \cdot \bar{a}_i$	s	Eve
s	1	1	$\frac{1}{2} \cdot f(i) \cdot (1 - \bar{a}_i)$	c	Alice
s	0	1	$\frac{1}{2} \cdot (1 - f(i)) \cdot \bar{a}_i$	c	Alice
s	0	0	$\frac{1}{2} \cdot (1 - f(i)) \cdot (1 - \bar{a}_i)$	s	Eve

Legend: c = cover, s = stego

Given that Eve looks only at position i , her probability of winning is

$$\frac{1}{2} (f(i) + f(i)\bar{a}_i + (1 - f(i))(1 - \bar{a}_i)) \quad (5.9)$$

$$= \frac{1}{2} (f(i) + f(i)\bar{a}_i + 1 - \bar{a}_i - f(i) + f(i)\bar{a}_i) \quad (5.10)$$

$$= \frac{1}{2} (1 + 2f(i)\bar{a}_i - \bar{a}_i) \quad (5.11)$$

$$= \frac{1}{2} + \bar{a}_i \left(f(i) - \frac{1}{2} \right) \quad (5.12)$$

$$= \frac{1}{2} + \bar{a}_i \tilde{f}(i). \quad (5.13)$$

Hence Eve's total probability of winning is

$$\sum_{i=0}^{n-1} \bar{e}_i \left(\frac{1}{2} + \bar{a}_i \tilde{f}(i) \right) \quad (5.14)$$

$$= \frac{1}{2} + \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i). \quad (5.15)$$

And thus Eve's total expected game payoff is

$$\Pr(\text{Eve wins}) \cdot 1 + \Pr(\text{Eve loses}) \cdot (-1) \quad (5.16)$$

$$= \left(\frac{1}{2} + \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i) \right) \cdot 1 + \left(\frac{1}{2} - \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i) \right) \cdot (-1) \quad (5.17)$$

$$= 2 \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i). \quad (5.18)$$

With the zero-sum property, the total expected payoff for (Eve, Alice) is thus

$$\left(2 \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i), -2 \sum_{i=0}^{n-1} \bar{e}_i \bar{a}_i \tilde{f}(i) \right). \quad (5.19)$$

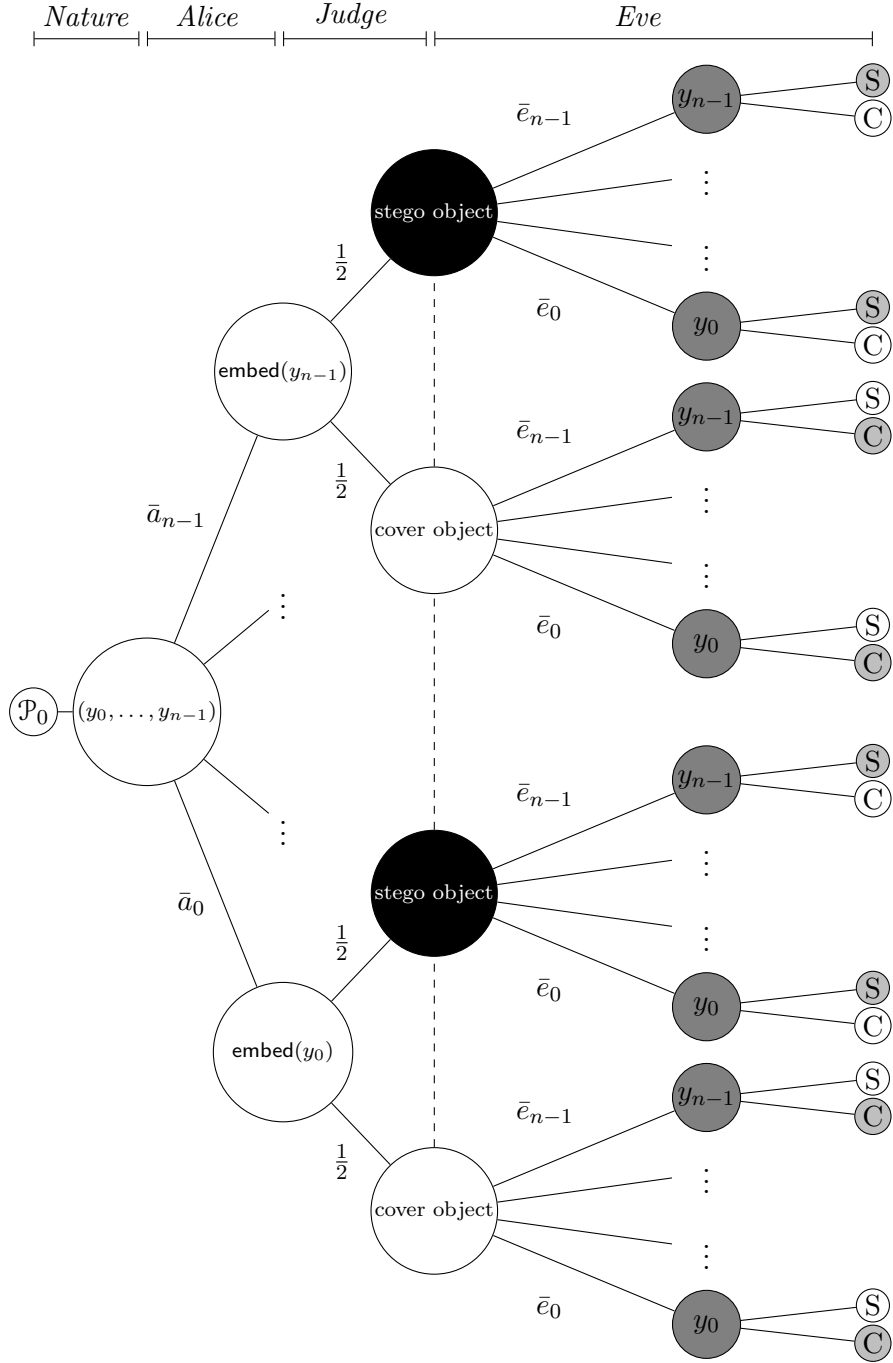
□

5.1.1.3 Solving the Game

We now turn our attention to the game's Nash equilibria.

5.1.1.3.1 Hiding One Bit

We start with analyzing the case of $k = 1$. This simplifies Alice's mixed strategy action space to a probability distribution over the set $\{0, \dots, n - 1\}$.



Cover generation *Embedding strategy* *Coin flip* *Query* *Decision*
 Figure 5.2: Extensive form of the game for $k = 1$. The dashed line indicates Eve's information set. The dark gray nodes represent Eve's query strategy and the light gray nodes are the situations in which Eve wins the game.

Lemma 5.1 (Exclusion of pure strategies). *There is no equilibrium in which either Alice or Eve assigns zero probability to any i .*

Proof. Assume Alice assigns zero probability to position i . Then Eve gains no advantage from assigning positive probability to position i . Hence, Eve's best response would assign zero probability to position i . But then Alice can completely eliminate Eve's advantage by assigning probability 1 to position i . So Alice is not in equilibrium.

Assume Eve assigns zero probability to position i , then Alice can completely eliminate Eve's advantage by assigning probability 1 to position i . But then Eve's best response would be assign probability 1 to position i . So Eve is not in equilibrium. \square

It is useful to quantify Eve's advantage from looking at one position and observing the most likely value. The following two definitions facilitate such quantification.

Definition 5.1 (Eve's Local Advantage). *Eve's local Advantage at position i is $\bar{a}_i \cdot \tilde{f}(i)$.*

Definition 5.2 (Eve's Total advantage). *Eve's total advantage is the weighted sum over all her local advantages at positions $0, \dots, n-1$, i. e., $\sum_{i=0}^{n-1} (\bar{e}_i \bar{a}_i \tilde{f}(i))$.*

Observe that from Theorem 5.1, Eve's expected game payoff is exactly twice her total advantage. Hence we may consider total advantage as a quantity of primary interest. Eve's primary objective is to increase her total advantage, while Alice's primary objective is to reduce it. Our next lemma characterizes the structure of possible equilibria in relation to Eve's local and total advantages.

Lemma 5.2 (Uniform local advantage condition). *A necessary condition for any equilibrium is that Eve's local advantage is uniform over $i = 0, \dots, n-1$.*

Proof. Suppose Eve's local advantage is not uniform. Then there is at least one position i where her local advantage is not as high as it is at some other position j . I. e., $\bar{a}_i \cdot \tilde{f}(i) < \bar{a}_j \cdot \tilde{f}(j)$. Eve can then strictly increase her total advantage by setting $\bar{e}_j = \bar{e}_j + \bar{e}_i$ and then setting $\bar{e}_i = 0$. The resulting difference in her total advantage will be $\bar{e}_i(\bar{a}_j \cdot \tilde{f}(j) - \bar{a}_i \cdot \tilde{f}(i))$, which is positive. So the situation is not an equilibrium. \square

This condition can actually be fulfilled.

Lemma 5.3 (Existence of Alice's unique strategy). *In any equilibrium, Alice's strategy to embed one bit is*

$$\bar{a}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}. \quad (5.20)$$

Proof. We start with the condition from Lemma 5.2,

$$\bar{a}_i \cdot \tilde{f}(i) = \bar{a}_j \cdot \tilde{f}(j) \quad \forall i \neq j. \quad (5.21)$$

This implies that there is a constant τ with $\bar{a}_i \cdot \tilde{f}(i) = \tau$ for each i , and hence $\bar{a}_i = \frac{\tau}{\tilde{f}(i)}$ for some τ .

Now by the probability axiom,

$$\sum_{i=0}^{n-1} \bar{a}_i = 1, \quad (5.22)$$

so that

$$\sum_{i=0}^n \frac{\tau}{\tilde{f}(i)} = 1, \quad (5.23)$$

and hence

$$\tau = \frac{1}{\sum_{i=0}^n \tilde{f}(i)}. \quad (5.24)$$

It follows that

$$\bar{a}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}. \quad (5.25)$$

I. e. the two constraints (5.21) and (5.22) completely determine \bar{a}_i . \square

Lemma 5.4 (Game outcome in equilibrium). *The game's outcome for (Eve, Alice) in equilibrium is*

$$\left(\frac{2}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \frac{-2}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}} \right). \quad (5.26)$$

Proof. Alice's strategy fixes Eve's total advantage, which in turn fixes Eve's payoff. As Alice has only one strategy in equilibrium, we know Eve's total advantage in equilibrium must be

$$\sum_{i=0}^{n-1} (\bar{e}_i \bar{a}_i \tilde{f}(i)) = \sum_{i=0}^{n-1} \left(\bar{e}_i \frac{\tilde{f}(i)}{\tilde{f}(i) \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}} \right) = \frac{1}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.27)$$

hence Eve's payoff in equilibrium is $\frac{2}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}$ and the result follows. \square

Turning now to Eve's strategy, we may construe her objective as preserving her total advantage.

Lemma 5.5 (Uniform weighted bias condition). *A necessary condition for any equilibrium is that $\bar{e}_i \cdot \tilde{f}(i)$ is uniform over $i = 0, \dots, n - 1$.*

Proof. Suppose Alice is playing her unique strategy in equilibrium from Lemma 5.3 and that, for the sake of contradiction, there exist $i \neq j$ with $\bar{e}_i \cdot \tilde{f}(i) < \bar{e}_j \cdot \tilde{f}(j)$. Then Alice can decrease Eve's total advantage by adopting a new strategy \bar{a}' with, $\bar{a}'_j = 0$; $\bar{a}'_i = \bar{a}_i + \bar{a}_j$; and $\bar{a}'_r = \bar{a}_r$ for $r \neq i, j$.

The difference in Eve's total advantage is

$$\sum_{r=0}^{n-1} \left(\bar{e}_r \bar{a}'_r \tilde{f}(r) \right) - \sum_{r=0}^{n-1} \left(\bar{e}_r \bar{a}_r \tilde{f}(r) \right) \quad (5.28)$$

$$= \bar{e}_i \bar{a}'_i \tilde{f}(i) + w_j \bar{a}'_j \tilde{f}(i) - (\bar{e}_i \bar{a}_i \tilde{f}(i) + \bar{e}_j \bar{a}_j \tilde{f}(i)) \quad (5.29)$$

$$= \bar{e}_i (\bar{a}_i + \bar{a}_j) \tilde{f}(i) - (\bar{e}_i \bar{a}_i \tilde{f}(i) + \bar{e}_j \bar{a}_j \tilde{f}(j)) \quad (5.30)$$

$$= \bar{e}_i \bar{a}_j \tilde{f}(i) - \bar{e}_j \bar{a}_j \tilde{f}(j) \quad (5.31)$$

$$= \bar{a}_j (\bar{e}_i \tilde{f}(i) - \bar{e}_j \tilde{f}(j)) \quad (5.32)$$

$$< 0.$$

So Alice would prefer to change strategies, in violation of the equilibrium condition. \square

Lemma 5.6 (Existence of Eve's unique strategy). *In any equilibrium for the one-bit case, Eve's probability \bar{e}_i of looking at position i must be the same as Alice's probability of embedding at position i :*

$$\bar{e}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}. \quad (5.33)$$

Proof. The formula follows from the uniform weighted bias condition: $\bar{e}_i \cdot \tilde{f}(i) = \bar{e}_j \cdot \tilde{f}(j)$ for all $i \neq j$; and the probability constraint on Eve's mixed strategy: $\sum_{j=0}^{n-1} \bar{e}_j = 1$. The argument that these conditions uniquely determine a function is given in Lemma 5.3. \square

Theorem 5.2 (Unique Nash equilibrium). *There is a unique Nash equilibrium for the one-bit game where Alice embeds in position i with probability*

$$\bar{a}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.34)$$

and Eve observes position i with probability

$$\bar{e}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.35)$$

and the expected payoff outcome for (Eve, Alice) is $\left(\frac{2}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, -\frac{2}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}} \right)$.

Proof. See Lemmas 5.3, 5.4, and 5.6. \square

5.1.1.3.2 Hiding k Bits

Lemma 5.7 (Alice's k -bit strategy). *In any equilibrium, Alice's mixed strategy distribution satisfies*

$$\bar{a}_i = \frac{k}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}. \quad (5.36)$$

Proof. First, any equilibrium must satisfy the uniform advantage condition, as the logic from Lemma 5.2 applies also in the k -bit case. Thus we have

$$\bar{a}_i \cdot \tilde{f}(i) = \bar{a}_j \cdot \tilde{f}(j) \quad \forall i \neq j. \quad (5.37)$$

Since we also have

$$\sum_{i=0}^{n-1} \bar{a}_i = k, \quad (5.38)$$

the a_i are completely determined as $\bar{a}_i = \frac{k}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}$. □

Remark 5.1. Note that in the k -bit case, depending on f , single values for \bar{a}_i can be larger than 1 when calculated with Equation (5.36).¹⁴ These are positions that Eve would assign weight 0, as she would gain more from other positions $j \neq i$. We can think of these positions as “gifts” for Alice, as Eve’s strategy to look at them is strictly dominated by a strategy that assigns higher probability to other positions. The logic that there exists no pure strategy equilibrium from Lemma 5.1 still applies.

Lemma 5.8 (Eve’s k -bit strategy). *In any equilibrium, Eve’s mixed strategy distribution is*

$$\bar{e}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}. \quad (5.39)$$

Proof. Eve’s strategy must satisfy the uniform weighted bias condition: $\bar{e}_i \cdot \tilde{f}(i)$ is uniform in i ; as the logic from Lemma 5.5 still applies in the k -bit case. Since we also have $\sum_{i=0}^{n-1} \bar{e}_i = 1$, these two conditions imply $\bar{e}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}$. □

Theorem 5.3 (k -bit Nash equilibria). *There is a Nash equilibrium for the k -bit game where Alice’s strategy satisfies*

$$\bar{a}_i = \frac{k}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.40)$$

and Eve observes position i with probability

$$\bar{e}_i = \frac{1}{\tilde{f}(i) \cdot \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.41)$$

and the expected payoff outcome for (Eve, Alice) is $\left(\frac{2k}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, -\frac{2k}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}} \right)$.

The equilibrium is unique up to the projection of Alice’s mixed strategy.

¹⁴We would like to thank Aron Laszka from Budapest University of Technology and Economics, Hungary, for pointing out this fact to us in private conversation and a working paper of his.

Proof. See Lemmas 5.7 and 5.8 for the strategies. For the payoffs, note that Eve's advantage in equilibrium is

$$\sum_{i=0}^{n-1} \left(\bar{e}_i \bar{a}_i \tilde{f}(i) \right) = \sum_{i=0}^{n-1} \left(\bar{e}_i \frac{k \tilde{f}(i)}{\tilde{f}(i) \sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}} \right) = \frac{k}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}, \quad (5.42)$$

so that Eve's payoff in equilibrium is $\frac{2k}{\sum_{j=0}^{n-1} \frac{1}{\tilde{f}(j)}}$. \square

The following two corollaries are easily observable.

Corollary 5.1. *Eve's mixed strategy in equilibrium is independent of the number of embedded bits.*

Corollary 5.2. *Eve's expected payoff in equilibrium increases linearly with the number of embedded bits.*

Corollary 5.1 stipulates that the steganalyst's equilibrium strategy does not depend on the number of embedded bits. This is a handy property for the construction of detectors, where no knowledge of the hidden message length must be assumed. Corollary 5.2 states that if the detector follows the equilibrium strategy, its success rate increases linearly with the number of embedded bits. This deviates from the square root law of steganographic capacity, which predicts asymptotically quadratic advantage even for homogeneous covers [57]. The reason for this difference is that our detector is constrained to a locally optimal decision rule.

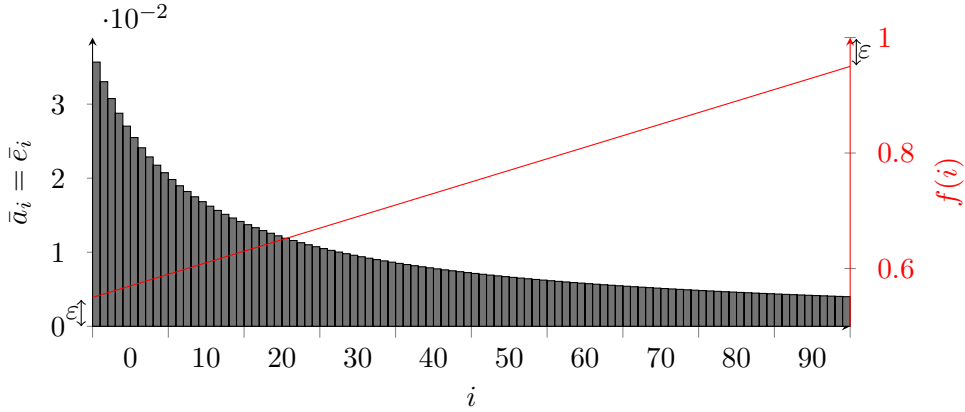


Figure 5.3: Equilibrium strategies for $\varepsilon \gg 0$ and a linear function f

5.1.1.4 Numerical Illustration

Figures 5.3 and 5.4 display numerical examples of the equilibrium in this instantiation of our game-theoretical framework with parameters $k = 1$ and $n = 100$.

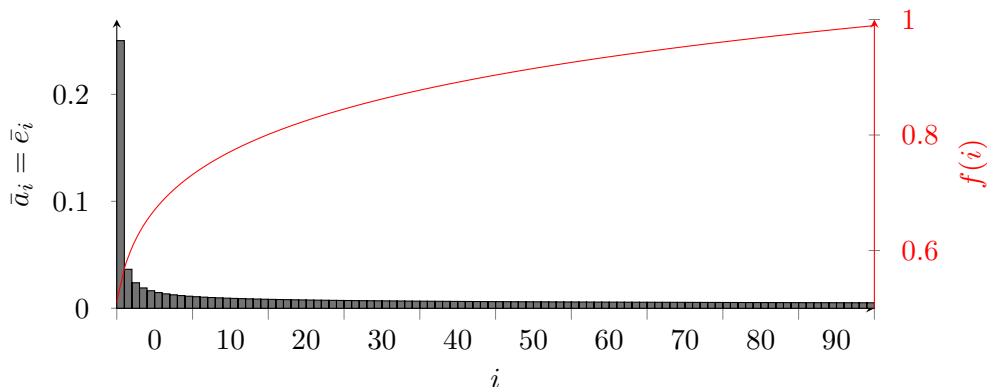


Figure 5.4: Equilibrium strategies for $\varepsilon \approx 0$ and a non-linear function f

The red line shows the suitability function $f(i)$; note the right-hand scale. From this it follows that on the left side there are the least predictable, i.e., most uncertain, positions and on the right side the positions are almost perfectly informative. The gray bars display Alice’s and Eve’s identical optimal strategies (left-hand scale). The higher a bar, the higher is the probability for Alice and Eve to choose this position for embedding and querying, respectively. Note the different scales of the left axis in both figures.

In Figure 5.3, the parameter ε is set relatively high and the suitability function f is linear. We see that the equilibrium strategies reflect the decrease in suitability by decreasing monotonically. Figure 5.4 is more realistic. It shows a small ε and a non-linear suitability function f with the majority of positions being relatively unsuitable, just like large homogeneous areas in natural images. It follows the intuition that the value of \bar{a}_0 is even higher than in Figures 5.3, as this position is more suitable and the other positions are mainly less suitable.

Both figures show that the value of \bar{a}_0 is at its maximum. This illustrates the claimed advantage of content-adaptive embedding over random uniform embedding if the cover source produces heterogeneous covers. Nonetheless, the fact that $\bar{a}_i > 0$ for all i suggests that the steganographer should potentially use every available position and not only the least predictable ones, as it would be the case in naïve adaptive embedding.

5.1.2 Powerful Steganalyst and Fixed Net Embedding

In this instantiation of the game-theoretic framework we loosen the constraints on Eve and allow her to obtain information from the whole object she examines. By this, her strategy becomes a function of whole sequences \mathbf{y} instead of probabilities to examine single positions within the sequences. We even allow her to calculate possible stego distribution functions \mathcal{P}_1 and then to perform a likelihood ratio test (LRT), the information-theoretically optimal strategy.

This instantiation originates from joint work with Aron Laszka, Benjamin Johnson,

and Jens Grossklags. To maintain consistency with the original publication, we use a slightly different notation in this section. To describe this game-theoretic model, we explicitly specify the set of states that the world can be in, the set of choices available to the players, and the set of consequences as a result of these choices. Because the game is a randomized extension of a deterministic game, we first present the structure of the deterministic game, and follow up afterwards with details of the randomization.

The event space Ω is the set $\{0, 1\}^n \times \{c, s\}$. An event consists of two parts: a binary sequence $y \in \{0, 1\}^n$ and a steganographic state $z \in \{c, s\}$, where c stands for *cover* and s for *stego*. The binary sequence represents the object Eve observes on the communication channel. The steganographic state tells whether or not a message is embedded in the sequence. In the randomized game, neither of these two states is known by the players until after they make their choices. To define payoffs for the finite game, we simply assume that some event has been chosen by Nature (including the coin flip of the Judge) so that the world is in some fixed state (\mathbf{y}, z) . Alice embeds a secret message of length k into the binary sequence \mathbf{y} ; Nature determines whether the original cover or the modified stego object appears on the communication channel; Eve observes the sequence appearing on the channel and makes a decision as to whether or not it contains a message.

5.1.2.1 Strategies

Alice's (pure strategy) choice is to select a size- k subset I of $\{0, \dots, n-1\}$, which represents the positions into which she embeds her encoded message, by flipping the value of the given sequence at each of the positions in I . By this, Alice would perform random uniform embedding if she chooses every I with the same probability and naïve adaptive embedding if she always chooses $I = \{0, \dots, k-1\}$.

Eve's (pure strategy) choice is to select a subset E_s of $\{0, 1\}^n$, which represents the set of sequences that she classifies as stego objects, i.e., sequences containing a secret message. Objects in $E_c := \{0, 1\}^n \setminus E_s$ are classified as cover objects, i.e., sequences not containing a secret message. Eve's strategy space in this model differs from the canonical detection strategies (cf. Definition 4.18) in that she has to decide beforehand which sequences are cover and which are stego objects. As we assume Eve to be able to perform a LR test between sequences, i.e., she is a powerful steganalyst, this is the information-theoretic optimal strategy, as defined in Definition 4.19.

5.1.2.1.1 Consequences

Suppose that Alice chooses a pure strategy $I \subseteq \{0, \dots, n-1\}$, Eve chooses a pure strategy $E_s \subseteq \{0, 1\}^n$, and Nature chooses a binary sequence \mathbf{y} and a steganographic state z . Then, Eve wins 1 if she classifies \mathbf{y} correctly, i.e., either she says stego and Nature chose stego, or she says cover and Nature chose cover, and she loses 1 if her classification is wrong. The game is zero-sum so that Alice's payoff is the negative of Eve's payoff. Table 5.7 (on p. 131) formalizes the possible outcomes as a zero-sum payoff matrix.

5.1.2.1.2 Randomization

In the fully randomized game, we have distributions on binary sequences and steganographic states. We also have randomization in the players' strategies. To describe the nature of the randomness, we start by defining two random variables on our event space Ω . Let $Y : \Omega \rightarrow \{0, 1\}^n$ be the random variable which takes an event to its binary sequence and let $S : \Omega \rightarrow \{c, s\}$ be the random variable which takes an event to its steganographic state. We proceed through the rest of this section by first describing the structure of the distribution on Ω ; next describing the two players' mixed strategies; and finally, by giving the players' payoffs as a consequence of their mixed strategies.

5.1.2.1.3 Steganographic States

Our results describing equilibria for this model carry through with arbitrary prior probabilities; so we replace the equal prior assumption with the notations p_s and p_c in several subsequent formulas. Note however, that with highly unequal priors, the game may trivialize because the prior probabilities can dominate other incentives. For this reason, we do require equal priors for some structural theorems; and we also use equal priors in our numerical illustrations. The event $S = s$ happens when Nature chooses the steganographic state to be stego; and this event occurs with probability p_s . We also define $\Pr_\Omega[S = c] := p_c = 1 - p_s$. From Eve's perspective, p_s is the prior probability that she observes a stego sequence on the communication channel.

5.1.2.1.4 Binary Sequences

The distribution on binary sequences depends on the value of the steganographic state. If $S = c$, then the steganographic state is cover, and Y is distributed according to the *cover distribution* \mathcal{P}_0 ; if $S = s$, then the steganographic state is stego, and Y is distributed according to a *stego distribution* \mathcal{P}_1 .

With this notation in hand, we may define, for any event $(\mathbf{Y} = \mathbf{y}, S = z)$:

$$\begin{aligned} \Pr_\Omega[(\mathbf{y}, z)] &= \Pr_\Omega[S = z] \cdot \Pr_\Omega[\mathbf{Y} = \mathbf{y} | S = z] \\ &= \begin{cases} p_c \cdot \mathcal{P}_0[\mathbf{Y} = \mathbf{y}] & \text{if } z = c \\ p_s \cdot \mathcal{P}_1[\mathbf{Y} = \mathbf{y}] & \text{if } z = s \end{cases}. \end{aligned} \quad (5.43)$$

We will define the distribution \mathcal{P}_1 after describing the players' mixed strategies.

5.1.2.1.5 Players' Mixed Strategies

In a mixed strategy, Alice can probabilistically embed into any given subset of positions, by choosing a probability distribution over size- k subsets of $\{0, \dots, n-1\}$. To describe a mixed strategy, for each $I \subseteq \{0, \dots, n-1\}$, we let \bar{a}_I denote the probability that Alice embeds into each of the positions in I .

A mixed strategy for Eve is a probability distribution over subsets of $\{0, 1\}^n$. Suppose that Eve's mixed strategy assigns probability e_T to each subset $T \subseteq \{0, 1\}^n$. Overloading notation slightly, we define $e : \{0, 1\}^n \rightarrow [0, 1]$ via

$$e(\mathbf{y}) = \sum_{T \subseteq \{0, 1\}^n : \mathbf{y} \in T} e_T . \quad (5.44)$$

Each $e(\mathbf{y})$ gives the total probability for the binary sequence \mathbf{y} that Eve classifies the sequence as stego object. Note that this “projected” representation of Eve's mixed strategy given in Equation (5.44) requires specifying 2^n real numbers, whereas the canonical representation of her mixed strategy using the notation e_T would require specifying 2^{2^n} real numbers. For this reason, we prefer to use the projection representation. Fortunately, the projected representation contains enough information to determine both players' payoffs; and the mapping from the canonical representation to the projected representation is surjective¹⁵ so that we may express results using the simpler representation without loss of generality.

5.1.2.2 Embedding Impact

The stego distribution \mathcal{P}_1 depends on Alice's embedding strategy. Let $I \subseteq \{0, \dots, n-1\}$, and for each $\mathbf{y} \in \{0, 1\}^n$ let \mathbf{y}_I denote the binary sequence obtained from \mathbf{y} by flipping the bits at all the positions in I . The stego distribution is obtained from the cover distribution by adjusting the likelihood that each \mathbf{y} occurs, assuming that for each I , with probability \bar{a}_I Alice flips the bits of \mathbf{y} in all the positions in I .

More formally, suppose that Alice embeds into each subset $I \subseteq \{0, \dots, n-1\}$ with probability \bar{a}_I . We then have

$$\begin{aligned} \mathcal{P}_1[\mathbf{Y} = \mathbf{y}] &= \sum_I \bar{a}_I \mathcal{P}_0[\mathbf{Y} = \mathbf{y}_I] \\ &= \sum_I \bar{a}_I \cdot \prod_{i \notin I} \mathcal{P}_0[Y_i = y_i] \cdot \prod_{i \in I} \mathcal{P}_0[Y_i = 1 - y_i] \\ &= \sum_I \bar{a}_I \cdot \prod_{i \notin I} \left(1 - f(i) + 2y_i f(i)\right) \cdot \prod_{i \in I} \left(f(i) - 2y_i f(i)\right). \end{aligned} \quad (5.45)$$

5.1.2.3 Payoff

In the full game, the expected payoff for Eve can be written as:

$$\begin{aligned} u(\text{Eve}) &= \Pr_{\Omega}[\mathbf{Y} \in E_s \text{ and } S = s] && \text{(true positive)} \\ &+ \Pr_{\Omega}[\mathbf{Y} \in E_c \text{ and } S = c] && \text{(true negative)} \\ &- \Pr_{\Omega}[\mathbf{Y} \in E_s \text{ and } S = c] && \text{(false positive)} \\ &- \Pr_{\Omega}[\mathbf{Y} \in E_c \text{ and } S = s] && \text{(false negative)} \end{aligned} \quad (5.46)$$

¹⁵The proof of surjectivity follows directly from using induction on n .

and this can be further computed as:

$$\begin{aligned}
u(\text{Eve}) &= p_s \mathcal{P}_1[\mathbf{Y} \in E_s] + p_c \mathcal{P}_0[\mathbf{Y} \in E_c] - p_c \mathcal{P}_0[\mathbf{Y} \in E_s] - p_s \mathcal{P}_1[\mathbf{Y} \in E_c] \\
&= \sum_{\mathbf{y} \in \{0,1\}^n} \left[e(\mathbf{y}) p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] \right. \\
&\quad \left. + (1 - e(\mathbf{y})) p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}] \right. \\
&\quad \left. - (1 - e(\mathbf{y})) p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] \right. \\
&\quad \left. - e(\mathbf{y}) p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}] \right] \\
&= \sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1) \\
&\quad \cdot (p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] - p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}]). \tag{5.47}
\end{aligned}$$

The terms $\mathcal{P}_0[\mathbf{Y} = \mathbf{y}]$ and $\mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}]$ are defined in Equations (5.4) (p. 69) and (5.45), respectively. Note that we write $\mathcal{P}_1 = \mathcal{P}_1^{(\bar{a})}$ to clarify that the distribution \mathcal{P}_1 depends on Alice's mixed strategy. In summary, Eve's payoff is the probability that her classifier is correct minus the probability that it is incorrect; and the game is zero-sum so that Alice's payoff is exactly the negative of Eve's payoff.

5.1.2.4 Solving the Game

In this section, we present our analytical results. We begin by describing best response strategies for each player. Next, we describe in formal notation the minmax strategies for each player. Finally, we present several theorems which give structural constraints on the game's Nash equilibria.

To compute best responses for Alice and Eve, we assume that the other player is playing a fixed strategy, and determine the strategy for Alice (or Eve) which minimizes (or maximizes) the payoff in Equation (5.47), as appropriate.

5.1.2.4.1 Alice's Best Response

Given a fixed strategy e for Eve, Alice's goal is to minimize the payoff in Equation (5.47). However, since she has no control over the cover distribution \mathcal{P}_0 , this goal can be simplified to that of minimizing

$$\begin{aligned}
&\sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1) \cdot p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] \\
&= p_s \sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1) \cdot \sum_{I \subseteq \{0, \dots, n-1\}} \bar{a}_I \mathcal{P}_0[\mathbf{Y} = \mathbf{y}_I] \\
&= p_s \sum_{I \subseteq \{0, \dots, n-1\}} \bar{a}_I \sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1) \cdot \mathcal{P}_0[\mathbf{Y} = \mathbf{y}_I] .
\end{aligned}$$

This formula is linear in Alice's choice variables, so she can minimize its value by putting all her probability on the sum's least element. A best response for Alice is thus to play a pure strategy I that minimizes

$$\sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1) \cdot \mathcal{P}_0[\mathbf{Y} = \mathbf{y}]. \quad (5.48)$$

Of course, several different I might simultaneously minimize this sum. In this case, Alice's best response strategy space may also include a mixed strategy that distributes her embedding probabilities randomly among such I .

5.1.2.4.2 Eve's Best Response

Given a fixed strategy for Alice, Eve's goal is to maximize her payoff as given in Equation (5.47). So, for each \mathbf{y} , she should choose $e(\mathbf{y})$ to maximize the term of the sum corresponding to \mathbf{y} . Specifically, if $p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] - p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}] > 0$, then the best choice is $e(\mathbf{y}) = 1$; and if the strict inequality is reversed, then the best choice is $e(\mathbf{y}) = 0$. If the inequality is an equality, then Eve may choose any value for $e(\mathbf{y}) \in [0, 1]$ and still be playing a best response.

Formally, her optimal decision rule is the following LRT:

$$e(\mathbf{y}) = \begin{cases} 1 & \text{if } \frac{\Pr_{\Omega}[S=s|\mathbf{Y}=\mathbf{y}]}{\Pr_{\Omega}[S=c|\mathbf{Y}=\mathbf{y}]} > 1, \\ 0 & \text{if } \frac{\Pr_{\Omega}[S=s|\mathbf{Y}=\mathbf{y}]}{\Pr_{\Omega}[S=c|\mathbf{Y}=\mathbf{y}]} < 1, \\ \text{any } p \in [0, 1] & \text{if } \frac{\Pr_{\Omega}[S=s|\mathbf{Y}=\mathbf{y}]}{\Pr_{\Omega}[S=c|\mathbf{Y}=\mathbf{y}]} = 1. \end{cases} \quad (5.49)$$

For a fixed sequence \mathbf{y} , the condition for classifying \mathbf{y} as stego can be rewritten as:

$$\begin{aligned} 1 &< \frac{\Pr_{\Omega}[S = s|\mathbf{Y} = \mathbf{y}]}{\Pr_{\Omega}[S = c|\mathbf{Y} = \mathbf{y}]} \\ &= \frac{\Pr_{\Omega}[\mathbf{Y} = \mathbf{y}]}{\Pr_{\Omega}[\mathbf{Y} = \mathbf{y}]} \cdot \frac{\Pr_{\Omega}[S = s|\mathbf{Y} = \mathbf{y}]}{\Pr_{\Omega}[S = c|\mathbf{Y} = \mathbf{y}]} \\ &= \frac{\Pr_{\Omega}[S = s]}{\Pr_{\Omega}[S = c]} \cdot \frac{\Pr_{\Omega}[\mathbf{Y} = \mathbf{y}|S = s]}{\Pr_{\Omega}[\mathbf{Y} = \mathbf{y}|S = c]} \\ &= \frac{p_s \mathcal{P}_1[\mathbf{Y} = \mathbf{y}]}{p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}]} \\ &= \frac{p_s \sum_I \bar{a}_I \cdot \prod_{i \notin I} (1 - f(i) + 2y_i \tilde{f}(i)) \cdot \prod_{i \in I} (f(i) - 2y_i \tilde{f}(i))}{p_c \prod_{i=0}^{n-1} (1 - f(i) + 2y_i \tilde{f}(i))} \\ &= \frac{p_s}{p_c} \sum_I \bar{a}_I \prod_{i \in I} \left(\frac{f(i) - 2y_i \tilde{f}(i)}{1 - f(i) + 2y_i \tilde{f}(i)} \right) \\ &= \frac{p_s}{p_c} \sum_I \bar{a}_I \prod_{i \in I} \left(\frac{f(i)}{1 - f(i)} - y_i \frac{2\tilde{f}(i)}{f(i)(1 - f(i))} \right). \end{aligned} \quad (5.50)$$

Note that Eve's decision rule is written as a multilinear polynomial inequality of degree at most k in the binary sequence \mathbf{y} , and that the number of terms in the formula is $\binom{n}{k}$. When k is a constant relative to n (as it typically is in practical applications), then $\binom{n}{k}$ is polynomial in n , and Eve's optimal decision rule can be applied for each binary sequence in time that is polynomial in the length of the sequence.

5.1.2.4.3 Maxmin and Minmax Strategies

As Eve wants to maximize her payoff and Alice wants to minimize, we describe her maxmin and minmax strategy, respectively.

Eve's maxmin strategy is given by

$$\operatorname{argmax}_e \left(\min_I \left(\sum_{\mathbf{y} \in \{0,1\}^n} (2e(\mathbf{y}) - 1)(p_s \mathcal{P}_0[\mathbf{Y} = \mathbf{y}_I] - p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}]) \right) \right); \quad (5.51)$$

while Alice's minmax strategy is given by

$$\operatorname{argmin}_{\bar{a}} \left(\max_{E_s} \left(\sum_{\mathbf{y} \in E_s} (p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}] - p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}]) \right) + \sum_{\mathbf{y} \in E_c} (p_c \mathcal{P}_0[\mathbf{Y} = \mathbf{y}] - p_s \mathcal{P}_1^{(\bar{a})}[\mathbf{Y} = \mathbf{y}]) \right). \quad (5.52)$$

Each maxmin or minmax strategy can be determined (recursively) as the solution to a linear program involving the payoff matrix for Alice's and Eve's pure strategies. Unfortunately, Eve's pure strategy space has size 2^{2^n} so it is computationally intractable to find the maxmin strategies using this method even for $n = 5$.

5.1.2.4.4 Nash Equilibria

In this section, we present structural constraints for Nash equilibria. We begin with a lemma showing that Eve's classifier in a specific type of equilibrium must respect the canonical partial ordering on binary sequences. We conclude the section with a conjecture about Alice's equilibrium strategy.

Lemma 5.9. *Define a partial ordering on $\{0, 1\}^n$ by $\mathbf{y} < \mathbf{z}$ iff $y_i \leq z_i$ for $i = 0, \dots, n-1$ and $y_i < z_i$ for at least one i . Then whenever Alice's embedding strategy satisfies the constraint $\frac{p_s}{p_c} \sum_I \bar{a}_I \prod_{i \in I} \left(\frac{f(i)}{1-f(i)} - y_i \frac{2\bar{f}(i)}{f(i)(1-f(i))} \right) \neq 1$ for the sequence \mathbf{y} , the following condition holds:*

- ▶ If Eve classifies \mathbf{y} as stego and $\mathbf{u} < \mathbf{y}$, then Eve classifies \mathbf{u} as stego too.
- ▶ If Eve classifies \mathbf{y} as cover and $\mathbf{y} < \mathbf{z}$, then Eve classifies \mathbf{z} as cover too.

Proof. Suppose Eve classifies \mathbf{y} as stego. Then from the conditions on Eve's best response (Equations (5.49) and (5.50)), we have that $\frac{p_s}{p_c} \sum_I \bar{a}_I \prod_{i \in I} \left(\frac{f(i)}{1-f(i)} - y_i \frac{2\bar{f}(i)}{f(i)(1-f(i))} \right) \geq 1$;

and by the hypothesis of the lemma, the inequality is strict. Suppose $\mathbf{u} < \mathbf{y}$. Then the value of $\frac{p_s}{p_c} \sum_I \bar{a}_I \prod_{i \in I} \left(\frac{f(i)}{1-f(i)} - z_i \frac{2\tilde{f}(i)}{f(i)(1-f(i))} \right)$ is at least the value of the same expression with \mathbf{y} replacing \mathbf{u} . So, this value is also greater than 1, and Eve also classifies \mathbf{u} as stego. The proof of the reverse direction is analogous. \square

This lemma implies that in any Nash equilibrium, the set of all binary sequences can be divided into three disjoint sets, *low sequences* which Eve’s likelihood test proscribes a clear value of stego, *high sequences* which Eve’s test proscribes as clearly cover, and a small set of *boundary sequences* on which Eve’s behavior is not obviously constrained. Furthermore, changing 0s to 1s in a clearly-cover sequence keeps it cover, and changing 1s to 0s in a clearly-stego sequence keeps it stego.

Next, we state a conjecture about Alice’s strategy in an equilibrium.

Conjecture 5.1. *Assume equal priors, so that $p_s = p_c = \frac{1}{2}$ and a reasonable $f(i)$. In a Nash equilibrium, Alice uses every $i \in \{0, \dots, n-1\}$ with non-zero probability.*

For homogeneous $f(i)$ there are simple counter-examples to the conjecture; however, it is important to note that for homogeneous $f(i)$ the definition of adaptive embedding itself is not sensible.

Here, we frame a proof outline for this conjecture. Assume a Nash equilibrium with \bar{a} and \bar{e} as the strategies of Alice and Eve, respectively. To obtain a contradiction, suppose that $i \in \{0, \dots, n-1\}$ is such that $\bar{a}_I = 0$ for every I containing i . If \mathbf{y} is any sequence that Eve’s optimal decision rule classifies as either clearly cover or clearly stego, then Eve’s behavior does not depend on the value of \mathbf{y} at position i . However, if there are “indifferent” sequences \mathbf{z} that Eve’s likelihood test proscribes as cover or stego with equal probability, we cannot rule out that Eve may take the position i into account for \mathbf{z} . This remains true even though her likelihood test does not proscribe an outcome based on i , and even though she is playing a best response to Alice who is not using position i . Our avenue to proceed is to demonstrate a violation of the equilibrium condition by showing how Alice can increase her payoff by using position i . Toward this end, we can show that, by shifting her embedding probability to sets containing i from sets not containing i , Alice will increase Eve’s misclassification probability for sequences that are not on her “indifference boundary”. However, it is possible that Eve gains enough advantage from conditioning on i when the special boundary sequences occur to offset this disadvantage. It seems to us that this possibility hinges on structural properties of the sequence $f(i)$.

In the following section, we explicitly compute all equilibria in the case of length-two sequences and a message length of $k = 1$. Note that in this case, Conjecture 5.1 holds.

5.1.2.5 Solution and Numerical Illustration for $n = 2$ and $k = 1$

In this section, we instantiate our model with the special case of flipping a single bit ($k = 1$) in sequences of length two ($n = 2$). In this setting, Alice’s pure strategy space is $\{\{0\}, \{1\}\}$; and since $\bar{a}_{\{1\}} = 1 - \bar{a}_{\{0\}}$, her mixed strategy space can be represented by a single value $\bar{a}_0 = \bar{a}_{\{0\}} \in [0, 1]$. Eve’s pure strategy space is represented by the set of

all $[0, 1]$ -valued functions on $\{00, 01, 10, 11\}$. Throughout this section we assume equal priors, i.e., $p_c = p_s = \frac{1}{2}$.

5.1.2.5.1 Alice's Minmax Strategy

To compute Alice's minmax strategy, we first divide Alice's strategy space into three regions based on Eve's best response:

Lemma 5.10. *The following table gives Eve's best response for each sequence \mathbf{y} as a function of \bar{a}_0 .*

Alice's strategy		Eve's best response			
		$\mathbf{y} =$			
		00	01	10	11
$\bar{a}_0 < \tau_1$		<i>s</i>	<i>c</i>	<i>s</i>	<i>c</i>
	$\tau_1 < \bar{a}_0 < \tau_2$	<i>s</i>	<i>s</i>	<i>s</i>	<i>c</i>
	$\tau_2 < \bar{a}_0$	<i>s</i>	<i>s</i>	<i>c</i>	<i>c</i>

where $\tau_1 = \frac{(1-f(0))2\tilde{f}(1)}{f(0)+f(1)-1}$ and $\tau_2 = \frac{f(0)2\tilde{f}(1)}{f(0)+f(1)-1}$.

Proof. We prove Eve's optimal decision for the four realizations separately.

00: Eve always classifies 00 as stego.

$$\begin{aligned} \mathcal{P}_0[\mathbf{Y} = 00] &= \\ (1-f(0))(1-f(1)) &< \bar{a}_0 f(0)(1-f(1)) + (1-\bar{a}_0)(1-f(0))f(1) \\ &= \mathcal{P}_1(\bar{a}_0)[\mathbf{Y} = 00], \end{aligned}$$

since $(1-f(0))(1-f(1)) < f(0)(1-f(1))$ and $(1-f(0))(1-f(1)) < (1-f(0))f(1)$.

01: Eve classifies 01 as cover when $\bar{a}_0 < \frac{(1-f(0))2\tilde{f}(1)}{f(0)+f(1)-1} := \tau_1$.

$$\begin{aligned} \mathcal{P}_0[\mathbf{Y} = 01] &= \\ (1-f(0))f(1) &\stackrel{!}{>} \bar{a}_0 f(0)f(1) + (1-\bar{a}_0)(1-f(0))(1-f(1)) \\ &= \mathcal{P}_1(\bar{a}_0)[\mathbf{Y} = 01] && \Leftrightarrow \\ (1-f(0))(f(1)-1+f(1)) &> \bar{a}_0(f(0)f(1)-1+f(0)+f(1)-f(0)f(1)) && \Leftrightarrow \\ \frac{(1-f(0))2\tilde{f}(1)}{f(0)+f(1)-1} &> \bar{a}_0 \end{aligned}$$

10: Eve classifies 10 as cover when $\bar{a}_0 > \frac{f(0)2\tilde{f}(1)}{f(0)+f(1)-1} := \tau_2$.

$$\begin{aligned}
 \mathcal{P}_0[\mathbf{Y} = 10] &= \\
 f(0)(1-f(1)) &\stackrel{!}{>} \bar{a}_0(1-f(0))(1-f(1)) + (1-\bar{a}_0)f(0)f(1) \\
 &= \mathcal{P}_1(\bar{a}_0)[\mathbf{Y} = 10] && \Leftrightarrow \\
 f(0)(1-f(1)) - f(0)f(1) &> \bar{a}_0(1-f(0)-f(1) + f(0)f(1) - f(0)f(1)) && \Leftrightarrow \\
 \frac{-f(0)2\tilde{f}(1)}{1-f(0)-f(1)} &< \bar{a}_0
 \end{aligned}$$

11: Eve always classifies 11 as cover.

$$\begin{aligned}
 \mathcal{P}_0[\mathbf{Y} = 00] &= \\
 f(0)f(1) &> \bar{a}_0(1-f(0))f(1) + (1-\bar{a}_0)f(0)(1-f(1)) \\
 &= \mathcal{P}_1(\bar{a}_0)[\mathbf{Y} = 00],
 \end{aligned}$$

since $f(0)f(1) > (1-f(0))f(1)$ and $f(0)f(1) > f(0)(1-f(1))$.

Finally, $\tau_1 < \tau_2$ always holds, since $(1-f(0)) < f(0)$. \square

Theorem 5.4. *The strategy $(\tau_2, 1 - \tau_2)$ is a minmax strategy for Alice.*

Proof. First, for each region, we compute the derivative of Alice's payoff as a function of \bar{a}_0 given that Eve always uses her best response. Then, we have that Alice's payoff is

- ▶ strictly increasing when $\bar{a}_0 < \tau_1$,
- ▶ strictly decreasing when $\bar{a}_0 > \tau_2$,
- ▶ and, when $\tau_1 \leq \bar{a}_0 \leq \tau_2$, it is strictly increasing if $f(0) \neq f(1)$, and it is constant if $f(0) = f(1)$.

Thus, we have that $\bar{a}_0 = \tau_2$ always attains the maximum. \square

Note that embedding uniformly into both positions ($\bar{a}_0 = \frac{1}{2}$) is optimal only if the biases are uniform ($\tilde{f}(0) = \tilde{f}(1)$); and embedding only in the first position would be optimal only if the bias of the first position were zero ($\tilde{f}(0) = 0$) or if the bias of the second position were one half ($\tilde{f}(1) = 1/2$).

Figure 5.5 depicts Eve's error rates and the resulting overall misclassification rate as a function of Alice's strategy $(\bar{a}_0, 1 - \bar{a}_0)$. Figure 5.5(a) shows a homogeneous f , while Figure 5.5(b) shows a heterogeneous f . It can be seen that neither the false positive rate (dashed line) nor the false negative rate (dotted line) is continuous and that the discontinuities occur at the points τ_1 and τ_2 , the points where Eve changes her optimal decision rule. Nonetheless, the overall misclassification rate (solid line) is continuous, which leads to the conclusion that this rate leverages out the discontinuities and thus is a good measure of the overall accuracy of Eve's detector.

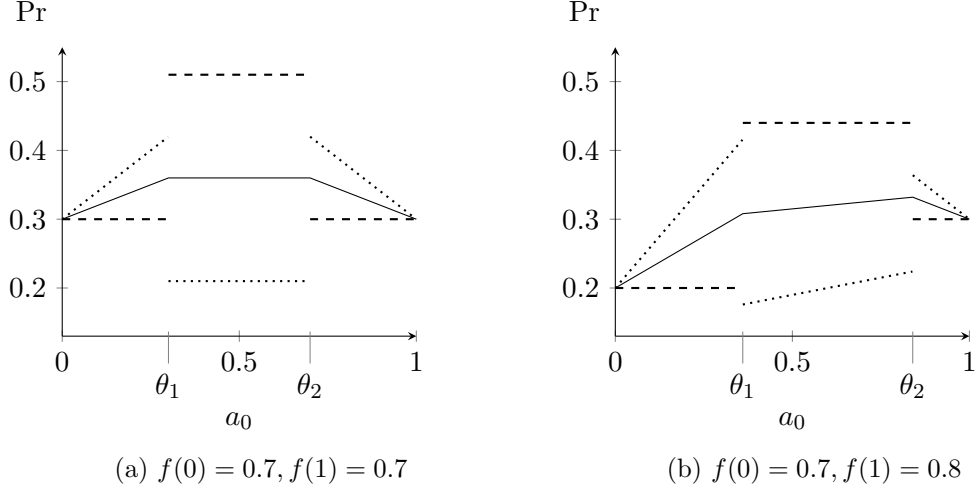


Figure 5.5: Eve's false positive rate (dashed line), false negative rate (dotted line) and her overall misclassification rate (solid line) as a function of a_0 , assuming that Eve plays a best response to Alice.

5.1.2.5.2 Eve's Maxmin Strategy

Theorem 5.5. *Eve's maxmin strategy e_{maxmin} is $e_{maxmin}(00) = e_{maxmin}(01) = 1$, $e_{maxmin}(11) = 0$, and*

$$e_{maxmin}(10) = p = \frac{2\tilde{f}(0)}{f(0) + f(1) - 1}. \quad (5.53)$$

Proof. Since the game is zero sum, Eve's strategy is a maxmin strategy if Alice's minmax strategy is a best response to it [87]. Therefore, it suffices to show that Alice has no incentives for deviating from her own minmax strategy when Eve uses e_{maxmin} . Alice's best response to e_{maxmin} is

$$\begin{aligned} & \operatorname{argmax}_{\bar{a}_0 \in [0,1]} \left\{ -\mathcal{P}_1(\bar{a}_0) [\mathbf{Y} = 00] - \mathcal{P}_1(\bar{a}_0) [\mathbf{Y} = 01] \right. \\ & \quad \left. + (1 - 2p)\mathcal{P}_1(\bar{a}_0) [\mathbf{Y} = 10] + \mathcal{P}_1(\bar{a}_0) [\mathbf{Y} = 11] \right\} \\ & = \operatorname{argmax}_{\bar{a}_0 \in [0,1]} \left\{ -\bar{a}_0 f(0)(1 - f(1)) - (1 - \bar{a}_0)(1 - f(0))f(1) \right. \\ & \quad - \bar{a}_0 f(0)f(1) - (1 - \bar{a}_0)(1 - f(0))(1 - f(1)) \\ & \quad + (1 - 2p)[\bar{a}_0(1 - f(0))(1 - f(1)) + (1 - \bar{a}_0)f(0)f(1)] \\ & \quad \left. + \bar{a}_0(1 - f(0))f(1) + (1 - \bar{a}_0)f(0)(1 - f(1)) \right\} \end{aligned}$$

$$= \operatorname{argmax}_{\bar{a}_0 \in [0,1]} \left\{ \bar{a}_0 [2 - 4f(0) - 2p(1 - f(0) - f(1))] + \operatorname{const}(f, p) \right\}.$$

If $p = \frac{2\tilde{f}(0)}{f(0)+f(1)-1}$, then the value of the above optimization problem does not depend on \bar{a}_0 . Consequently, Alice has no incentives for deviating from her minmax strategy. \square

5.1.3 Powerful Steganalyst and Independent Embedding

This instantiation originates from joint work with Aron Laszka, Benjamin Johnson, and Jens Grossklags. To maintain consistency with the original publication, we use a similar notation as in the previous section.

In this section we replace the embedding rate of exactly k bit for Alice by a message of expected length of k and thus independent embedding. We derive analytical results on the minmax strategies and the existence of pure-strategy Nash equilibria.

Let S denote the random variable taking values in $\{c, s\}$ that represents whether the sequence was drawn from the cover or stego distribution, respectively.

5.1.3.1 Strategies

Alice wants to hide messages of expected length k into a cover object of length- n ; so she may choose any set of n probabilities that sum to k . Each \bar{a}_i represents the probability that Alice changes the value of the sequence at position i .

As above, Eve wants to optimally classify sequences as either stego cover object by performing a LRT; so she may choose a probability for each length- n binary sequence. Each $e(\mathbf{y})$ represents the probability that Eve classifies the sequence \mathbf{y} as stego.

5.1.3.2 Embedding Impact

In the stego distribution, Alice flips the value of the sequence in each position i with probability \bar{a}_i , so that $Y_i = 1$ with probability $f(i)(1 - \bar{a}_i) + (1 - f(i))\bar{a}_i$. Since positions in the stego distribution are also independent,

$$\mathcal{P}_1 = \Pr[\mathbf{Y} = \mathbf{y} | S = s] = \prod_{i:y_i=1} (f(i) - 2\bar{a}_i\tilde{f}(i)) \cdot \prod_{i:y_i=0} (1 - f(i) + 2\bar{a}_i\tilde{f}(i)). \quad (5.54)$$

5.1.3.3 Payoff

To formalize the game payoff, we assume equal priors over cover and stego objects. The game payoff is then determined by the probability over all 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 1.¹⁶

¹⁶Note that this is not a zero-sum game, but a constant-sum game. In game theory the strategies are insensitive to any positive affine transformation of the payoffs [64], so all the definitions for zero-sum games also hold for constant-sum games.

5.1.3.4 Solving the Game

5.1.3.4.1 Eve's Best Response

Given a fixed embedding strategy for Alice, Eve must classify each sequence as cover or stego object. Since she knows both of these distributions, she can perform a likelihood ratio test to determine her optimal decision [27]. This test gives a deterministic decision rule whenever the two likelihoods are unequal. When they are equal for a given sequence \mathbf{y} , Eve's decision at that \mathbf{y} does not affect the probability that her classifier is correct; so in this case, any randomized function over cover and stego objects is a best response.

Theorem 5.6. *A best response strategy of Eve is given by the decision rule*

$$DR(\mathbf{y}) = \begin{cases} c & \text{if } \sum_i w_i y_i > \Upsilon \\ s & \text{if } \sum_i w_i y_i < \Upsilon \\ c \text{ or } s & \text{if } \sum_i w_i y_i = \Upsilon \end{cases} \quad (5.55)$$

where for each i ,

$$w_i = \log \frac{f(i)(1 - f(i) + 2\bar{a}_i \tilde{f}(i))}{(f(i) - 2\bar{a}_i \tilde{f}(i))(1 - f(i))} \quad \text{and} \quad (5.56)$$

$$\Upsilon = \sum_i \log \frac{1 - f(i) + 2\bar{a}_i \tilde{f}(i)}{1 - f(i)}. \quad (5.57)$$

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

$$DR(\mathbf{y}) = \begin{cases} c & \text{if } \frac{\Pr[S=c|\mathbf{Y}=\mathbf{y}]}{\Pr[S=s|\mathbf{Y}=\mathbf{y}]} > 1 \\ s & \text{if } \frac{\Pr[S=c|\mathbf{Y}=\mathbf{y}]}{\Pr[S=s|\mathbf{Y}=\mathbf{y}]} < 1 \\ c \text{ or } s & \text{if } \frac{\Pr[S=c|\mathbf{Y}=\mathbf{y}]}{\Pr[S=s|\mathbf{Y}=\mathbf{y}]} = 1. \end{cases}$$

The condition for cover can be expressed using f and \bar{a} as follows:

$$\begin{aligned} 1 &< \frac{\Pr[S = c | \mathbf{Y} = \mathbf{y}]}{\Pr[S = s | \mathbf{Y} = \mathbf{y}]} \\ &= \frac{\Pr[S = c] \Pr[\mathbf{Y} = \mathbf{y} | S = c]}{\Pr[S = s] \Pr[\mathbf{Y} = \mathbf{y} | S = s]} \\ &= \frac{\frac{1}{2} \cdot \prod_i \Pr[Y_i = y_i | S = c]}{\frac{1}{2} \cdot \prod_i \Pr[Y_i = y_i | S = s]} \\ &= \prod_{i:y_i=1} \frac{f(i)}{f(i) - 2\bar{a}_i \tilde{f}(i)} \cdot \prod_{i:y_i=0} \frac{1 - f(i)}{1 - f(i) + 2\bar{a}_i \tilde{f}(i)} \\ 0 &< \sum_{i:y_i=1} \log \frac{f(i)}{f(i) - 2\bar{a}_i \tilde{f}(i)} + \sum_{i:y_i=0} \log \frac{1 - f(i)}{1 - f(i) + 2\bar{a}_i \tilde{f}(i)} \end{aligned}$$

$$\begin{aligned}
 &= \sum_i \left(y_i \log \frac{f(i)}{f(i) - 2\bar{a}_i \tilde{f}(i)} + (1 - y_i) \log \frac{1 - f(i)}{1 - f(i) + 2\bar{a}_i \tilde{f}(i)} \right) \\
 &= \sum_i y_i \log \frac{f(i)(1 - f(i) + 2\bar{a}_i \tilde{f}(i))}{(f(i) - 2\bar{a}_i \tilde{f}(i))(1 - f(i))} + \sum_i \log \frac{1 - f(i)}{1 - f(i) + 2\bar{a}_i \tilde{f}(i)} \\
 &\Leftrightarrow \sum_i y_i \log \frac{f(i)(1 - f(i) + 2\bar{a}_i \tilde{f}(i))}{(f(i) - 2\bar{a}_i \tilde{f}(i))(1 - f(i))} \geq \sum_i \log \frac{1 - f(i) + 2\bar{a}_i \tilde{f}(i)}{1 - f(i)}.
 \end{aligned}$$

□

5.1.3.4.2 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 $e(\mathbf{y})$ is Eve's probability for classifying \mathbf{y} as stego, then Alice's best response strategy is to choose an \bar{a} satisfying $\sum_i \bar{a}_i = k$ and maximizing

$$\begin{aligned}
 &\sum_{\mathbf{y} \in \{0,1\}^n} (1 - e(\mathbf{y})) \Pr[\mathbf{Y} = \mathbf{y} | S = s] = \\
 &\sum_{\mathbf{y} \in \{0,1\}^n} (1 - e(\mathbf{y})) \prod_{i: y_i=1} (f(i) - 2\bar{a}_i \tilde{f}(i)) \prod_{i: y_i=0} (1 - f(i) + 2\bar{a}_i \tilde{f}(i)). \tag{5.58}
 \end{aligned}$$

To get some leverage from this formula, consider Alice's best response strategy and any pair \bar{a}_i, \bar{a}_j that are interior values of $(0, 1)$. Alice's payoff cannot increase if she adjusts her strategy by simultaneously increasing \bar{a}_i and decreasing \bar{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 \bar{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 (\bar{a}_i, \bar{a}_j) taking interior values in $(0, 1)$ in terms of the remaining \bar{a}_m .

$$\begin{aligned}
 \bar{a}_i - \bar{a}_j = & \frac{\sum_{\mathbf{y}} (1 - e(\mathbf{y})) \prod_{m \neq i, j} \Pr[Y_m = y_m | S = s] \cdot \left((1 - 2y_i) 2\tilde{f}(i)(2y_j \tilde{f}(j) + 1 - f(j)) \right. \\
 & \left. - (1 - 2y_j) 2\tilde{f}(j)(2y_i \tilde{f}(i) + 1 - f(i)) \right)}{\sum_{\mathbf{y}} (1 - e(\mathbf{y})) \prod_{m \neq i, j} \Pr[Y_m = y_m | S = s] \cdot \left(4\tilde{f}(0)\tilde{f}(1)(1 - 2y_i)(1 - 2y_j) \right)} \tag{5.59}
 \end{aligned}$$

This set of constraints can be solved for at least some small n . We illustrate the structure of the solution in the following subsection by considering the special case of two positions.

5.1.3.5 Solution and Numerical Illustration for $n = 2$ and $k = 1$

For this subsection, we restrict our analysis to the case of changing a single bit ($k = 1$) in covers of length two ($n = 2$). Again, Alice's strategy is fully specified by the probability \bar{a}_0 . Eve's strategy is specified as a vector $(e(00), e(01), e(10), e(11))$.

5.1.3.5.1 Alice's Minmax Strategy

Alice's minmax 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 can be found in Appendix C.

Lemma 5.11. *Eve always classifies sequence 00 as stego and sequence 11 as cover.*

Lemma 5.12. *Eve classifies the sequence 01 as cover when $\bar{a}_0 \leq \tau_1$, and she classifies the sequence 10 as cover when $\bar{a}_0 \geq \tau_2$, where*

$$\tau_1 = \frac{(f(0) - 1)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1)}{4\tilde{f}(0)\tilde{f}(1)} + \frac{\sqrt{\left[(1 - f(0))\tilde{f}(1) + \tilde{f}(0)(1 - f(1))\right]^2 - 8\tilde{f}(0)\tilde{f}(1)^2(f(0) - 1)}}{4\tilde{f}(0)\tilde{f}(1)}$$

and

$$\tau_2 = \frac{f(0)\tilde{f}(1) + \tilde{f}(0)f(1) - \sqrt{\left[f(0)\tilde{f}(1) + \tilde{f}(0)f(1)\right]^2 - 8f(0)\tilde{f}(0)\tilde{f}(1)^2}}{4\tilde{f}(0)\tilde{f}(1)}.$$

Lemma 5.13. *It always holds that $\tau_1 < \tau_2$.*

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

Theorem 5.7. *Given a fixed strategy for Alice, Eve's optimal decision for each binary sequence \mathbf{y} is given by:*

Alice's strategy		Eve's best response			
		$\mathbf{y} =$			
		00	01	10	11
$\bar{a}_0 \leq$	τ_1	<i>s</i>	<i>c</i>	<i>s</i>	<i>c</i>
	$\tau_1 \leq \bar{a}_0 \leq$	<i>s</i>	<i>s</i>	<i>s</i>	<i>c</i>
	$\tau_2 \leq \bar{a}_0$	<i>s</i>	<i>s</i>	<i>c</i>	<i>c</i>

Proof. It follows immediately from Lemmas 5.11, 5.12, and 5.13. \square

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

Lemma 5.14. *Alice's payoff is increasing for $\bar{a}_0 \in [0, \tau_1]$ and decreasing for $\bar{a}_0 \in [\tau_2, 1]$.*

Lemma 5.15. *The first derivative of Alice's payoff for $\bar{a}_0 \in [\tau_1, \tau_2]$ is*

$$\left. \frac{\partial u(\text{Alice})}{\partial \bar{a}_0} \right|_{\tau_1 \leq \bar{a}_0 \leq \tau_2} = -16\bar{a}_0\tilde{f}(0)\tilde{f}(1) + 4\left(f(0)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1)\right)$$

and the second derivative is $-16\tilde{f}(0)\tilde{f}(1)$.

Theorem 5.8. *Alice's minimax strategy is*

$$\bar{a}_0 = \begin{cases} a_{\max} & \text{when } a_{\max} \leq \tau_2 \\ \tau_2 & \text{when } \tau_2 < a_{\max} \end{cases}, \quad (5.60)$$

where a_{\max} denotes $\frac{f(0)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1)}{4\tilde{f}(0)\tilde{f}(1)}$.

Proof. From Lemma 5.14, we have that Alice's minimax strategy satisfies $\tau_1 \leq a_1 \leq \tau_2$. This strategy must be a local maximum for her payoff over $[\tau_1, \tau_2]$. 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 \bar{a}_0 , which gives us a_{\max} .

- It can be shown that $a_{\max} \geq \tau_1$.
- If $a_{\max} \leq \tau_2$, the local maximum is attained at a_{\max} . Thus, Alice's minimax strategy is $(a_{\max}, 1 - a_{\max})$.
- If $\tau_2 < a_{\max}$, the local maximum is attained at the endpoint τ_2 . Thus, Alice's minimax strategy is $(\tau_2, 1 - \tau_2)$.

\square

5.1.3.5.2 Nash equilibria

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

Theorem 5.9. *A Nash equilibrium with a deterministic strategy for Eve exists if and only if $a_{\max} \leq \tau_2$.*

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 $a_{\max} \leq \tau_2$ as both strategies are best responses. Second, we have to show that no equilibrium with a deterministic strategy for Eve can exist if $\tau_2 < a_{\max}$:

- Alice's best response to the strategy (s, c, s, c) is $\bar{a}_0 = 1$; 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 $\bar{a}_0 = a_{\max}$; however, since $a_{\max} \notin [\tau_1, \tau_2]$, Eve's response is not (s, s, s, c) , but either (s, c, s, c) or (s, s, c, c) .
- ▶ Alice's best response to (s, s, c, c) is $\bar{a}_0 = 0$; but, Eve's 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 $a_{\max} \leq \tau_2$, 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 find a mixed strategy for Eve in the case of $a_{\max} > \tau_2$.

Theorem 5.10. *Eve's maxmin strategy is*

$$\begin{aligned}
 e(00) &= 1 \\
 e(01) &= 1 \\
 e(10) &= \begin{cases} 1 & \text{if } a_{\max} \leq \tau_2 \\ \frac{\tilde{f}(0)}{\sqrt{(f(0)-f(1))^2 - 4f(0)\tilde{f}(0)\tilde{f}(1)(f(1)-1)}} & \text{otherwise.} \end{cases} \\
 e(11) &= 0
 \end{aligned}$$

Proof. Eve is playing a maxmin strategy when she forces Alice to play her minmax 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 minmax strategy. □

Figure 5.6 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 5.7 shows Eve's classification error rates as a function of \bar{a}_0 for two different examples of f . The example f in Figure 5.7(a) yields a deterministic strategy equilibrium, while the f in Figure 5.7(b) yields a randomized strategy equilibrium. Both figures reveal that neither the false positive rate 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 τ_1 and τ_2 where Eve switches her optimal strategy (see Lemma 5.12).

For the practical steganalyst, these results give direction to the optimal detection of strategic embedding. In particular, Eve's optimal classifier should be monotone in the cover's predictability metric; and a deterministic classifier can be sub-optimal for covers with heterogeneous predictability.

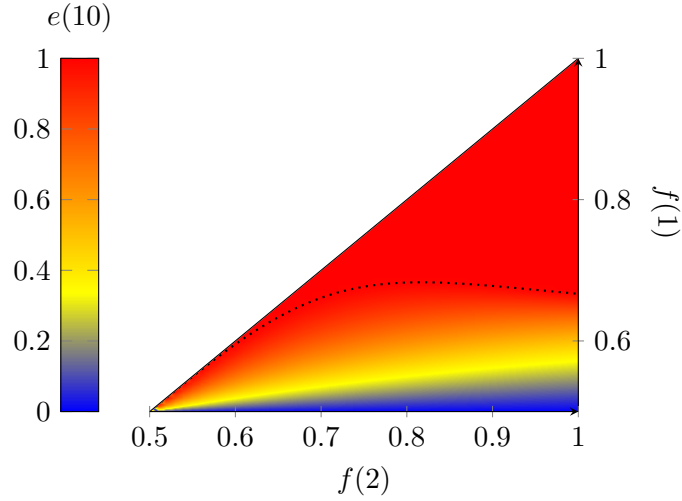


Figure 5.6: The value of $e(10)$ in Eve's minimax strategy as function of f .

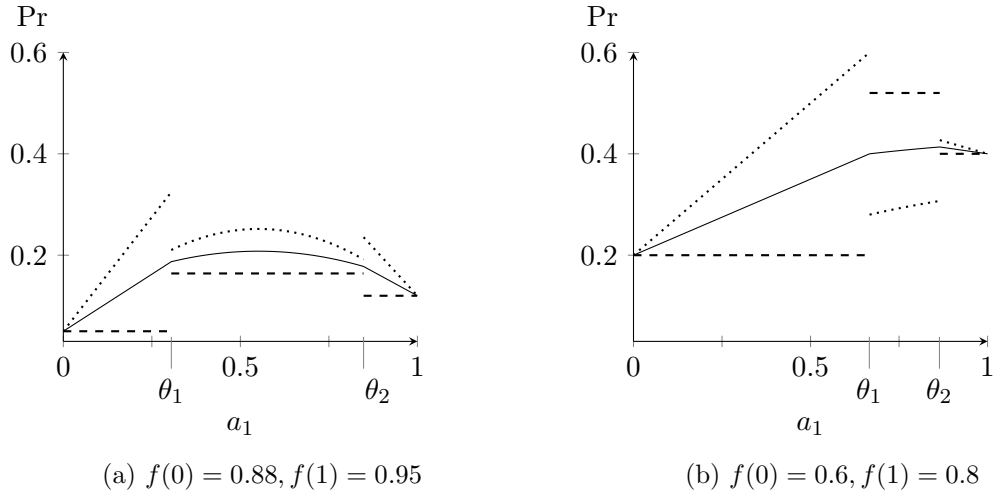


Figure 5.7: Eve's false positive rate (dashed line), Eve's false negative rate (dotted line), and Alice's payoff (solid line) as a function of \bar{a}_0 .

5.1.4 Summary

In this section we have examined three different models which all share cover objects of arbitrary length. In all models, steganographer and steganalyst were able to know the order established by the suitability exactly.

Summarizing the different models:

- ▶ first, we restricted the steganalyst to base her decision rule on one position only,
- ▶ this restriction was removed, but we differentiated between a steganographer
 - ▷ who has to embed exactly k bits, and
 - ▷ who has to embed independently with an expectation of k bits.

We have shown that in all three set-ups, the steganographer randomizes her strategy over all possible embedding positions, unless there are either perfectly uncertain ($f(i) = 1/2$) or perfectly informative ($f(i) = 1$) positions $\{i\}$.

One exceptionality of the first instantiation in this section was the existence of positions the steganalyst would never query, as they are dominated by querying other positions. These positions are “gifts” for the steganographer, as she will always embed in them, even if they are not perfectly uncertain.

A uniqueness of the third instantiation was that the optimal strategy of the steganalyst is to randomize her decision for the cover objects where her likelihood-ratio test does not yield a definite decision.

5.2 Cover Models with Two Embedding Positions

Instead of using an abstract function f we can also characterize \mathcal{P}_0 by a probability mass function (PMF) that measures the probability of the different symbol values to occur in a cover object.

The simplest model to study adaptive embedding consists of a source of heterogeneous covers of exactly two symbols ($n = 2$), $\mathbf{x}^{(0)} = (x_0^{(0)}, x_1^{(0)})$, in which Alice makes one embedding change ($k = 1$). To reduce the number of case distinctions, it is convenient to model covers ordered by decreasing suitability $\mathbf{y}^{(0)} = (y_0^{(0)}, y_1^{(0)})$. By symmetry, this is w.l.o.g. if we assume perfect recovery. Imperfect recovery can be modeled by flipping the two symbols with probability $1 - r$, where $r \in [0, 1]$ is the *recovery rate*.

The restriction to $n = 2$ symbols permits an interpretation of larger heterogeneous covers with independent symbols if they can be partitioned into two parts of equal size and suitability. The game is then repeated for each pair of heterogeneous symbols.

Alice embeds with probability \bar{a} into $y_0^{(0)}$ and with probability $1 - \bar{a}$ into $y_1^{(0)}$. With perfect recoverability, a value of $\bar{e} = 1$ means Eve examines $y_0^{(\bar{a})}$, the more suitable symbol, and $\bar{e} = 0$ means she examines $y_1^{(1-\bar{a})}$. More generally, we model Eve’s choice such that she can either examine $\hat{y}_0^{(\bar{a})}$ or $\hat{y}_1^{(1-\bar{a})}$, but not both at the same time. We justify this by the observation that Eve has no knowledge of the global distribution and

thus has to use imperfect local rules, thereby discarding some evidence. This is similar to the set-up in Section 5.1.1, where Eve is only allowed to query one position.

With this restrictions on the players, we rule out the possibility of information-theoretic optimal embedding and detection, as stated in Definition 4.19.

This set-up has several advantages. First, the simplifications allow us to draw this instantiation of the adaptive steganography game in extensive form (Figure 5.8).

Second, the assumption that the ordered symbols $\mathbf{y}^{(0)}$ are independent is a common (and possibly realistic) simplification because reordering the cover by the adaptivity criterion likely removes Markov-properties [20]. Of course, this does not prevent Eve from exploiting Markov-properties stemming from the cover in the *unordered* stego object $\mathbf{x}^{(1)}$. One way to interpret this is that she exhausts this information source when recovering the adaptivity criterion.

For both models with two embedding positions we choose LSB replacement as the embedding operation. $\text{LSBR}(x)$ can be expressed by

$$\text{LSBR}(x) := x + (-1)^x \quad \Rightarrow \quad \text{LSBR}^{-1}(x) = \text{LSBR}(x). \quad (5.61)$$

Here we see that LSBR is a special case for the biased bits used in the abstract f in the previous sections.

Lemma 5.16. *Under the assumption that $\mathcal{P}_0 \neq \mathcal{P}_1$ for LSB replacement, i. e., LSB replacement does not preserve the cover distribution perfectly, there is no equilibrium in pure strategies.*

Proof. The proof is the same as in Lemma 5.1 (on page 75), reduced to 2 positions. \square

5.2.1 Linear Increasing PMF

In this model we assume that the cover generation follows from a linearly increasing PMF. Although the PMF holds for arbitrary bit lengths, we consider the special case of positions with two bits ($\ell = 2$), when we need to enumerate all possibilities. The extension to higher bit lengths is straightforward.

5.2.1.1 Cover Generation

We need a model to represent some (simplified) conditions of heterogeneous cover sources. For this, we want to have one parameter m_i to adjust the level of heterogeneity. Now, the distribution \mathcal{P}_0 according to which the two ordered symbols $y_0^{(0)}$ and $y_1^{(0)}$ are realised, is a discrete bivariate distribution of $f_{m_0}^{(0)}$ (the PMF of $y_0^{(0)}$) and $f_{m_1}^{(0)}$ (the PMF of $y_1^{(0)}$) with $m_0 \neq m_1$ (if $m_0 = m_1$, we model a homogeneous cover). Here, m_i measures the suitability for embedding. A value of $m_i = 0$ indicates a uniform distribution (i. e., maximal entropy) and allows perfect steganography. With increasing m_i , the entropy and the suitability for embedding decrease. As we assume that $y_0^{(0)}$ is more suitable for embedding, we define $m_0 \leq m_1$.

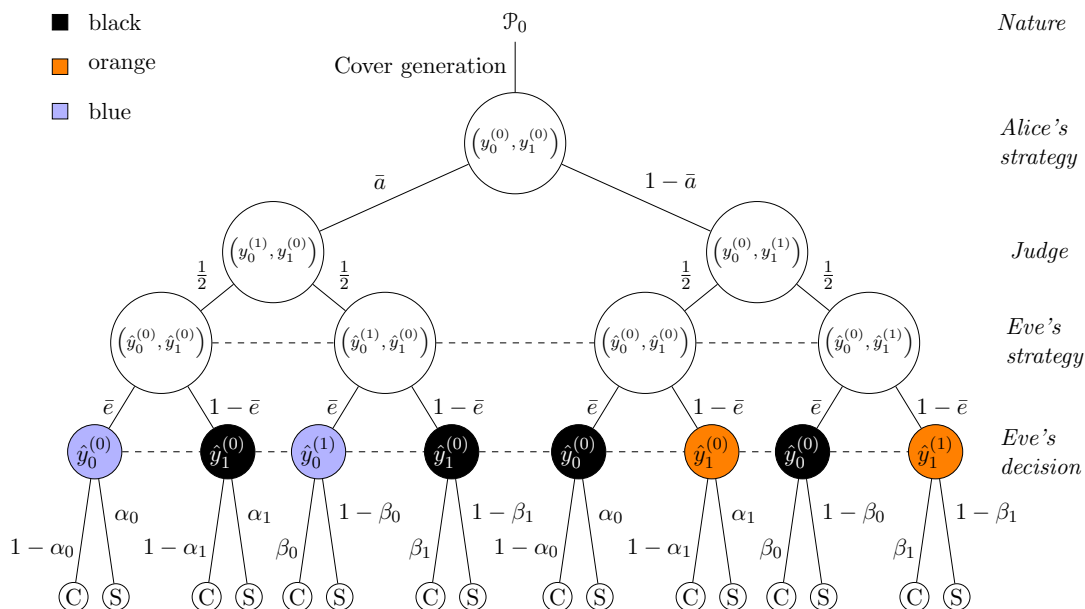


Figure 5.8: Extensive form of the instantiated adaptive steganography game. The dashed lines visualize Eve’s information sets: she does not know which of the connected nodes has been reached. α (β) is the false positive (false negative) rate of Eve’s detection decision.

So the joint PMF of the cover generation $f^{(0)}(y_0, y_1)$ is given by

$$f^{(0)}(y_0, y_1) = f_{m_0}^{(0)}(y_0) \cdot f_{m_1}^{(0)}(y_1). \quad (5.62)$$

To fulfill the requirements from above, we model the family of probability mass functions depending on m_i as

$$f_{m_i}^{(0)}(u) = (2^\ell - u)m_i + \frac{1 - \left(\sum_{j=1}^{2^\ell} j\right) m_i}{2^\ell}, \quad u \in \{0, \dots, 2^\ell - 1\}, \text{ with} \quad (5.63)$$

$$m_i \in \left[0; \left(\sum_{j=1}^{2^\ell - 1} j\right)^{-1}\right), \text{ and therefore: } m_i \in \left[0; \frac{1}{6}\right) \text{ for } \ell = 2. \quad (5.64)$$

Equation (5.63) ensures that the sum of masses equals 1 and the masses for the different symbol values are strictly decreasing. The constraints in Equation (5.64) ensure that the PMF is never negative. Note that the interval has to be open. Otherwise the value $u = 2^\ell - 1$ would have zero mass. This would allow detection with certainty whenever this value occurs in a stego object after LSB flipping.

Figure 5.9 visualizes our cover generation model. For two fixed values of m_0 , it shows the corresponding PMFs depending on m_1 . A lower value of m_0 in the homogeneous

case means a higher entropy. A bigger difference between m_0 and m_1 indicates a higher level of heterogeneity within the cover. As can be seen, by changing m_0 and m_1 , the entropy and the level of heterogeneity change simultaneously.

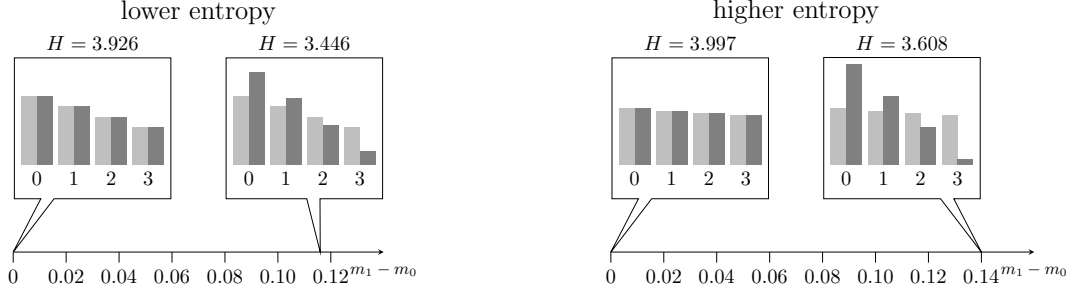


Figure 5.9: Cover generation model with increasing levels of heterogeneity from left to right. $f_{m_0}^{(0)}$ is light gray, $f_{m_1}^{(0)}$ is dark gray. Left: $m_0 = 0.05, m_1 \in \{0.05, 0.165\}$. Right: $m_0 = 0.01, m_1 \in \{0.01, 0.15\}$

5.2.1.2 Embedding Impact

Let $f_{m_i}^{(1)}$ be the PMF resulting from always embedding in $y_i^{(0)}$. Then, for single symbol values u it holds that:

$$f_{m_i}^{(0)}(u) = \Pr(u|\text{Cover}) \quad \text{and} \quad f_{m_i}^{(1)}(u) = \Pr(u|\text{Stego}). \quad (5.65)$$

As we are interested in the distribution after embedding $\mathcal{P}1$, we now proceed by examining the distribution after embedding in $y_0^{(0)}$ with probability \bar{a} and embedding in $y_1^{(0)}$ with probability $1 - \bar{a}$.

Now, in our model, where we always embed, it holds that

$$f_{m_i}^{(1)}(u) = f_{m_i}^{(0)}(\text{emb}^{-1}(u)), \quad u \in \{0, \dots, 2^\ell - 1\}. \quad (5.66)$$

This yields the following lemma about $f_{m_i}^{(1)}(u)$.

Lemma 5.17. *In our model, the PMF $f_{m_i}^{(1)}(u)$ is*

$$f_{m_i}^{(1)}(u) = \begin{cases} f_{m_i}^{(0)}(u + 1), & : u \equiv 0 \pmod{2} \\ f_{m_i}^{(0)}(u - 1), & : u \equiv 1 \pmod{2} \end{cases} \quad (5.67)$$

$$= \begin{cases} f_{m_i}^{(0)}(u) - m_i, & : u \equiv 0 \pmod{2} \\ f_{m_i}^{(0)}(u) + m_i, & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.68)$$

Proof. From Equation (5.66) we know that:

$$\begin{aligned}
 f_{m_i}^{(1)}(u) &= f_{m_i}^{(0)}(emb^{-1}(u)) \\
 &= f_{m_i}^{(0)}(u + (-1)^u) \\
 &= \begin{cases} f_{m_i}^{(0)}(u + 1), & : u \equiv 0 \pmod{2} \\ f_{m_i}^{(0)}(u - 1), & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.69)
 \end{aligned}$$

And with Equation (5.63):

$$\begin{aligned}
 f_{m_i}^{(1)}(u) &= \begin{cases} (2^\ell - (u + 1))m_i + \frac{1 - \left(\sum_{j=1}^{2^\ell} j\right)m_i}{2^\ell}, & : u \equiv 0 \pmod{2} \\ (2^\ell - (u - 1))m_i + \frac{1 - \left(\sum_{j=1}^{2^\ell} j\right)m_i}{2^\ell}, & : u \equiv 1 \pmod{2} \end{cases} \\
 &= \begin{cases} f_{m_i}^{(0)}(u) - m_i, & : u \equiv 0 \pmod{2} \\ f_{m_i}^{(0)}(u) + m_i, & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.70)
 \end{aligned}$$

□

As Lemma 5.16 excludes both pure strategies, we get a mixed strategy and thus a mixture distribution of the kind,

$$\mathcal{P}_1 = f^{(1)}(y_0, y_1) = \bar{a} \left(f_{m_0}^{(1)}(y_0) \cdot f_{m_1}^{(0)}(y_1) \right) + (1 - \bar{a}) \left(f_{m_0}^{(0)}(y_0) \cdot f_{m_1}^{(1)}(y_1) \right). \quad (5.71)$$

To quantify the overall information Eve can potentially gain from the embedding function, we can numerically calculate the KLD between $f^{(0)}$ and $f^{(1)}$ as a benchmark with the information-theoretic optimal strategies from Definition 4.19.

5.2.1.3 Eve's Decision: Optimal Local Detector

The parameter on which Eve's choice relies is \bar{e} . Conveniently, as will be shown in this paragraph, the false positive rate equals the false negative rate in our model. So we have only one variable of interest, the *equal error rate (EER)*.

Recall that we have a strictly decreasing PMF and thus for \mathcal{P}_0 it holds that,

$$f_{m_i}^{(0)}(0) > f_{m_i}^{(0)}(1) > f_{m_i}^{(0)}(2) > f_{m_i}^{(0)}(3). \quad (5.72)$$

Therefore, we know from Lemma 5.17 that in pure strategies it holds that,

$$f_{m_i}^{(1)}(1) > f_{m_i}^{(1)}(0) > f_{m_i}^{(1)}(3) > f_{m_i}^{(1)}(2). \quad (5.73)$$

So, Eve's decision rule $DR(u)$ between C (for cover) and S (for stego) follows.

Lemma 5.18. *Eve's optimal decision rule for individual symbol values u is:*

$$DR(u) = \begin{cases} \text{C}, & : u \equiv 0 \pmod{2} \\ \text{S}, & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.74)$$

Proof. The decision rule implements the *maximum a posteriori* (MAP) estimation, which can be found, for example, in [27]. With $\Pr(C) = \Pr(S) = \mu = 1/2$, the MAP estimation minimizes the decision errors by calculating:

$$\hat{q} = \arg \max_q \Pr(q|x) = \arg \max_q \Pr(x|q) \cdot \Pr(q). \quad (5.75)$$

With $q \in \{C, S\}$ and $x = u$, this results in

$$\begin{aligned} \hat{q} &= \arg \max_q \Pr(u|q) \cdot \mu \\ &\stackrel{\text{Eq. (5.65)}}{=} \max \left\{ f_{m_i}^{(0)}(u), f_{m_i}^{(1)}(u) \right\} \\ &= \begin{cases} C, & : u \equiv 0 \pmod{2} \\ S, & : u \equiv 1 \pmod{2}, \end{cases} \end{aligned} \quad (5.76)$$

because of Equations (5.72) and (5.73). \square

Thus, in our case with $n = \ell = 2$, Eve's decides for "C" whenever she sees a symbol with value 0 or 2, and "S" for values 1 and 3.

5.2.1.4 Error Rates and Payoff

5.2.1.4.1 Error Rates

Let α_i and β_i be Eve's false positive and false negative rate, respectively, for $f_{m_i}^{(0)}$ and $f_{m_i}^{(1)}$. By Lemma 5.18, her true positive rate ($1 - \alpha_i$), and consequently the false positive rate, is aggregated between the cases where her decision yields "C" and the same holds for the true negative rate ($1 - \beta_i$) in all other cases.

Lemma 5.19. *In our model, Eve's false positive rate α_i equals her false negative rate β_i and thus is called equal error rate EER_i .*

$$EER_i = \alpha_i = \beta_i = \frac{1}{2} - m_i, \quad (5.77)$$

for $i \in \{0, 1\}$.

Proof. As mentioned above, Eve's true positive and true negative rate can be calculated as follows:

True Positives $TP(x_j)$:

$$x_j = 0 : TP(0) = \frac{f_{m_i}^{(0)}(0)}{f_{m_i}^{(0)}(0) + f_{m_i}^{(1)}(0)} = \frac{f_{m_i}^{(0)}(0)}{f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(1)} \quad (5.78)$$

$$x_j = 2 : TP(2) = \frac{f_{m_i}^{(0)}(2)}{f_{m_i}^{(0)}(2) + f_{m_i}^{(1)}(2)} = \frac{f_{m_i}^{(0)}(2)}{f_{m_i}^{(0)}(2) + f_{m_i}^{(0)}(3)} \quad (5.79)$$

$$\Rightarrow (1 - \alpha_i) = (f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(1)) \cdot TP(0) + (f_{m_i}^{(0)}(2) + f_{m_i}^{(0)}(3)) \cdot TP(2) \quad (5.80)$$

$$= f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(2) \quad (5.81)$$

True Negatives $TN(x_j)$:

$$x_j = 1 : TN(1) = \frac{f_{m_i}^{(1)}(1)}{f_{m_i}^{(0)}(1) + f_{m_i}^{(1)}(1)} = \frac{f_{m_i}^{(0)}(0)}{f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(1)} = TP(0) \quad (5.82)$$

$$x_j = 3 : TN(3) = \frac{f_{m_i}^{(1)}(3)}{f_{m_i}^{(0)}(3) + f_{m_i}^{(1)}(3)} = \frac{f_{m_i}^{(0)}(2)}{f_{m_i}^{(0)}(2) + f_{m_i}^{(0)}(3)} = TP(2) \quad (5.83)$$

$$\Rightarrow (1 - \beta_i) = (f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(1)) \cdot TN(1) + (f_{m_i}^{(0)}(2) + f_{m_i}^{(0)}(3)) \cdot TN(3) \quad (5.84)$$

$$= f_{m_i}^{(1)}(1) + f_{m_i}^{(1)}(3) = f_{m_i}^{(0)}(0) + f_{m_i}^{(0)}(2) = (1 - \alpha_i) \quad (5.85)$$

$$\stackrel{Eq.(5.63)}{\Leftrightarrow} (1 - \alpha_i) = (1 - \beta_i) = 4 \cdot m_i + \frac{1 - 10m_i}{4} + 2 \cdot m_i + \frac{1 - 10m_i}{4} \quad (5.86)$$

$$= 6 \cdot m_i + 2 \cdot \frac{1 - 10m_i}{4} = \frac{2 \cdot m_i + 1}{2} = m_i + \frac{1}{2} \quad (5.87)$$

$$\Rightarrow EER_i = \alpha_i = \beta_i = \frac{1}{2} - m_i. \quad (5.88)$$

for $i \in \{0, 1\}$. □

Equation (5.77) is intuitive, as values of $m_i = 0$ indicate a uniform distribution. In this case \mathcal{P}_1 would equal \mathcal{P}_0 , i. e., the same distribution before and after embedding. Therefore the false positive and false negative rate would be $1/2$, i. e., random guessing. Furthermore, it follows our initial thoughts that a higher value of m_i implies a better detectability, which materializes in a lower EER .

Corollary 5.3. *The worst case for Eve would be Alice choosing $a \in \{0, 1\}$ and she herself choosing $e = 1 - a$ because by this, her decision would be merely guessing, i. e., $EER = 1/2$.*

Proof. If Eve chooses $e = 1 - a$ and $a \in \{0, 1\}$, it holds that Alice always embeds in $p_a^{(0)}$ and by this never into $p_e^{(0)}$. From Eq. (5.66) it follows that $f_{m_a}^{(1)}(u) = f_{m_a}^{(0)}(emb^{-1}(u))$, but $f_{m_e}^{(1)}(u) = f_{m_e}^{(0)}(u)$, as there is no embedding in $p_e^{(0)}$. Therefore, it holds that:

$$\left. \begin{array}{l} x_j \in \{0, 2\} : TP(x_j) \\ x_j \in \{1, 3\} : TN(x_j) \end{array} \right\} = \frac{f_{m_e}^{(0)}(x_j)}{f_{m_e}^{(0)}(x_j) + f_{m_e}^{(1)}(x_j)} = \frac{f_{m_e}^{(0)}(x_j)}{f_{m_e}^{(0)}(x_j) + f_{m_e}^{(0)}(x_j)} = \frac{f_{m_e}^{(0)}(x_j)}{2 \cdot f_{m_e}^{(0)}(x_j)} = \frac{1}{2}. \quad (5.89)$$

□

This confirms Lemma 5.16 that there is no equilibrium in pure strategies, as with every pure strategy, one of the players would benefit from changing her strategy to the opposite. Now we are in the position to solve the game and to identify equilibria in mixed strategies.

Table 5.3: Game outcome with correct recovery in different states of the world

Alice's choice	Eve's choice	Probability	Perfect/Correct recovery	
			$EEER$	Reason
$y_0^{(0)}$	$\hat{y}_0^{(1)}$	$\bar{a} \cdot \bar{e}$	$\frac{1}{2} - m_0$	Lemma 5.19, $i = 0$
$y_0^{(0)}$	$\hat{y}_1^{(0)}$	$\bar{a} \cdot (1 - \bar{e})$	$\frac{1}{2}$	Corollary 5.3
$y_1^{(0)}$	$\hat{y}_0^{(0)}$	$(1 - \bar{a}) \cdot \bar{e}$	$\frac{1}{2}$	Corollary 5.3
$y_1^{(0)}$	$\hat{y}_1^{(1)}$	$(1 - \bar{a}) \cdot (1 - \bar{e})$	$\frac{1}{2} - m_1$	Lemma 5.19, $i = 1$

5.2.1.4.2 Payoff Function

The $EEER$ can be seen as the payoff function in our zero-sum game. As it is Alice's intention to perform least detectable steganography, her goal is to maximize the $EEER$. It is Eve's goal to maximize her detection rate and thus, to minimize the $EEER$.

From Figure 5.8, the occurrence probabilities in Table 5.3 and the $EEER$ the payoff function $\chi(\bar{a}, \bar{e})$ for mixed strategies can be derived and equals the overall $EEER$. It is stated in the following corollary.

Lemma 5.20. *In our model, the payoff function in mixed strategies is*

$$\chi(\bar{a}, \bar{e}) = \frac{1}{2} - (\bar{a} \cdot \bar{e} \cdot m_0 + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot m_1) \quad (5.90)$$

Proof. Figure 5.8 shows that the (colored) nodes of Eve's decision can be classified into three different types.

- (1) Alice changes $y_0^{(0)}$ and Eve anticipates it (blue nodes in Figure 5.8). This situation occurs with probability $\bar{a} \cdot \bar{e}$. When faced with a situation like this, we know from Equation (5.77) that Eve's $EEER$ equals α_0 ($= \beta_0$).
- (2) Alice changes $y_1^{(0)}$ and Eve, again, anticipates it (orange nodes in Figure 5.8). The occurrence probability of this situation is $(1 - \bar{a}) \cdot (1 - \bar{e})$. Again, we know the payoff from Equation (5.77), which is α_1 ($= \beta_1$).
- (3) Alice changes $y_i^{(0)}$, but Eve examines the wrong embedding position (black nodes in Figure 5.8). This situation occurs with probability $(1 - \bar{a}) \cdot \bar{e}$ (for Alice embedding in $y_0^{(0)}$, but Eve examining $\hat{y}_1^{(1)}$) and $\bar{a} \cdot (1 - \bar{e})$ (for Alice embedding in $y_1^{(0)}$, but Eve examining $\hat{y}_0^{(1)}$). Here, we know from Corollary 5.3 that Eve's decision rule is no better than random guessing and thus has an $EEER$ of $1/2$.

Table 5.3 summarizes the respective probabilities of occurrence, payoffs, and justifications. In combination, this leads to the following expression for $\chi(\bar{a}, \bar{e})$:

$$\begin{aligned}\chi(\bar{a}, \bar{e}) &= (\bar{a} \cdot \bar{e}) \cdot \alpha_0 + ((1 - \bar{a}) \cdot \bar{e} + \bar{a} \cdot (1 - \bar{e})) \cdot \frac{1}{2} + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot \alpha_1 \\ &= (\bar{a} \cdot \bar{e}) \cdot \alpha_0 + \frac{\bar{a} + \bar{e} - 2\bar{a}\bar{e}}{2} + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot \alpha_1.\end{aligned}\quad (5.91)$$

From Lemma 5.19 we know that $\alpha_i = 1/2 - m_i$ and thus:

$$\begin{aligned}\chi(\bar{a}, \bar{e}) &= (\bar{a} \cdot \bar{e}) \cdot \left(\frac{1}{2} - m_0\right) + \frac{\bar{a} + \bar{e} - 2\bar{a}\bar{e}}{2} + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot \left(\frac{1}{2} - m_1\right) \\ &= \frac{\bar{a} \cdot \bar{e}}{2} - \bar{a} \cdot \bar{e} \cdot m_0 + \frac{\bar{a} + \bar{e} - 2\bar{a} \cdot \bar{e}}{2} \\ &\quad + \frac{1}{2} - \frac{\bar{a}}{2} - \frac{\bar{e}}{2} + \frac{\bar{a} \cdot \bar{e}}{2} - (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot m_1\end{aligned}\quad (5.92)$$

$$= \frac{1}{2} - (\bar{a} \cdot \bar{e} \cdot m_0 + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot m_1)\quad (5.93)$$

□

Remark 5.2. Note that in the pathological case of $m_0 = m_1 = 0$, i. e., a homogeneous cover source with perfect steganography possible in both symbols, it holds that $\chi(\bar{a}, \bar{e}) = 1/2$. Particularly, $\chi(\bar{a}, \bar{e})$ is independent of \bar{a} and \bar{e} . Such situations do not require game theory and thus are out of this thesis' scope and excluded in the following analysis.

5.2.1.5 Solving the Game

5.2.1.5.1 Equilibrium Strategies

Nash equilibria in two-player games are tuples of mixed strategies (\bar{a}^*, \bar{e}^*) such that no player can (strictly) increase her pay-off by unilaterally deviating from her equilibrium strategy [67]. To find a Nash equilibrium we look for a strategy that makes the opponent indifferent, i. e., a strategy where she cannot influence the pay-off by changing her strategy. We find such strategies by taking partial derivatives of the pay-off function, $\chi(\bar{a}, \bar{e})$ with regard to the opponent's strategy and setting them to zero. Then we show that these strategies indeed constitute a unique equilibrium, which happens to be symmetric.

Theorem 5.11. *In this model, there exists a unique symmetric Nash equilibrium in mixed strategies. In this equilibrium it holds that:*

$$\bar{a}^* = \bar{e}^* = \frac{m_1}{m_0 + m_1}\quad (5.94)$$

Proof. The partial derivatives of the payoff function are,

$$\frac{\partial \chi(\bar{a}, \bar{e})}{\partial \bar{a}} = -(m_0 + m_1) \cdot \bar{e} + m_1 \quad (5.95)$$

$$\frac{\partial \chi(\bar{a}, \bar{e})}{\partial \bar{e}} = -(m_0 + m_1) \cdot \bar{a} + m_1. \quad (5.96)$$

Setting both derivatives to zero yields Equation (5.94):

$$-(m_0 + m_1) \cdot \bar{e} + m_1 \stackrel{!}{=} 0 \Leftrightarrow \bar{e}^* = \frac{m_1}{m_0 + m_1} \quad (5.97)$$

$$-(m_0 + m_1) \cdot \bar{a} + m_1 \stackrel{!}{=} 0 \Leftrightarrow \bar{a}^* = \frac{m_1}{m_0 + m_1}. \quad (5.98)$$

To see that \bar{a}^* is an equilibrium strategy, we combine Equations (5.90) and (5.94):

$$\chi(\bar{a}^*, \bar{e}) = \frac{1}{2} - \left(\bar{e} \cdot \frac{m_1}{m_0 + m_1} \cdot m_1 \right) + \left(1 - \bar{e} - \frac{m_1}{m_0 + m_1} + \frac{m_1}{m_0 + m_1} \cdot \bar{e} \right)$$

Focusing only on the terms containing \bar{e} , yields:

$$\begin{aligned} & \bar{e} \cdot \left[\frac{m_1 \cdot m_0}{m_0 + m_1} + \frac{m_1^2}{m_0 + m_1} - m_1 \right] \\ = & \bar{e} \cdot \left[\frac{m_1 \cdot m_0}{m_0 + m_1} + \frac{m_1^2}{m_0 + m_1} - \frac{m_1 \cdot m_0 + m_1^2}{m_0 + m_1} \right] \quad (5.99) \\ = & \bar{e} \cdot 0. \quad (5.100) \end{aligned}$$

As the same holds for $\chi(\bar{a}, \bar{e}^*)$, both $\chi(\bar{a}^*, \bar{e})$ and $\chi(\bar{a}, \bar{e}^*)$ are independent of the opponent's strategy. Thus, $\forall \bar{a}, \bar{e} \in [0, 1] : \chi(\bar{a}^*, \bar{e}^*) = \chi(\bar{a}^*, \bar{e}) = \chi(\bar{a}, \bar{e}^*)$, and thus (\bar{a}^*, \bar{e}^*) is a Nash equilibrium.

A quick check that no combination of pure strategies is a Nash equilibrium (for $m_0 > 0$) establishes the uniqueness of (\bar{a}^*, \bar{e}^*) . The symmetry is obvious as $\bar{a}^* = \bar{e}^*$. \square

Corollary 5.4. *Only if the given cover source is homogeneous, i. e., $m_0 = m_1$, Alice's best strategy is random uniform embedding (strategy a) from Definition 4.17).*

Proof. The 'if' condition follows from the fact that for $m_0 = m_1$, it holds that:

$$\bar{a}^* = \frac{m_1}{m_0 + m_1} = \frac{1}{2}. \quad (5.101)$$

Alice changes each of the two symbols with probability $\bar{a} = 1/2$. With $k = 1$ and $n = 2$, this fulfills the definition of random uniform embedding.

If $m_0 < m_1 < 1$, it holds that:

$$\bar{a}^* = \frac{m_1}{m_0 + m_1} > \frac{m_1}{2m_1} = \frac{1}{2}. \quad (5.102)$$

This proves the 'only-if' condition. \square

Corollary 5.5. *Only if one of the symbols in the cover allows for perfect steganography, then Alice's best strategy is naïve adaptive embedding (strategy b) from Definition 4.17).*

Proof. Perfect steganography is only possible if the PMF of at least k symbols is invariant to embedding. Inserting the formal condition $m_0 = 0$ into the equilibrium condition:

$$\bar{a}^* = \frac{m_1}{m_0 + m_1} = 1. \quad (5.103)$$

Alice always changes the better suitable symbol. This fulfills the definition of naïve adaptive embedding. Whenever $m_0 > 0$ it follows that

$$\bar{a}^* = \frac{m_1}{m_0 + m_1} < 1. \quad (5.104)$$

This proves the 'only-if' condition. \square

From the uniqueness of the equilibrium and the preceding corollaries follows another property of our model.

Corollary 5.6. *If $m_0 > 0$, there are no dominated strategies and thus no dominant strategy equilibria (DSE) in our model.*

Proof. From Corollary 5.5 it follows that, unless $m_0 = 0$, the equilibrium given in Theorem 5.11 defines strategies that put positive probability on every pure strategy. Such an equilibrium is called *completely mixed equilibrium* and only exists if there is no pure or mixed strategy of any player that is strictly or weakly dominated by a convex combination of her other strategies [74]. Therefore, there are no dominant strategies and thus no dominant strategy equilibria. \square

It is easy to see that in the corner case $m_0 = 0$, the pure strategies $\bar{a}^* = \bar{e}^* = 1$ are dominant pure strategies and form a dominant strategy equilibrium.

5.2.1.5.2 Payoff in Equilibrium

Inserting the optimal strategies into $\chi(\bar{a}^*, \bar{e}^*)$ yields the equilibrium *EER*.

Corollary 5.7. *In the equilibrium it holds that the *EER* is,*

$$EER^* = \chi\left(\frac{m_1}{m_0 + m_1}, \frac{m_1}{m_0 + m_1}\right) = \frac{1}{2} - \frac{m_0 \cdot m_1}{m_0 + m_1}. \quad (5.105)$$

Proof. Equation (5.93) can be rearranged to

$$\chi(\bar{a}, \bar{e}) = \frac{1}{2} - ((m_0 + m_1) \cdot (\bar{a} \cdot \bar{e}) - m_1 \cdot \bar{a} - \bar{e} \cdot m_1 + m_1), \quad (5.106)$$

and using $\bar{e} = \bar{a} = \bar{a}^* = \frac{m_1}{m_0+m_1}$ from Theorem 5.11 we obtain,

$$\chi(\bar{a}^*, \bar{a}^*) = \frac{1}{2} - ((m_0 + m_1) \cdot (\bar{a}^*)^2 - 2 \cdot m_1 \cdot \bar{a}^* + m_1) \quad (5.107)$$

$$= \frac{1}{2} - \left((m_0 + m_1) \cdot \left(\frac{m_1}{m_0 + m_1} \right)^2 - \frac{2 \cdot m_1^2}{m_0 + m_1} + m_1 \right) \quad (5.108)$$

$$= \frac{1}{2} - \left(m_1 - \frac{m_1^2}{m_0 + m_1} \right) = \frac{1}{2} - \frac{m_0 \cdot m_1}{m_0 + m_1}. \quad (5.109)$$

□

With this unique value for \bar{a}^* , we say a steganographer performs *optimal adaptive steganography* in the model with a linear increasing PMF. It is always less detectable than a steganographer who performs naïve adaptive steganography.

A closer look at the equilibrium strategies in our model reveals that they are *equalizer strategies*.

Corollary 5.8. *The equilibrium strategies \bar{a}^* , respectively \bar{e}^* are equalizer strategies.*

Proof. From the proof of Theorem 5.11 we know that $\chi(\bar{a}^*, \bar{e}^*) = \chi(\bar{a}^*, \bar{e}) = \chi(\bar{a}, \bar{e}^*)$. Thus, if Alice plays her equilibrium strategy \bar{a}^* , Eve's strategy \bar{e} does not influence the pay-off and vice versa. From this property it follows that \bar{a}^* and \bar{e}^* are equalizer strategies. □

5.2.2 Constant Ratio PMF

In this section we assume a model where the PMF of the cover generation has a constant ration between the different values.

The specific cover model is more realistic as the cover generalization can be tied to a (discretized) Laplace distribution and its analytical solution is more general, as we can explicitly calculate the KLD for given ratio parameters.

5.2.2.1 Cover Generation and Justification

As natural covers' entropy may be below its maximum, symbol values differ in their probability of occurrence. To reflect this, let $f_{t_i}^{(0)} : \mathbb{X} \rightarrow [0, 1]$ be a family of probability mass functions (PMFs),

$$f_{t_i}^{(0)}(u) = \Pr(y_i^{(0)} = u) := \frac{(t_i)^u}{d_i}, \quad (5.110)$$

with parameter $t_i \geq 1$ and normalizing constant $d_i = \frac{1-t_i^{2^\ell}}{1-t_i}$. Observe that the probabilities of values $0, \dots, 2^\ell - 1 \in \mathbb{X}$ are increasing by a constant ratio. In the limit case, $t_i = 1$ creates a uniform distribution (i. e., maximum entropy). The entropy decreases with increasing t_i .

Now extending to $n = 2$ independent cover symbols, we restrict the parameter ranges of t_0 and t_1 to $1 \leq t_0 \leq t_1$. This allows us to generate homogenous (for $t_0 = t_1$) and heterogenous (for $t_0 < t_1$) covers with ordered suitability. (Corollary 5.9 in Sect. 5.2.2.3 will prove the very last assertion.)

Although our cover generation model is very simple and in fact artificial [7], several reasons justify its specific choice.

First, note that the PMF for individual symbols asymptotically converges to (the left half of) a discretized Laplace distribution, which is known to model the marginal distribution of real transform-coded covers reasonably well [62]. The PMF of a mean-free discretized Laplacian distribution with scale parameter p is given by [44]:

$$g_p(u) = \frac{p-1}{p+1} \cdot p^{|u|}, \quad p \in (0, 1), \quad u \in \mathbb{Z}. \quad (5.111)$$

Looking at the left half only, $u \leq 0$, simplifies it to:

$$g_p(u) = \frac{p-1}{p+1} \cdot p^{-u}. \quad (5.112)$$

As $p < 1$, we substitute $t_i := \frac{1}{p}$ in Equation (5.110) to obtain

$$f_{\frac{1}{p}}(u) = \frac{\left(\frac{1}{p}\right)^u}{d_i} = \frac{1}{d_i} \cdot p^{-u}. \quad (5.113)$$

For $t_i = \frac{1}{p}$ fixed, $\mathcal{O}(g_p)$ and $\mathcal{O}(f_{\frac{1}{p}})$ give the asymptotic equivalence in tails as u (and ℓ) go to infinity:

$$\begin{aligned} g_p(u) &\in \mathcal{O}(p^{-u}), \\ f_{\frac{1}{p}}(u) &\in \mathcal{O}(p^{-u}). \end{aligned} \quad (5.114)$$

Second, independent cover symbols imply that the entropy of the cover source is the sum of the entropy of its symbols. This way, we can vary the heterogeneity of the cover source by adjusting t_i while (numerically) enforcing constant entropy.

5.2.2.2 Embedding Impact

Let $f_{t_i}^{(1)}$ be the family of PMFs resulting from always embedding in $y_i^{(0)}$. Then, for individual values u it holds:

$$f_{t_i}^{(0)}(u) = \Pr(u \mid \text{Cover}) \text{ and } f_{t_i}^{(1)}(u) = \Pr(u \mid \text{Stego}). \quad (5.115)$$

In the cover model, we can find an analytical expression for \mathcal{P}_1 by examining the distribution after embedding in $y_0^{(0)}$ with probability \bar{a} and embedding in $y_1^{(0)}$ with probability $1 - \bar{a}$.

As we always change one symbol, it holds that

$$f_{t_i}^{(1)}(u) = f_{t_i}^{(0)}(\text{emb}^{-1}(u)). \quad (5.116)$$

This yields the following lemma about $f_{t_i}^{(1)}(u)$, the marginal distributions of \mathcal{P}_1 .

Lemma 5.21. *The PMF of stego symbols $f_{t_i}^{(1)}(u)$ is*

$$f_{t_i}^{(1)}(u) = f_{t_i}^{(0)}(u) \cdot t_i^{(-1)^u}. \quad (5.117)$$

Proof. After inserting Eq. (5.61) into Eq. (5.116),

$$f_{t_i}^{(1)}(u) = f_{t_i}^{(0)}(\text{emb}^{-1}(u)) = f_{t_i}^{(0)}(u + (-1)^u), \quad (5.118)$$

we use the definition of Eq. (5.110) and rearrange,

$$= \frac{t_i^{u+(-1)^u}}{d_i} = f_{t_i}^{(0)}(u) \cdot t_i^{(-1)^u}. \quad (5.119)$$

□

If Alice plays a mixed strategy with parameter \bar{a} , the joint distribution \mathcal{P}_1 after embedding is a mixture of the kind:

$$\begin{aligned} \mathcal{P}_1(\mathbf{y}) &= \Pr(y_0 = u, y_1 = v) \\ &= \bar{a} \left(f_{t_0}^{(1)}(u) \cdot f_{t_1}^{(0)}(v) \right) + (1 - \bar{a}) \left(f_{t_0}^{(0)}(u) \cdot f_{t_1}^{(1)}(v) \right). \end{aligned} \quad (5.120)$$

Remark 5.3. *With this cover model and embedding operation, perfect steganography is only possible if $t_0 = 1$.*

Whenever $t_0 > 1$, some simple algebra shows that \mathcal{P}_0 and \mathcal{P}_1 differ. Note that this is necessary but not sufficient to rule out the possibility of perfect steganography. Even if \mathcal{P}_0 and \mathcal{P}_1 are not the same, the marginal distributions for one symbol (the more suitable) may be equal.

5.2.2.3 Heterogeneity

The definition of heterogeneity (Definition 4.14) is indirectly based on the KLD. There is an easy way to calculate it for our model.

Lemma 5.22. *The Kullback–Leibler divergence between \mathcal{P}_0 and $\mathcal{P}_{(y_i)}$ can be calculated as follows:*

$$\text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_i)}) = \log t_i \cdot \frac{t_i - 1}{t_i + 1}. \quad (5.121)$$

Proof. We carry out the proof for $\mathcal{P}_{(y_0)}$. So, we insert $\bar{a} = 1$ into Eq. (5.120), simplify, and then expand using Eq. (5.119):

$$\mathcal{P}_{(y_0)}(u, v) = \frac{t_0^{u+(-1)^u} \cdot t_1^v}{d_0 \cdot d_1}. \quad (5.122)$$

We will use shorthand $\mathbb{X}_0 \subset \mathbb{X}$ for the set of all even elements in \mathbb{X} , and $\mathbb{X}_1 = \mathbb{X} \setminus \mathbb{X}_0$. (The subscript indicates the LSB.) Now starting from the definition of KLD (cf. Definition 2.9):

$$\begin{aligned} \text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_0)}) &= \\ &= \sum_{u \in \mathbb{X}} \sum_{v \in \mathbb{X}} \mathcal{P}_0(u, v) \cdot \log \frac{\mathcal{P}_0(u, v)}{\mathcal{P}_{(y_0)}(u, v)} \end{aligned} \quad (5.123)$$

$$\begin{aligned} &= \sum_{v \in \mathbb{X}} \left(\sum_{u \in \mathbb{X}_0} \frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \log \left(\frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \cdot \frac{d_0 \cdot d_1}{t_0^{u+1} \cdot t_1^v} \right) \right. \\ &\quad \left. + \sum_{u \in \mathbb{X}_1} \frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \log \left(\frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \cdot \frac{d_0 \cdot d_1}{t_0^{u-1} \cdot t_1^v} \right) \right) \end{aligned} \quad (5.124)$$

$$= \sum_{v \in \mathbb{X}} \left(\sum_{u \in \mathbb{X}_0} \frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \log \frac{1}{t_0} + \sum_{u \in \mathbb{X}_1} \frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \log t_0 \right) \quad (5.125)$$

$$= \sum_{v \in \mathbb{X}} \sum_{u \in \mathbb{X}} (-1)^{u+1} \cdot \frac{t_0^u \cdot t_1^v}{d_0 \cdot d_1} \log t_0 \quad (5.126)$$

$$= \log t_0 \cdot \frac{1}{d_0 \cdot d_1} \cdot \sum_{u \in \mathbb{X}} (-1)^{u+1} \cdot t_0^u \cdot \underbrace{\sum_{v \in \mathbb{X}} t_1^v}_{=d_1} \quad (5.127)$$

$$= \log t_0 \cdot \frac{1}{d_0} \cdot (-1) \cdot \sum_{u=0}^{2^\ell-1} (-t_0)^u. \quad (5.128)$$

Now using a closed form for the sum of the geometric series:

$$= \log t_0 \cdot \frac{1 - t_0}{1 - t_0^{2^\ell}} \cdot (-1) \cdot \frac{1 - (-t_0)^{2^\ell}}{1 - (-t_0)} \quad (5.129)$$

$$= \log t_0 \cdot \frac{t_0 - 1}{t_0 + 1}. \quad (5.130)$$

The proof for $\text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_1)})$ is analogous □

As the symbols are independent, the amount of distortion introduced by embedding, as measured by the KLD, only depends on the PMF of the symbol used for embedding.

Corollary 5.9. *If $t_0 < t_1$, then $y_0^{(0)}$ is more suitable for embedding than $y_1^{(0)}$.*

Proof. If $t_0 < t_1$, then $\log t_0 \cdot \frac{t_0-1}{t_0+1} < \log t_1 \cdot \frac{t_1-1}{t_1+1}$, hence by Lemma 5.22:
 $\text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_0)}) < \text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_1)})$. □

Remark 5.4. *The difference in the KLD between (1) changing only the least suitable and (2) changing only the best suitable symbol is a metric to quantify the heterogeneity of a cover source: $\Delta \text{KLD} = \text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_1)}) - \text{KLD}(\mathcal{P}_0, \mathcal{P}_{(y_0)})$.*

Note that this metric depends on the embedding operation, like our notions of heterogeneity and suitability.

The histograms in Figure 5.10 show examples of two different parameterizations of the cover source with a fixed alphabet of four values ($\ell = 2$). The smaller parameter t_i , the closer is the distribution to a uniform distribution and the less detectable is the embedding operation LSBR (as indicated by the arrows). Figure 5.10(a) shows a homogeneous cover source. Only for heterogeneous cover sources (Figure 5.10(b)), Alice can take advantage of adaptively choosing more suitable positions. This advantage increases with the level of heterogeneity.

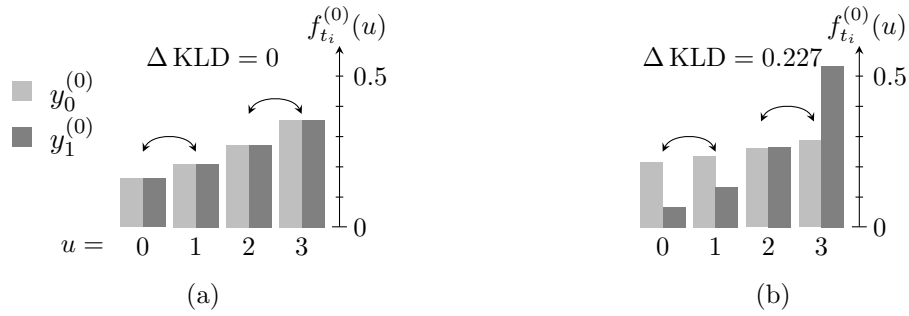


Figure 5.10: Example histograms of the cover source for $n = \ell = 2$. Compare the more suitable (brighter bars) to the less suitable (darker bars) position for a: (a) homogeneous ($t_0 = t_1 = 1.3$); (b) heterogeneous ($t_0 = 1.1, t_1 = 2$) cover source. The arrows indicate which values are exchanged by the LSBR embedding operation.

5.2.2.4 Eve's Decision: Optimal Local Detector

For this analysis, we equip Eve with the locally optimal decision rule, specific to the embedding operation LSBR and the cover generation model.

Eve's decision rule $\text{DR}(u)$ between C (for cover) and S (for stego) follows from the *maximum a posteriori* (MAP) estimation [27, for example], and the fairness of the Judge ($\mu = 1/2$).

Lemma 5.23. *Eve's locally optimal decision rule when examining an individual symbol and finding value u is:*

$$\text{DR}(u) = \begin{cases} \text{S} & : u \equiv 0 \pmod{2} \\ \text{C} & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.131)$$

Proof. MAP estimation minimizes the decision errors by using Bayes' theorem:

$$\hat{q} = \arg \max_q \Pr(q | u) = \arg \max_q \Pr(u | q) \cdot \Pr(q). \quad (5.132)$$

With $q \in \{C, S\}$, we obtain

$$\hat{q} = \arg \max_q \Pr(u | q) \cdot \mu \quad (5.133)$$

$$\stackrel{\text{Eq. (5.115)}}{=} \arg \max \left\{ C : f_{t_i}^{(0)}(u), S : f_{t_i}^{(1)}(u) \right\}, \quad (5.134)$$

now using Lemma 5.21 and dividing element-wise by $f_{t_i}^{(0)}(u)$,

$$= \arg \max \left\{ C : 1, S : t_i^{(-1)^u} \right\}, \quad (5.135)$$

$$= \begin{cases} S & : u \equiv 0 \pmod{2} \\ C & : u \equiv 1 \pmod{2}. \end{cases} \quad (5.136)$$

The last identity follows from the fact that $t_i \geq 1$. If $t_i = 1$, Eve is indifferent, but the rule is still optimal in the sense that she cannot do better than random guessing. \square

Remember that fixing the embedding operation (in Sect. 5.2.2.2) and this detector generally precludes both Alice and Eve from using the information-theoretical optimal strategies (cf. Definition 4.19) (unless $t_i = 1$). This is intentional to reflect the hardness of reaching these goals in practice. It allows us to analyze the players' strategies under knowledge and computational constraints.

5.2.2.5 Error Rates and Payoff

5.2.2.5.1 Error Rates

As mentioned in Section 4.4.2, Eve's error rates quantify steganographic security. In our model, the error rates depend on the parameters t_i . Let α_i (β_i) be Eve's false positive (false negative) probability when applying DR on $f_{t_i}^{(0)}$ ($f_{t_i}^{(1)}$). We use Eve's *average error rate* (under equal priors) $\text{AER} = (\alpha_i + \beta_i)/2$ to measure steganographic security in this analysis.

Lemma 5.24. *If Eve investigates the same position $i \in \{0, 1\}$ that Alice has changed for embedding, then*

$$\text{AER} = \frac{1}{t_i + 1}. \quad (5.137)$$

Proof. False positives occur if DR classifies a symbol drawn from $f_{t_i}^{(0)}$ as "S".

$$\alpha_i = \sum_{u=0}^{2^{(\ell-1)}-1} f_{t_i}^{(0)}(2u) \stackrel{\text{Eq. (5.110)}}{=} \sum_{u=0}^{2^{(\ell-1)}-1} \frac{(t_i)^{2u}}{d_i} \quad (5.138)$$

$$= \frac{\frac{t_i^{2^\ell} - 1}{t_i^2 - 1}}{\frac{t_i^{2^\ell} - 1}{t_i - 1}} = \frac{t_i - 1}{t_i^2 - 1} = \frac{1}{t_i + 1}. \quad (5.139)$$

False negatives occur if DR classifies a symbol drawn from $f_{t_i}^{(1)}$ as “C”.

$$\beta_i = \sum_{u=0}^{2^{(\ell-1)}-1} f_{t_i}^{(1)}(2u+1) \quad (5.140)$$

Rewriting in terms of $f_{t_i}^{(0)}$ (with the help of Lemma 5.21):

$$= \sum_{u=0}^{2^{(\ell-1)}-1} \frac{f_{t_i}^{(0)}(2u+1)}{t_i} \stackrel{\text{Eq. (5.110)}}{=} \sum_{u=0}^{2^{(\ell-1)}-1} \frac{(t_i)^{2u+1}}{d_i \cdot t_i}. \quad (5.141)$$

After reducing t_i from the right hand side of Eq. (5.141), the term equals the right hand side of Eq. (5.138) and it follows that

$$\text{AER} := \frac{\alpha_i + \beta_i}{2} = \frac{1}{t_i + 1}. \quad (5.142)$$

□

Equation (5.137) is intuitive, as the error probability is $1/2$ (random guessing) for the boundary case $t_i = 1$; uniform i. i. d. where LSBR is undetectable. It also illustrates Corollary 5.9 because higher values of t_i imply less suitability for embedding, which leads to a lower AER, and vice versa.

Corollary 5.10. *The worst case for Eve is Alice choosing $a \in \{0, 1\}$ and she herself choosing $e = 1 - a$. In this case, her decision is no better than random guessing, i. e., $\text{AER} = 1/2$.*

Proof. If $e = 1 - a$, Eve’s decision rule is always applied to symbols drawn from the (marginal) cover distribution. For every symbol $u \in \mathbb{X}$, let bias $\tilde{f}_u \in [0, 1]$ be the probability that any probabilistic decision rule (including DR from Lemma 5.23) returns S for (stego) upon finding value u . Then,

$$\text{AER} | _u = \frac{\alpha | _u + \beta | _u}{2} = \frac{\tilde{f}_u + (1 - \tilde{f}_u)}{2} = \frac{1}{2}. \quad (5.143)$$

$\text{AER} | _u$ is independent of u , hence $\text{AER} = 1/2$. □

In this section, we have instantiated and explained a concrete cover model, embedding operation, and detector within the general framework of Section 4.4. Now we are in the position to solve this instantiation of our game and find its equilibrium.

With all components of the framework instantiated, we first derive the payoff function and then solve the game for Nash equilibria. Throughout this section we assume that Eve can perfectly recover the order of the suitability of the embedding positions; formally: $\hat{\mathbf{y}}^{(\bar{a})} = \mathbf{y}^{(0)}$.

5.2.2.5.2 Payoff Function

Being agnostic about detailed cost assumptions, we devise a zero-sum game with the AER determining the payoffs. Alice wants to perform least detectable steganography, hence she tries to maximize the AER. Eve's goal is to maximize her detection rate, hence she tries to minimize the AER. Consequently, Alice's utility is her expected AER, and Eve's utility is her expected $-$ AER. Expectations are taken over realizations of random variables governed by Nature as well as the realizations of the players' strategies A and E .

Table 5.4 lists all possible states (in rows), the associated AER for two different scenarios (column blocks), and how we obtain it. Note that each row aggregates both possible outcomes of the Judge's coin flip and the AER combines both error rates.

Lemma 5.25. *The expected AER in mixed strategies is*

$$\begin{aligned} \chi(\bar{a}, \bar{e}) = & 1 - \frac{\bar{a} + \bar{e}}{2} + \left(\frac{1}{t_0 + 1} - \frac{t_1}{t_1 + 1} \right) \cdot \bar{a}\bar{e} \\ & - (1 - \bar{a} - \bar{e}) \cdot \left(\frac{t_1}{t_1 + 1} \right). \end{aligned} \quad (5.144)$$

Proof. Figure 5.8 shows that the (colored) nodes of Eve's decision can be classified into three different types.

- (1) Alice changes $y_0^{(0)}$ and Eve anticipates it (blue nodes in Figure 5.8). This situation occurs with probability $\bar{a} \cdot \bar{e}$. When faced with a situation like this, we know from Equation (5.137) that Eve's AER equals $\frac{1}{t_0}$.
- (2) Alice changes $y_1^{(0)}$ and Eve, again, anticipates it (orange nodes in Figure 5.8). The occurrence probability of this situation is $(1 - \bar{a}) \cdot (1 - \bar{e})$. Again, we know the payoff from Equation (5.137), which is $\frac{1}{t_1 + 1}$.
- (3) Alice changes $y_i^{(0)}$, but Eve examines the wrong embedding position (black nodes in Figure 5.8). This situation occurs with probability $(1 - \bar{a}) \cdot \bar{e}$ (for Alice embedding in $y_0^{(0)}$, but Eve examining $\hat{y}_1^{(1)}$) and $\bar{a} \cdot (1 - \bar{e})$ (for Alice embedding in $y_1^{(0)}$, but Eve examining $\hat{y}_0^{(1)}$). Here, we know from Corollary 5.10 that Eve's decision rule is no better than random guessing and thus has an AER of $1/2$.

Table 5.4 summarizes the respective probabilities of occurrence, payoffs, and justifications.

Table 5.4: Game outcome with correct recovery in different states of the world

Alice's choice	Eve's choice	Probability	Perfect/Correct recovery	
			AER	Reason
$y_0^{(0)}$	$\hat{y}_0^{(1)}$	$\bar{a} \cdot \bar{e}$	$\frac{1}{t_0+1}$	Lemma 5.24, $i = 0$
$y_0^{(0)}$	$\hat{y}_1^{(0)}$	$\bar{a} \cdot (1 - \bar{e})$	$\frac{1}{2}$	Corollary 5.10
$y_1^{(0)}$	$\hat{y}_0^{(0)}$	$(1 - \bar{a}) \cdot \bar{e}$	$\frac{1}{2}$	Corollary 5.10
$y_1^{(0)}$	$\hat{y}_1^{(1)}$	$(1 - \bar{a}) \cdot (1 - \bar{e})$	$\frac{1}{t_1+1}$	Lemma 5.24, $i = 1$

In combination, this leads to the following expression for $\chi(\bar{a}, \bar{e})$:

$$\begin{aligned}
 \chi(\bar{a}, \bar{e}) &= (\bar{a}\bar{e}) \cdot \left(\frac{1}{t_0+1} \right) \\
 &\quad + \frac{\bar{a} \cdot (1 - \bar{e})}{2} + \frac{(1 - \bar{a}) \cdot \bar{e}}{2} \\
 &\quad + (1 - \bar{a})(1 - \bar{e}) \cdot \left(\frac{1}{t_1+1} \right). \tag{5.145}
 \end{aligned}$$

Equation (5.144) follows from rearranging Equation (5.145). \square

Remark 5.5. *Note that in the pathological case of $t_0 = t_1 = 1$, i. e., a homogeneous cover source with perfect steganography possible in both symbols, it holds that $\chi(\bar{a}, \bar{e}) = 1/2$. Particularly, $\chi(\bar{a}, \bar{e})$ is independent of \bar{a} and \bar{e} . Such situations do not require game theory and thus are out of this thesis' scope and excluded in the following analysis.*

5.2.2.6 Solving the Game

5.2.2.6.1 Equilibrium Strategies

With the same method as in Section 5.2.1.5, we find a unique equilibrium that is symmetric.

Theorem 5.12. *There exists a unique symmetric Nash equilibrium in mixed strategies. In this equilibrium it holds that:*

$$\bar{a}^* = \bar{e}^* = \frac{(1 - t_1)(1 + t_0)}{2(1 - t_0 t_1)}. \tag{5.146}$$

Proof. The partial derivatives of the expected AER function are:

$$\frac{\partial \chi(\bar{a}, \bar{e})}{\partial \bar{a}} = -\frac{1}{2} + \left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1}\right) \cdot \bar{e} + \frac{t_1}{t_1 + 1}, \quad (5.147)$$

$$\frac{\partial \chi(\bar{a}, \bar{e})}{\partial \bar{e}} = -\frac{1}{2} + \left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1}\right) \cdot \bar{a} + \frac{t_1}{t_1 + 1}. \quad (5.148)$$

Setting both derivatives to zero yields Equation (5.146):

$$\left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1}\right) \cdot \bar{e}^* \stackrel{!}{=} \frac{1}{2} - \frac{t_1}{t_1 + 1} \quad (5.149)$$

$$\Leftrightarrow \bar{e}^* = \frac{(1 - t_1)(1 + t_0)}{2 \cdot (1 - t_0 t_1)}, \text{ and}$$

$$\left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1}\right) \cdot \bar{a}^* \stackrel{!}{=} \frac{1}{2} - \frac{t_1}{t_1 + 1} \quad (5.150)$$

$$\Leftrightarrow \bar{a}^* = \frac{(1 - t_1)(1 + t_0)}{2 \cdot (1 - t_0 t_1)}.$$

To see that \bar{a}^* is an equilibrium strategy, we combine Equations (5.144) and (5.146):

$$\begin{aligned} \chi(\bar{a}^*, \bar{e}) &= \frac{1}{t_1 + 1} + \left(\frac{t_1 - 1}{2(t_1 + 1)}\right) \cdot \left(\frac{(1 - t_1)(t_0 + 1)}{2(1 - t_0 t_1)}\right) \\ &\quad + \left(\frac{t_1 - 1}{2(t_1 + 1)}\right) \bar{e} \\ &\quad + \left(\frac{1 - t_0 t_1}{(t_0 + 1)(t_1 + 1)}\right) \cdot \left(\frac{(1 - t_1)(t_0 + 1)}{2(1 - t_0 t_1)}\right) \bar{e}. \end{aligned} \quad (5.151)$$

Considering only the terms containing \bar{e} :

$$\bar{e} \cdot \left(\frac{t_1 - 1}{2(t_1 + 1)} + \frac{1 - t_1}{2(t_1 + 1)}\right) = \bar{e} \cdot 0. \quad (5.152)$$

As the same holds for $\chi(\bar{a}, \bar{e}^*)$, both $\chi(\bar{a}^*, \bar{e})$ and $\chi(\bar{a}, \bar{e}^*)$ are independent of the opponent's strategy. Thus, $\forall \bar{a}, \bar{e} \in [0, 1] : \chi(\bar{a}^*, \bar{e}^*) = \chi(\bar{a}^*, \bar{e}) = \chi(\bar{a}, \bar{e}^*)$, and thus (\bar{a}^*, \bar{e}^*) is a Nash equilibrium.

A quick check that no combination of pure strategies is a Nash equilibrium (for $t_0 > 1$) establishes the uniqueness of (\bar{a}^*, \bar{e}^*) . The symmetry is obvious as $\bar{a}^* = \bar{e}^*$. \square

The following corollaries state two direct implications for the design of more secure embedding functions.

Corollary 5.11. *Only if the given cover source is homogeneous, i. e., $t_0 = t_1$, Alice's best strategy is random uniform embedding (strategy a) from Definition 4.17).*

Proof. The 'if' direction follows from the fact that for $t_0 = t_1$, it holds that:

$$\bar{a}^* = \frac{(1 - t_1)(1 + t_0)}{2 \cdot (1 - t_0 t_1)} = \frac{(1 - t_0)(1 + t_0)}{2 \cdot (1 - t_0^2)} = \frac{1}{2}. \quad (5.153)$$

Alice changes each of the two symbols with probability $\bar{a} = 1/2$. With $k = 1$ and $n = 2$, this fulfills the definition of random uniform embedding.

If $t_0 < t_1$, it holds that:

$$\bar{a}^* = \frac{(1-t_1)(1+t_0)}{2 \cdot (1-t_0t_1)} = \frac{1}{2} \cdot \underbrace{\left(\frac{\overbrace{t_0 - t_1}^{<0} + (1-t_0t_1)}{1-t_0t_1} \right)}_{>1} > \frac{1}{2}. \quad (5.154)$$

This proves the ‘only-if’ direction. \square

Corollary 5.12. *Only if one of the symbols in the cover allows for perfect steganography, then Alice’s best strategy is naïve adaptive embedding (strategy b) from Definition 4.17).*

Proof. Perfect steganography is only possible if the PMF of at least k symbols is invariant to embedding. Inserting the formal condition, $t_0 = 1$ (from Remark 5.3), into the equilibrium condition:

$$\bar{a}^* = \frac{(1-t_1)(1+t_0)}{2 \cdot (1-t_0t_1)} = \frac{(1-t_1) \cdot 2}{2 \cdot (1-t_1)} = 1. \quad (5.155)$$

Alice always changes the better suitable symbol. This fulfills the definition of naïve adaptive embedding. Whenever $t_0 > 1$ it follows that

$$t_0(t_1 + 1) > t_1 + 1 \quad \Leftrightarrow \quad t_0t_1 - 1 > t_1 - t_0. \quad (5.156)$$

Rewriting Equation (5.146) yields:

$$\bar{a}^* = \frac{1}{2} + \frac{1}{2} \cdot \underbrace{\left(\frac{t_1 - t_0}{t_0t_1 - 1} \right)}_{<1} < 1. \quad (5.157)$$

This proves the ‘only-if’ condition. \square

From the uniqueness of the equilibrium and the preceding corollaries follows another property of our model.

Corollary 5.13. *If $t_0 > 1$, there are no dominated strategies and thus no dominant strategy equilibria (DSE) in our model.*

Proof. From Corollary 5.12 it follows that, unless $t_0 = 1$, the equilibrium given in Theorem 5.12 defines strategies that put positive probability on every pure strategy. Such an equilibrium is called *completely mixed equilibrium* and only exists if there is no pure or mixed strategy of any player that is strictly or weakly dominated by a convex combination of her other strategies [74]. Therefore, there are no dominant strategies and thus no dominant strategy equilibria. \square

It is easy to see that in the corner case $t_0 = 1$, the pure strategies $\bar{a}^* = \bar{e}^* = 1$ are dominant pure strategies and form a dominant strategy equilibrium.

5.2.2.6.2 Payoff in Equilibrium

Now, that we determined the equilibrium strategies for Alice, respectively Eve, we can calculate the payoff in equilibrium.

Corollary 5.14. *The expected AER in equilibrium is*

$$\chi(\bar{a}^*, \bar{e}^*) = \frac{(t_0 + 1)(t_1 + 1) - 4}{4(t_0 t_1 - 1)}. \quad (5.158)$$

This corollary follows directly from inserting the equilibrium conditions (Theorem 5.12) into Lemma 5.25.

A closer look at the equilibrium strategies in our model reveals that they are *equalizer strategies*.

Corollary 5.15. *The equilibrium strategies \bar{a}^* , respectively \bar{e}^* are equalizer strategies.*

Proof. From the proof of Theorem 5.12 we know that $\chi(\bar{a}^*, \bar{e}^*) = \chi(\bar{a}^*, \bar{e}) = \chi(\bar{a}, \bar{e}^*)$. Thus, if Alice plays her equilibrium strategy \bar{a}^* , Eve's strategy \bar{e} does not influence the pay-off and vice versa. From this property it follows that \bar{a}^* and \bar{e}^* are equalizer strategies. \square

This yields the following corollary.

Corollary 5.16. *If Alice (Eve) plays her equilibrium strategy, she balances Eve's (Alice's) advantage over choosing a specific position and creates a uniform local advantage.*

Proof. The corollary follows directly from the fact that equalizer strategies make the other player indifferent between the strategies of the opponent [74]. In our model, this means that the local advantage is the same for every position, i. e., a uniform local advantage. \square

5.2.3 Imperfect Recoverability

In this section we build upon the set-up and the results of the previous sections. Especially the assumption of Eve being able to perfectly recover the order of possible embedding positions is unrealistic. We have seen in Table 3.2 from Section 3.3.5 that the adaptivity criteria used in practice differ in their recoverability from the stego object. As the recovery rate is approximately constant for a given adaptivity criterion, it is sensible to grant both players full information about the (average) recovery rate r .

In our models with two positions, the correct definition for the recovery rate r , corresponding to Definition 3.7 from Section 3.3.3, is as follows:

Definition 5.3 (Recovery Rate for Two Embedding Positions).

The recovery rate r is the probability that Eve can correctly recover the order of the symbols, i. e., $\hat{\mathbf{y}}^{(1)} = \mathbf{y}^{(1)}$. With two embedding positions, this implies that with probability $(1 - r)$ she assumes the wrong order, i. e., $\hat{y}_i^{(\bar{a})} = y_{(1-i)}^{(0)}$ for $i \in \{0, 1\}$.

Table 5.5: Game outcome with incorrect recovery in different states of the world

Alice's choice	Eve's choice	Probability	Incorrect recovery		
			Reality	EER	Reason
$y_0^{(0)}$	$\hat{y}_0^{(1)}$	$\bar{a} \cdot \bar{e}$	$y_1^{(0)}$	$\frac{1}{2}$	Corollary 5.3
$y_0^{(0)}$	$\hat{y}_1^{(0)}$	$\bar{a} \cdot (1 - \bar{e})$	$y_0^{(1)}$	$\frac{1}{2} - m_0$	Lemma 5.19, $i = 0$
$y_1^{(0)}$	$\hat{y}_0^{(0)}$	$(1 - \bar{a}) \cdot \bar{e}$	$y_1^{(1)}$	$\frac{1}{2} - m_1$	Lemma 5.19, $i = 1$
$y_1^{(0)}$	$\hat{y}_1^{(1)}$	$(1 - \bar{a}) \cdot (1 - \bar{e})$	$y_0^{(0)}$	$\frac{1}{2}$	Corollary 5.3

It is easy to see that with this definition of the recovery rate, the payoff function $\chi_r(\bar{a}, \bar{e})$ is symmetric around $r = \frac{1}{2}$ for both models with two embedding positions. If it holds that $r < 1/2$, Eve flips the output of Equation (5.74) for the model with linear PMF or Equation (5.131) for the constant ratio PMF and thus gets $r' = 1 - r$.

The experimental results in Section 3.3.6 suggest that when Alice uses naïve adaptive embedding in heterogeneous covers, Eve's performance is positively associated with the recoverability of the adaptivity criterion. In this section we want to ascertain if imperfect recovery also influences both players' game-theoretic optimal strategies.

Note that a recovery rate of $r = \frac{1}{2}$ indicates (conditionally) perfect SSI, as Eve is not able to recover the order better than random guessing. For such a situation, the payoff is probably not influenced by Eve's strategy.

5.2.3.1 Imperfect Recovery with Linear Increasing PMF

With the introduction of imperfect recoverability, i. e., $1/2 < r < 1$, we need to adjust the payoff function from Equation (5.90).

Lemma 5.26. *The payoff function in this model with recovery rate r is:*

$$\begin{aligned} \chi_r(\bar{a}, \bar{e}) = & r \cdot \left(\frac{1}{2} - (\bar{a} \cdot \bar{e} \cdot m_0 + (1 - \bar{a}) \cdot (1 - \bar{e}) \cdot m_1) \right) \\ & + (1 - r) \cdot \left(\frac{1}{2} - \bar{a} \cdot (1 - \bar{e}) \cdot m_0 - (1 - \bar{a}) \cdot \bar{e} \cdot m_1 \right). \end{aligned} \quad (5.159)$$

Proof. Imperfect recovery is modeled by a mixture of correct and incorrect recovery. The payoff function from Lemma 5.20 holds with probability r for the case of correct recovery. And with probability $(1 - r)$, the payoff function is given by the terms in columns 4–6 of Table 5.5 for the case of incorrect recovery. \square

Theorem 5.13. *There exists a unique (asymmetric) Nash equilibrium in mixed strategies for $r \neq 1/2$. In this equilibrium it holds that:*

$$\bar{a}_r^* = \frac{m_1}{m_0 + m_1} \quad (5.160)$$

$$\bar{e}_r^* = \frac{m_0 - r(m_0 + m_1)}{(1 - 2r)(m_0 + m_1)}. \quad (5.161)$$

Proof. The partial derivatives of the payoff function are:

$$\frac{\partial \chi_r(\bar{a}, \bar{e})}{\partial \bar{e}} = (1 - 2r)(m_0 + m_1)\bar{a} - (1 - 2r)m_1, \quad (5.162)$$

$$\frac{\partial \chi_r(\bar{a}, \bar{e})}{\partial \bar{a}} = (1 - 2r)(m_0 + m_1)\bar{e} - m_0r(m_0 + m_1). \quad (5.163)$$

Setting both derivatives to zero yields the strategies.

Inserting \bar{a}_r^* in the partial derivative of the second term of Eq. (5.159) (factor $(1 - r)$), which describes the case when Eve is not able to recover the order of the positions, eliminates all factors containing \bar{e} in this term. The same was already shown for the first term of Eq. (5.159) (factor r) in the proof of Theorem 5.11. Some algebra shows that $\chi_r(\bar{a}, \bar{e}_r^*)$ is independent of \bar{a} as well and thus, with the same arguments as in the proof of Theorem 5.11, $(\bar{a}_r^*, \bar{e}_r^*)$ is a Nash equilibrium. \square

Notably, Alice follows the same strategy as with perfect recoverability, whereas Eve deviates from her strategy.

As can be seen from the new equilibrium strategies, Eve's strategy is not well-defined for $r = 1/2$. Thus, we handle this case separately.

Corollary 5.17. *In the case of $r = 1/2$ the payoff function $\chi_{\frac{1}{2}}$ is linear in \bar{a} and independent of \bar{e} . By this, Eve cannot influence the payoff. Alice's best strategy is $\bar{a} = 1$, i. e. naïve adaptive embedding. The payoff in equilibrium is always between $\frac{5}{12}$ and $\frac{1}{2}$, and depends only on m_0 .*

Proof. Inserting $r = 1/2$ into Equation (5.159), yields:

$$\chi_{\frac{1}{2}}(\bar{a}, \bar{e}) = \frac{1}{2}(1 - m_1) + \frac{m_1 - m_0}{2} \bar{a}, \quad (5.164)$$

which is linear in \bar{a} and independent of \bar{e} . Obviously, the slope $\frac{m_1 - m_0}{2}$ is positive whenever $m_0 < m_1$. Thus, the maximum is reached at $\bar{a} = 1$ and the payoff is:

$$\frac{5}{12} < \frac{1 - m_0}{2} \leq \frac{1}{2}, \quad (5.165)$$

as $\frac{1}{6} > m_0 \geq 0$ (cf. Equation (5.64)). \square

Table 5.6: Game outcome with incorrect recovery in different states of the world

Alice's choice	Eve's choice	Probability	Incorrect recovery		
			Reality	AER	Reason
$y_0^{(0)}$	$\hat{y}_0^{(1)}$	$\bar{a} \cdot \bar{e}$	$y_1^{(0)}$	$\frac{1}{2}$	Corollary 5.10
$y_0^{(0)}$	$\hat{y}_1^{(0)}$	$\bar{a} \cdot (1 - \bar{e})$	$y_0^{(1)}$	$\frac{1}{t_0+1}$	Lemma 5.24, $i = 0$
$y_1^{(0)}$	$\hat{y}_0^{(0)}$	$(1 - \bar{a}) \cdot \bar{e}$	$y_1^{(1)}$	$\frac{1}{t_1+1}$	Lemma 5.24, $i = 1$
$y_1^{(0)}$	$\hat{y}_1^{(1)}$	$(1 - \bar{a}) \cdot (1 - \bar{e})$	$y_0^{(0)}$	$\frac{1}{2}$	Corollary 5.10

Interpreted for realistic cover sources, this special case echos the obvious that if there is no leakage of information through the recoverability of the adaptivity criterion, there is no advantage for Eve if she tries to recover it.

For $r \neq 1/2$, we find that the equilibrium strategies are still equalizer strategies, and the game outcome is the same as in the case of perfect recovery.

Lemma 5.27. *With recovery rate r , the equilibrium strategies are equalizer strategies and the payoff in equilibrium is:*

$$\chi_r(\bar{a}_r^*, \bar{e}_r^*) = \frac{1}{2} - \frac{m_0 \cdot m_1}{m_0 + m_1}. \quad (5.166)$$

Proof. From the proof of Theorem 5.13 follows that the players cannot influence the pay-off when the other player uses her equilibrium strategy. Thus, \bar{a}_r^* and \bar{e}_r^* are equalizer strategies. The payoff follows from combining Equations (5.159), (5.162) and (5.163). \square

5.2.3.2 Imperfect Recovery with Constant Ratio PMF

With imperfect recoverability, i. e., $1/2 < r < 1$, we need to adjust the payoff function in Equation (5.144) in a similar way as above.

Lemma 5.28. *The payoff function in the model with recovery rate r is:*

$$\begin{aligned} \chi_r(\bar{a}, \bar{e}) = & r \cdot \left(1 - \frac{\bar{a} + \bar{e}}{2} + \left(1 - \frac{t_0}{t_0+1} - \frac{t_1}{t_1+1} \right) \cdot \bar{a}\bar{e} \right. \\ & \left. - (1 - \bar{a} - \bar{e}) \cdot \left(\frac{t_1}{t_1+1} \right) \right) \\ & + (1 - r) \cdot \left(\frac{1}{2} - \frac{\bar{a} + \bar{e}}{2} - \left(1 - \frac{t_0}{t_0+1} - \frac{t_1}{t_1+1} \right) \cdot \bar{a}\bar{e} \right) \end{aligned}$$

$$+ \left(1 - \frac{t_0}{t_0 + 1}\right) \cdot \bar{e} + \left(1 - \frac{t_1}{t_1 + 1}\right) \cdot \bar{a}. \quad (5.167)$$

Proof. Imperfect recovery is modeled by a mixture of correct and incorrect recovery. The payoff function from Corollary 5.25 holds with probability r for the case of correct recovery. And with probability $(1 - r)$, the payoff function is given by the terms in columns 4–6 of Table 5.6 for the case of incorrect recovery. \square

Theorem 5.14. *There exists a unique (asymmetric) Nash equilibrium in mixed strategies for $r \neq 1/2$. In this equilibrium it holds that:*

$$\bar{a}_r^* = \frac{(1 - t_1)(1 + t_0)}{2(1 - t_0 t_1)}, \quad (5.168)$$

$$\bar{e}_r^* = \frac{1}{2} - \frac{t_0 - t_1}{2(2r - 1)(t_0 t_1 - 1)}. \quad (5.169)$$

Proof. The partial derivatives of the payoff function are:

$$\begin{aligned} \frac{\partial \chi_r(\bar{a}, \bar{e})}{\partial \bar{e}} &= \frac{1}{2} + r \cdot \left(\frac{t_0}{t_0 + 1} + \frac{t_1}{t_1 + 1} - 1 \right) - \frac{t_0}{t_0 + 1} \\ &\quad + (2r - 1) \cdot \left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1} \right) \cdot \bar{a}, \end{aligned} \quad (5.170)$$

$$\begin{aligned} \frac{\partial \chi_r(\bar{a}, \bar{e})}{\partial \bar{a}} &= \frac{1}{2} - r + (2r - 1) \cdot \frac{t_1}{t_1 + 1} \\ &\quad + (2r - 1) \cdot \left(1 - \frac{t_0}{t_0 + 1} - \frac{t_1}{t_1 + 1} \right) \cdot \bar{e}. \end{aligned} \quad (5.171)$$

Setting both derivatives to zero yields the strategies.

Inserting \bar{a}_r^* in the partial derivative of the second term of Eq. (5.167) (factor $(1 - r)$), which describes the case when Eve is not able to recover the order of the positions, eliminates all factors containing \bar{e} in this term. The same was already shown for the first term of Eq. (5.167) (factor r) in the proof of Theorem 5.12. Some algebra shows that $\chi_r(\bar{a}, \bar{e}_r^*)$ is independent of \bar{a} as well and thus, with the same arguments as in the proof of Theorem 5.12, $(\bar{a}_r^*, \bar{e}_r^*)$ is a Nash equilibrium. \square

Just as in Theorem 5.13, Alice follows the same strategy as with perfect recoverability, whereas Eve deviates from her strategy. This could indicate that the game-theoretical optimal strategy of the steganographer is invariant to the recovery rate. This highlights again the importance of such a strategy and the superiority of it in comparison to random uniform and naïve adaptive embedding.

As can be seen from the new equilibrium strategies, Eve's strategy is not well-defined for $r = 1/2$. Thus, we handle this case separately.

Corollary 5.18. *In the case of $r = 1/2$ the payoff function $\chi_{\frac{1}{2}}$ is linear in \bar{a} and independent of \bar{e} . By this, Eve cannot influence the payoff. Alice's best strategy is $\bar{a} = 1$, i. e. naïve adaptive embedding. The payoff in equilibrium is always between $\frac{1}{4}$ and $\frac{1}{2}$, and depends only on t_0 .*

Proof. Inserting $r = 1/2$ into Equation (5.167), yields:

$$\chi_{\frac{1}{2}}(\bar{a}, \bar{e}) = \frac{t_1 + 3}{4(t_1 + 1)} + \left(\frac{1}{2(t_0 + 1)} - \frac{1}{2(t_1 + 1)} \right) \bar{a}, \quad (5.172)$$

which is linear in \bar{a} and independent of \bar{e} . Obviously, the slope $\frac{1}{2(t_0+1)} - \frac{1}{2(t_1+1)}$ is positive whenever $t_0 < t_1$. Thus, the maximum is reached at $\bar{a} = 1$ and the payoff is:

$$\frac{1}{4} < \frac{1}{4} + \frac{1}{2(t_0 + 1)} \leq \frac{1}{2}, \quad (5.173)$$

as $t_0 \geq 1$. □

Again, the same as for Corollary 5.17, interpreted for realistic cover sources, the case of $r = \frac{1}{2}$, echoes that if there is no leakage of information through the recoverability of the adaptivity criterion, there is no advantage for Eve if she tries to recover it. So, a cover source with that property would be the best for Alice.

For $r \neq 1/2$, we find that the equilibrium strategies are still equalizer strategies, and the game outcome is the same as in the case of perfect recovery.

Lemma 5.29. *With recovery rate r , the equilibrium strategies are equalizer strategies and the payoff in equilibrium is:*

$$\chi_r(\bar{a}_r^*, \bar{e}_r^*) = \frac{(t_0 + 1)(t_1 + 1) - 4}{4(t_0 t_1 - 1)}. \quad (5.174)$$

Proof. From the proof of Theorem 5.14 follows that the players cannot influence the pay-off when the other player uses her equilibrium strategy. Thus, \bar{a}_r^* and \bar{e}_r^* are equalizer strategies. The payoff follows from combining Equations (5.167), (5.168) and (5.169). □

5.2.4 Numerical Illustrations

In this section we numerically illustrate all the relevant parameters from the previous sections. We exclude the case of imperfect recovery from the illustrations, as we have seen that the payoff is independent of the recovery rate.

5.2.4.1 Numerical Illustration for Linear Increasing PMF

Our analysis of the instantiation where the cover source is modeled with a linear increasing PMF shows that the optimal distribution of embedding changes depends on the level of heterogeneity of the cover source. So, steganographer and steganalyst

both have to adjust their strategy to the cover source. The discussion of our results is facilitated by looking at numerical examples in Figures 5.11 to 5.13. As one requirement for our model was simplicity, we are able to calculate numerically the KLD as benchmark, which is infeasible for real-world cover sources. Furthermore, we show all plots for a cover source with low entropy (Figures 5.11(a), 5.12(a), and 5.13(a)) and with high entropy (Figures 5.11(b), 5.12(b), and 5.13(b)).

Figure 5.11 shows the optimal value of \bar{a}^* , once by numerically minimizing KLD (dashed line) and once the value found in the equilibrium (solid line). Figure 5.12 shows the KLD created by the values for \bar{a}^* from the figure above and Figure 5.13 shows the resulting EER . To recall how the corresponding PMFs look like, please refer to Figure 5.9 (on page 101). Figure 5.13 reveals that if Alice's goal was to minimize KLD, she would choose higher values for \bar{a}^* , i. e., embed with higher probability in the better suitable location. Furthermore, it can be seen in Figure 5.12 that the KLD generated by Alice's strategy in the equilibrium increases rapidly with an increasing level of heterogeneity. Nonetheless, Figure 5.13 shows that Alice's strategy in the equilibrium implicates a higher EER than in the situation with minimal KLD, and thus more secure steganography against the specific detector defined in our model. By this, both players could perform better, if the other would not follow the strategy in the equilibrium. So, it follows that if Alice tries to minimize the KLD and Eve anticipates this (still being bound to her specific detector), Eve's detection rate would increase and thus Alice would perform less secure steganography.

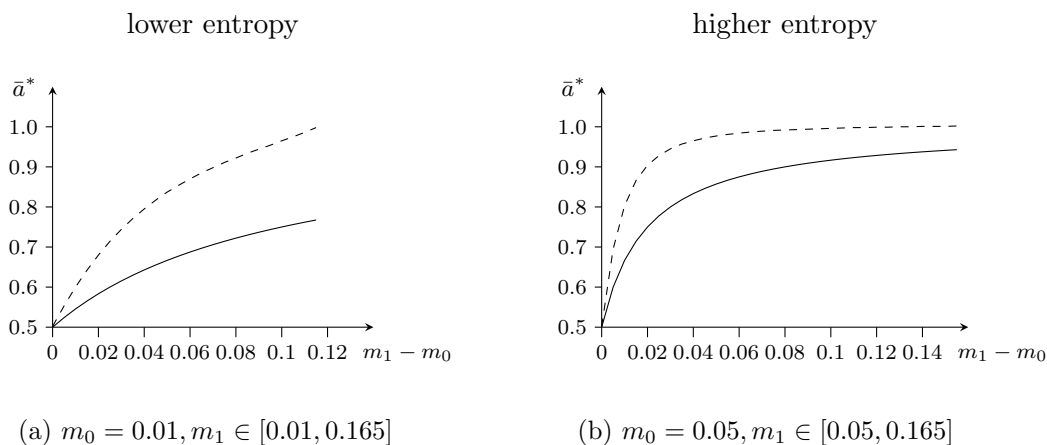


Figure 5.11: Optimal \bar{a}^* once with regard to minimal KLD (dashed line) and once with regard to the equilibrium of our game (solid line).

5.2.4.2 Numerical Illustration for Constant Ratio PMF

Here, we show plots of all important parameters of the instantiation with a cover source that is modeled with a constant ratio PMF, with the restrictions $t_0, t_1 \in [1, 4]$

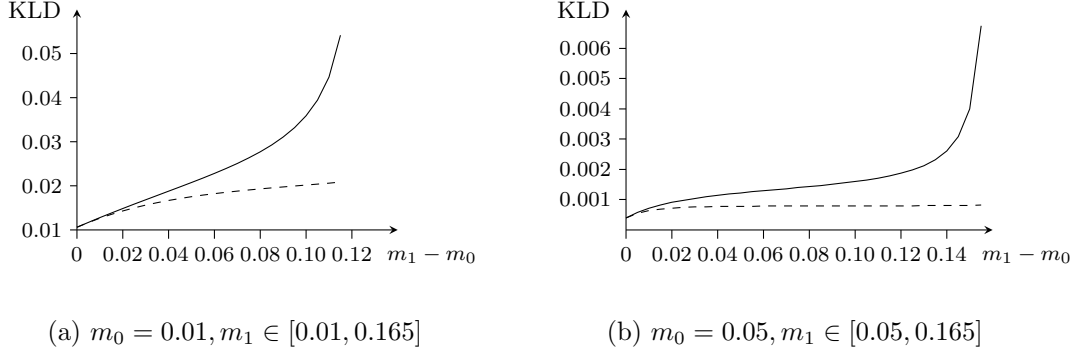


Figure 5.12: Optimal KLD once minimal achievable using LSB replacement (dashed line) and once with regard to \bar{a}^* in the equilibrium of our game (solid line). Note the different scales.

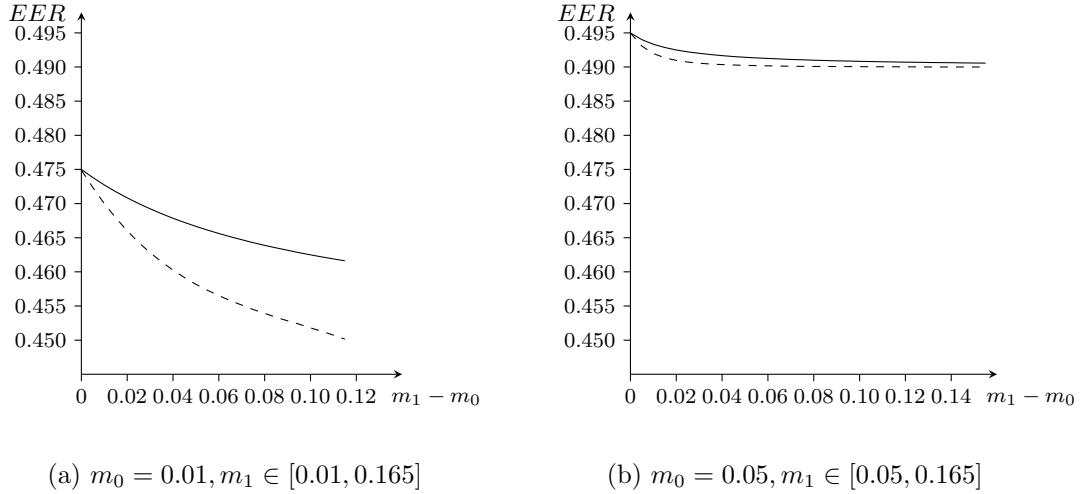


Figure 5.13: Optimal EER with optimal KLD and fixed detector (dashed line) and once in the equilibrium of our game (solid line).

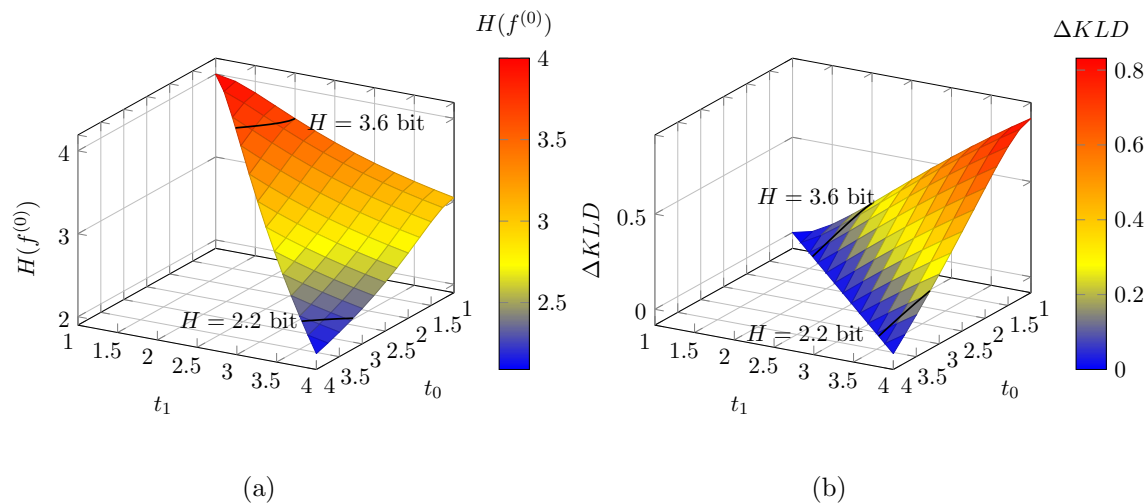


Figure 5.14: Entropy of cover generation as a function of the ratio parameters (left) and the level of heterogeneity, measured by the difference of the Kullback–Leibler divergence, $\Delta KLD = KLD(\mathcal{P}_0, \mathcal{P}_{(y_1)}) - KLD(\mathcal{P}_0, \mathcal{P}_{(y_0)})$ (right).

and $t_0 \leq t_1$. Figure 5.14(a) shows the entropy of the cover source as a function of the ratio parameters t_0 and t_1 . As intended with our model, there are several parameter combinations that yield the same entropy. Figure 5.14(b) depicts the level of heterogeneity $\Delta KLD = KLD(\mathcal{P}_0, \mathcal{P}_{(y_1)}) - KLD(\mathcal{P}_0, \mathcal{P}_{(y_0)})$, i. e., the difference of the KLD. Conforming with our expectations, ΔKLD is zero for a homogeneous cover source and rises for fixed t_0 with increasing t_1 . The black lines in Figure 5.14 indicate the entropy levels used in Figure 5.16. Figure 5.15(a) shows Alice’s optimal strategy \bar{a}^* as a function of t_0 and t_1 . Whenever $t_0 = t_1$, i. e., a homogeneous cover source, the optimal strategy is to use random uniform embedding, illustrating Corollary 5.11. When $t_0 = 1$, i. e., the possibility for perfect steganography, embedding should solely take place in $y_0^{(0)}$ (with the exception if $t_0 = t_1 = 1$, the uniform, homogeneous case), confirming Corollary 5.12. In all other cases, the values for \bar{a}^* are in the interval $(1/2, 1)$. Figure 5.15(b) shows the AER in the equilibrium, derived in Corollary 5.14. It can be seen that if one of the two symbols allows perfect steganography, the AER is always exactly $1/2$. It reaches its minimum for an homogeneous cover with a low entropy. As we fixed LSB replacement as embedding function, which is not the optimal embedding function, we do see different payoffs for cover sources with the same entropy. In Figure 5.16 we fixed the level of the entropy at two levels: high (solid lines) and low (dashed lines). Figure 5.16(a) shows Alice’s equilibrium strategy as a function of the level of heterogeneity and Figure 5.16(b) the equilibrium payoff. We see that for cover sources with low entropy both equilibrium strategy and payoff rise very slowly in comparison to a cover source with high entropy.

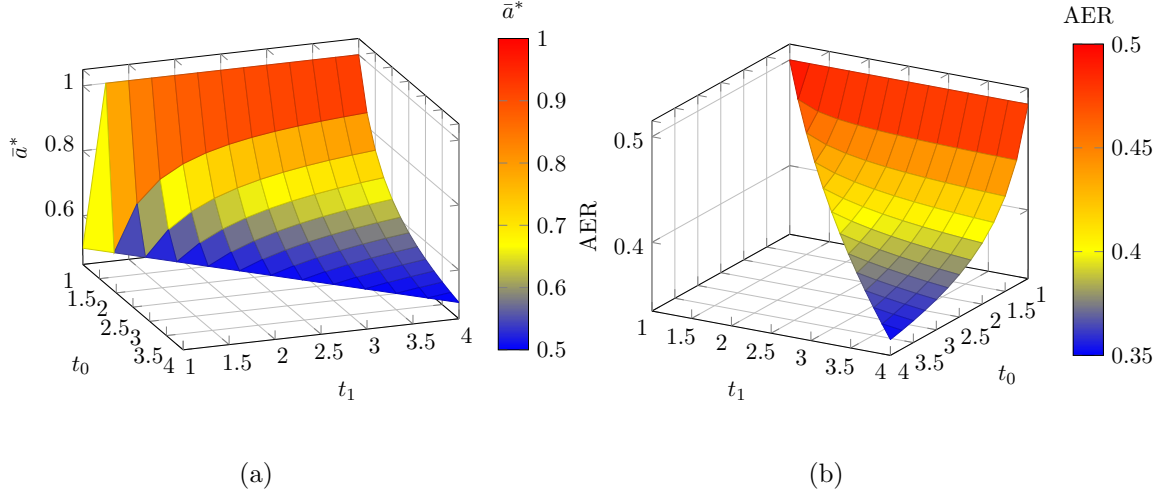


Figure 5.15: Optimal adaptive embedding strategy \bar{a}^* (left) and average error rate (AER) in equilibrium (right).

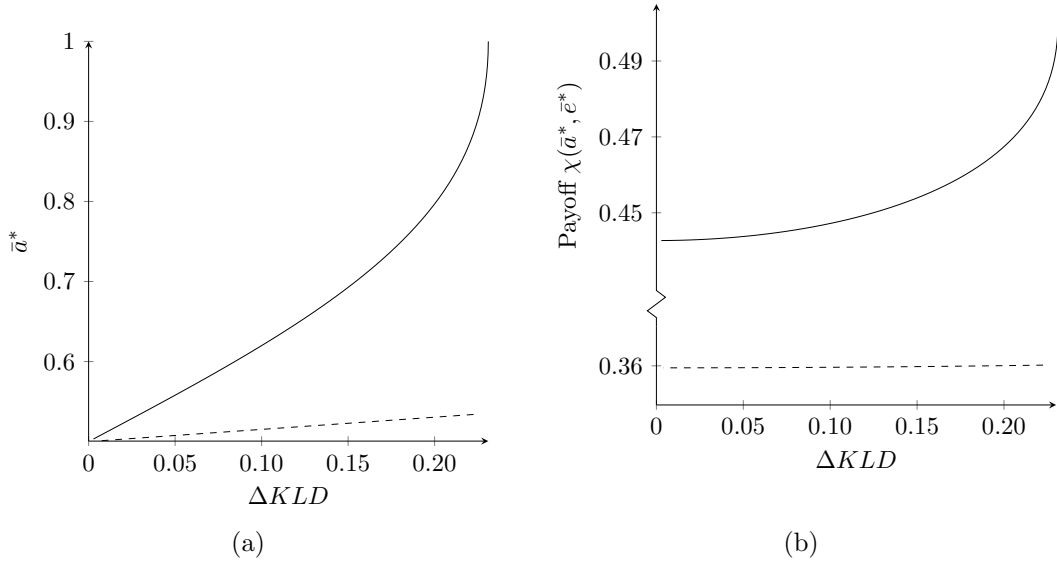


Figure 5.16: Equilibrium strategy \bar{a}^* (left) and equilibrium payoff in the $\chi(\bar{a}^*, \bar{e}^*)$ (right) as a function of the level of heterogeneity with constant entropy once low ($H = 2.2$ bit, dashed line) and once high ($H = 3.6$ bit, solid line).

5.2.4.3 Comparison

When we compare the numerical illustrations of both instantiations with two cover positions, we see several similarities. First, in both models the optimal adaptive

embedding strategy depends on the level of heterogeneity emitted by the cover source. Independent of the entropy of the cover source, for both models random uniform embedding is only optimal for a homogenous cover source, as can be seen in Figures 5.11 and 5.15(a). Furthermore, with increasing heterogeneity, the preference of the more suitable embedding position increases. Nonetheless, we also see that naïve adaptive embedding is never optimal.

Comparing Figures 5.11 and 5.16(a) reveals that for cover sources with a higher entropy the optimal adaptive embedding strategy rises faster than for a cover source with a low entropy. Finally, regarding the payoff, Figures 5.13 and 5.16(b) confirm that Alice profits most from a very heterogeneous cover source.

5.2.5 Type of Game

The purpose of this section is to facilitate the classification of the games introduced so far into the game-theoretical literature. A large part of this section originates from the correspondence with the anonymous reviewers of [78].

As introduced in Definition 4.2.1, a game with incomplete information, also called a Bayesian game, is characterized by the uncertainty of some, or all of the players about either payoffs or strategies of the other players. As we explicitly deal with uncertainty, although in a different context, it might suggest itself that we should model our game as a Bayesian game. In this section we explain how we circumvent the analysis of Bayesian Nash equilibria and why we are able to concentrate on (ordinary) Nash equilibria in our set-up.

According to the definition by Harsanyi [39], a Bayesian game is characterized as a

“... game with incomplete information where the players are uncertain about some important parameters of the game situation, such as payoff functions, the strategies available to various players, the information other players have about the game, etc. However, each player has a subjective probability distribution over the alternative possibilities.”

In principle, the games in this section can be formulated as Bayesian games and thus solved for Bayesian Nash Equilibria. Our cover generation process, Nature, draws covers according to the probability distribution \mathcal{P}_0 . For game-theoretic tractability, this distribution is assumed to be common knowledge. Together with the knowledge that the Judge forwards cover and stego objects with probability 1/2, we obtain the set of all possible types as the cross product of all possibilities. However, as we do not restrict the alphabet size of the cover symbols $(0, \dots, 2^\ell - 1)$ in the general framework, the number of different values drawn by Nature can be very large.

We use a different approach and incorporate the common knowledge about \mathcal{P}_0 , the embedding operation (hence, \mathcal{P}_1) and the Judge’s equal priors in Eve’s local decision rule. In Lemmas 5.74 and 5.23, we allow Eve to make locally (information-theoretically) optimal decisions. We obtain this optimality by the *maximum a posteriori* (MAP) estimation [27]. MAP estimation is the method of choice, not only in steganography, to

estimate the generating distribution of a new observation, if we have prior knowledge about the distributions ($\mathcal{P}_0, \mathcal{P}_1$, and the Judge in our case). As we assume the cover symbols to be a priori independent, the problem of finding the best statistical test for our simple models can be isolated from the problem of choosing embedding and detection positions.¹⁷

Technically, we aggregate the two *non-strategic* sources of randomness (Nature and Judge) and replace them with the first moment when calculating the *EER* in Lemmas 5.20 and 5.26, and the *AER* in Lemmas 5.25 and 5.28. *EER* and *AER* are common and valid metrics for steganographic security (Sec. 2.2.3). Averaging success rates over the realizations of Nature and Judge does not affect the payoff of risk-neutral players.

After having taken care of non-strategic randomness, we are left with the *strategic* choices of Alice and Eve, i. e., which position to embed and examine, respectively. In our model with the two positions y_0 and y_1 , this leaves 4 possible combinations, for which we calculate the error rates explicitly by using Eve’s decision function DR. Another source of *strategic* randomness is introduced by allowing mixed strategies in the resulting zero-sum matrix game with payoffs (“rates”) as given in Table 5.7.

For a fixed instance of the cover generation source, the parameters m_0 and m_1 , or t_0 and t_1 , respectively, are fixed as well. Thus, we have fixed rates ($\frac{1}{2} - m_i$ or $\frac{1}{t_i+1}$) in the cases where $\bar{a} = \bar{e} \in \{0, 1\}$.

Table 5.7: Payoff matrices of the bi-matrix games

(a) EER of linear increasing PMF				(b) AER of constant ratio PMF			
		Eve				Eve	
		y_0	y_1			y_0	y_1
Alice		$\frac{1}{2} - m_0$	$1/2$	Alice		$\frac{1}{t_0+1}$	$1/2$
		$1/2$	$\frac{1}{2} - m_1$			$1/2$	$\frac{1}{t_1+1}$

These games can be solved for Nash equilibria in mixed strategies, as a quick glance at the payoff matrices shows that there is no equilibrium in pure strategies in the general case of $0 < m_0 < m_1$, or $1 < t_0 < t_1$ (i. e., with strict inequalities).

To summarize the above: our set-up starts as a game with *incomplete information*, i. e., a Bayesian game: the players are uncertain about the cover realization. By introducing Nature and the Judge, we use the Harsanyi transformation [39] to rewrite the game as a game with *imperfect information*. Finally, aggregating the probability distributions of Nature and the Judge to a (frequentist) rate, the *EER*, or *AER*, transforms the set-up to a *simultaneous move game with perfect information*.

¹⁷Finding these tests for more practical scenarios is the subject of another stream of research, e. g. [18].

The games we introduce are characterized by Alice's objective to minimize the information flow to Eve. As the amount of available information is endogenous in our set-up, we do not have discrete information sets like in classical game theory. Our games might constitute a new class of games that could be called *information hiding games*.

5.2.6 Discussion and Summary

Summarizing this section, we have proven that

- ▶ both adaptive steganography games with two positions have a unique symmetric Nash equilibrium in equalizer strategies (Theorem 5.11 and Corollary 5.8 from Section 5.2.1; and Theorem 5.12 and Corollary 5.15 from Section 5.2.2)
- ▶ random uniform embedding is only optimal for homogeneous covers (Corollary 5.4 and Corollary 5.11); and
- ▶ naïve adaptive embedding is only optimal when perfect steganography is possible (Corollary 5.5 and Corollary 5.12).

The optimal embedding parameter for heterogeneous covers depends on the level of heterogeneity, albeit in a non-linear manner. Although the dependence is not linear, it is monotone in the amount of heterogeneity, i.e., the more heterogeneous the cover source is, the higher is Alice's preference for the better suitable position.

Remark 5.6. *In our model, the probability distribution of the cover source determines the location of the game-theoretic equilibrium and thus the equalizer strategy. As this distribution is unknowable in practice, Alice and Eve may not be able to find the exact equilibrium strategy.*

Furthermore, we have shown that the concept of equalizer strategies extends to the model with imperfect recovery. We show that both players' payoff does not depend on the recovery rate. It is solely determined by the heterogeneity emitted by the cover source. However, the recovery rate may matter if Alice deviates from (or does not know) her optimal strategy.

It is interesting to note that, excluding the corner case $r = \frac{1}{2}$, the equilibrium payoff of the game is independent of r . As the concept of equalizer strategies makes the players indifferent to the opponents' action, we conjecture that the equilibrium strategies completely balance the local advantage.

Furthermore, concerning steganographic security, we state the following corollary about the recovery rate r .

Corollary 5.19. *If Alice is able to find an equalizer strategy, the recovery rate is of no interest to her.*

Proof. With Alice playing an equalizer strategy, Eve cannot gain from a better or worse recoverability. Thus, it is of no interest to Alice. \square

This result is remarkable, as there might be realistic scenarios where an equalizer strategy is feasible and, for the reasons stated here, very desirable. This solution concept

has, to the best of our knowledge, not found much attention in the field of steganography research. Interesting open questions for further research are, if the equalizer strategies are always optimal for Alice and under which conditions do they exist.

5.3 Lessons Learned and Limitations

In this section we summarize the findings from all instantiations of our game-theoretical framework and show what we can learn from all of them regarding the construction of more secure adaptive steganography. Then, we present the limitations of our approaches in order to highlight what should not yet be concluded from our results and what are the most prominent research gaps that have to be closed to leverage our results.

5.3.1 Lessons Learned – Secure Adaptive Steganography

In this section we tie the findings from all instantiations of our game-theoretical framework to the definitions of steganographic side information and uncertainty from Section 3.1. From this we can deduce optimal strategies in different scenarios and, finally, outline directions towards a new embedding paradigm. We differentiate optimal strategies depending on the availability of perfect SSI and perfect uncertainty. Especially, the concept of equalizer strategies and the balancing of the opponent’s advantage induces interesting properties even for realistic cover sources where the generating distribution is unknown. Making the opponent indifferent to the own actions would automatically lead to an optimal embedding strategy.

Definition 5.4 (Optimal Embedding Strategies).

The steganographer’s embedding strategy is called optimal for ...

1. perfect uncertainty, *if she spreads the embedding changes uniformly across the perfectly uncertain positions,*
2. perfect steganographic side information, *if she chooses deterministically the most uncertain positions resulting from this perfect SSI, and*
3. neither perfect uncertainty nor perfect steganographic side information, *if she equalizes the steganalyst’s advantage over positions.*

Each of these definitions requires some reflection.

Remark 5.7 (Presence of Perfect Uncertainty). *If we have perfect uncertainty for at least k positions, we spread our embedding changes uniformly across these positions. If we have exactly k perfectly uncertain positions, we deterministically use them. In this case it does not matter whether the SSI we used to identify the embedding positions is reconstructible or not, as Eve might perfectly know in which positions we embed but cannot gain from that knowledge.*

This remark is supported by the game-theoretical findings in Section 5.1 when $f(i) = 1/2$ in binary sequences of length n and the corner cases $m_0 = 0$ and $t_0 = 1$ in Sections 5.2.1 and 5.2.2, respectively. Additionally, this remark can be tied to the cover composition model from [8], where the indeterministic part resembles our notion of perfectly uncertain positions (cf. Section 3.1.1.2), and where it is argued that the embedding changes should be confined solely to this indeterministic part.

Remark 5.8 (Presence of Perfect SSI). *If we have (unconditionally) perfect steganographic side information, Eve will get no information about the rule according to which we select the embedding positions. Thus, we choose deterministically the most uncertain positions resulting from this perfect SSI, i.e., we perform naïve adaptive embedding.*

This remark is backed up by the cases of $r = 1/2$ in Section 5.2.3, where the payoff was independent of Eve’s choice and was linear increasing in Alice’s choice. This states the obvious, if Eve cannot recover the order, she will not gain from trying so.

Remark 5.9 (Absence of Perfect Uncertainty and Perfect SSI). *If we can ensure neither perfect SSI nor enough¹⁸ perfectly uncertain positions, we have to randomize the selection of embedding positions over all positions, excluding only the perfectly informative positions. We deduce the embedding probability for single positions by taking the uncertainty, or suitability, into account. The goal in this situation is to create a uniform advantage for Eve by assigning a higher probability to more suitable embedding positions and a lower change probability to less suitable positions. This is the translation of the solution concept of equalizer strategies to empirical embedding functions.*

It is questionable if there exists either perfect SSI or perfect uncertainty for empirical cover sources. If they exist, game theory is not necessary to find the optimal strategies, as there would be no competition between Alice and Eve. Alice would always win. But, similar to the argument that \mathcal{P}_0 and \mathcal{P}_1 are unknowable in practice [7], it seems reasonable that they do not exist or at least that they are not tractable for real-world cover sources.

If we accept this, we should follow the solutions in our game-theoretical instantiations and equalize Eve’s advantage over all positions. Although our equilibrium strategies do depend on the knowledge of the cover distribution or the KLD between \mathcal{P}_0 and \mathcal{P}_1 , we can construct a new embedding paradigm called *equalizer embedding strategy* for empirical cover sources, given (partially) reconstructible SSI Θ and an adaptivity criterion $\zeta(\cdot, \theta)$ whose values are tied to uncertainty. We believe that these strategies, although they might not be perfectly optimal, help a steganographer to perform more secure adaptive steganography.

Definition 5.5 (Equalizer Embedding Strategy).

A steganographer uses an equalizing embedding strategy, depending on an adaptivity criterion $\zeta(\cdot, \theta)$, when she calculates the change probability λ_i of each position $x_i^{(0)}$ from

¹⁸In this context “enough” means that we have less than the k perfectly uncertain embedding positions.

a given cover realization $\mathbf{x}^{(0)}$ by:

$$\forall i \in \{0, \dots, n-1\} : \lambda_i = \zeta(x_i^{(0)}, \theta_i)^{-1}/d, \quad (5.175)$$

where d is a normalizing constant to ensure that $\sum_{i=0}^{n-1} \lambda_i = k$, the bit-length of the (encoded) message.

Such a strategy ensures a uniform advantage for Eve, as it holds that:

$$\forall i, j \in \{0, \dots, n-1\} : \lambda_i \cdot \zeta(x_i^{(0)}, \theta_i) = \lambda_j \cdot \zeta(x_i^{(0)}, \theta_j). \quad (5.176)$$

This strategy should always be followed when we cannot guarantee either enough perfectly uncertain embedding positions or unconditionally perfect SSI for every realization of the cover source. Furthermore, if all positions are equally uncertain, it ensures random uniform embedding. This is in line with our game-theoretical results, where the optimal strategy for homogeneous cover sources is random uniform embedding.

Figures 5.17 and 5.18 visualize the change probability and Eve's advantage for different embedding strategies, respectively. The solid blue line depicts an equalizing embedding strategy, the dashed blue line naïve adaptive embedding and the dotted blue line shows random uniform embedding. The red line shows exemplary values of an adaptivity criterion $\zeta(\cdot, \boldsymbol{\theta})$ where lower values indicate better suitability.

Finally, we like to note that the concept of equalizer embedding strategies has occurred in earlier game-theoretical models of steganography, although it was not recognized as such by the respective authors. In the example from Section 4.3.1, the author states that Eve's equilibrium strategy "consists of equalizing as many of the lowest order effective channel capacities as allowed by the distortion limit." [17, p. 9] Then, for the batch steganography example from Section 4.3.2, the optimal strategy to spread the secret message as thinly as possible over all cover objects is nothing else than an equalizer embedding strategy.

5.3.2 Limitations

One of the most obvious limitations of our game-theoretical models is that all of them rely on a number of assumptions. The models in Section 5.1 rely on the assumptions that covers consist of binary objects of arbitrary length and that the steganographer has full access to the side information that determines the suitability. Furthermore, the steganalyst can either query the side information or is able to perform a maximum likelihood test between the cover and the stego distribution. The models in Section 5.2 assume that covers consist of only two a priori independent positions and both players know the marginal cover distribution. Additionally, we assume that the steganographer replaces only one bit, while the steganalyst only inspects one position. So, many limitations apply when transferring our results to practical systems. We nevertheless think that a solid theory might help to guide the design of future embedding and detection functions with qualitative insights. One of the biggest research challenges towards this end seems to be the incorporation of non-trivial dependence structures in

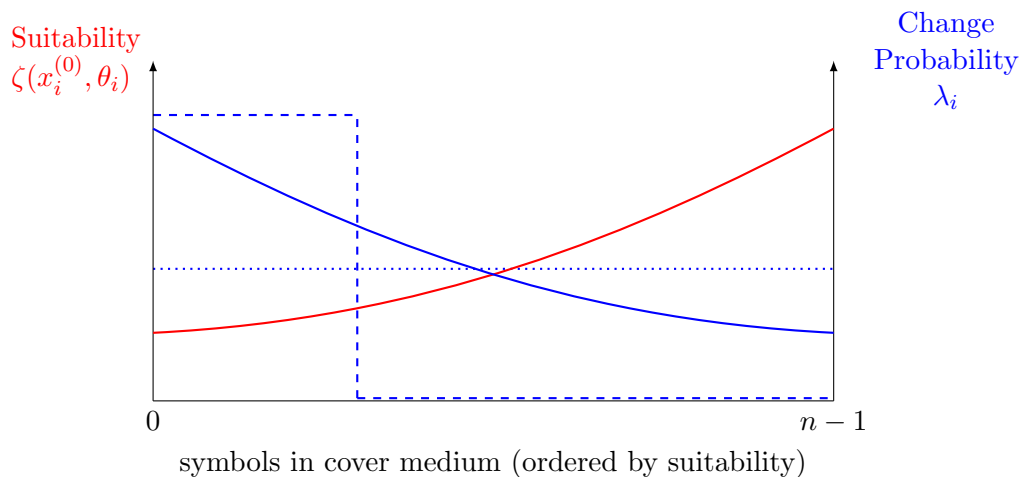


Figure 5.17: Comparison of change probabilities for the embedding strategies: equalizer embedding strategy (solid blue line), naïve adaptive embedding (dashed blue line), and random uniform embedding (dotted blue line). The red line shows the values of an adaptivity criterion $\zeta(\cdot)$ (lower values indicate better suitability).

the cover model. First observations in this directions show that such a game is more complex and the location of the game theoretical equilibrium might be hard to identify.

Then, in all the models we examined, the location of the game theoretical equilibrium, and thus the equalizer embedding strategy, depends on the parameters of the cover source \mathcal{P}_0 . Although, we believe that these equilibria exist in realistic settings, i.e., without the limiting assumptions mentioned in the last paragraph, it still might be computationally hard to find their exact location. We believe that even if the exact location of the equalizer strategies is not tractable, already an approximation might lead to more secure embedding functions. The same holds for steganalysis: even when the game theoretical optimal detection strategy is not known, the incorporation of the knowledge about likely embedding positions should always increase the detection performance, as suggested by recent results [81].

Another assumption we used in most of our models is that not only the cover distribution but also the exact values of the suitability are known. As already mentioned in Definition 4.16, all practical embedding algorithms use adaptivity criteria as an approximation of the authors' knowledge about how suitable embedding positions are. The incorporation and consequences of imperfect adaptivity criteria has to be taken into account to furthermore bring our game theoretical models one step closer to reality.

Finally, most of nowadays steganalysis methods rely on machine learning and high-dimensional feature sets. The adaption and validation of our framework for these detectors remains an open issue. It might be possible that these detectors consider adaptivity by default, or even intuitively adapt to a steganographer who plays an equalizer strategy. But, we cannot be sure about that before we find a way to incorporate them into the framework or find a way to implement such an equalizer strategy for

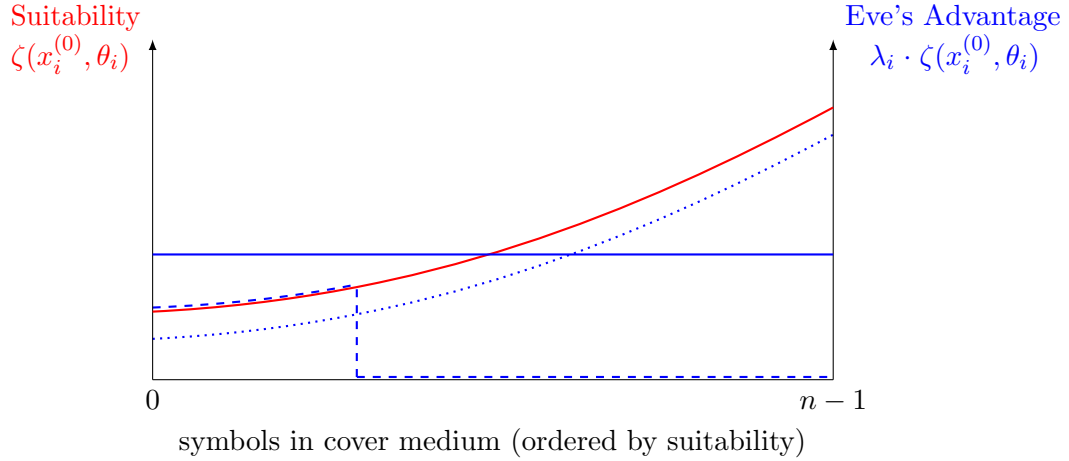


Figure 5.18: Comparison of Eve's advantage for the embedding strategies: equalizer embedding strategy (solid blue line), naïve adaptive embedding (dashed blue line), and random uniform embedding (dotted blue line). The red line shows the values of an adaptivity criterion $\zeta(\cdot)$ (lower values indicate better suitability).

realistic settings and benchmark it against machine-learning based detectors.

In general, we regard this stream of work as a step towards adding more theoretical rigor to practical steganography and steganalysis.

Chapter 6

Conclusion

6.1 Summary of Results

The contribution of this thesis is twofold. We establish the formalization of side-informed steganography and a game-theoretical analysis of this situation. The following sections summarize our main results.

6.1.1 Formalizing Side Information in Steganography

To the best of our knowledge, we are the first to formally define *steganographic side information* (SSI) (cf. Definition 3.1) as a source of information fully available to the steganographer to enhance her embedding strategy. This definition captures all possibilities in a steganographic communication system to utilize side information during the embedding. We hope to remove the inconsistent use of the term side information in the research community with our definition. Furthermore, we differentiate between unconditionally and conditionally perfect SSI, which is exclusively available to the steganographer, and (partially) reconstructible SSI, to explicitly state if, and under what conditions, the SSI is also available to the steganalyst.

We relate the definition of SSI to the definition of *uncertainty* (cf. Definition 3.2) on the side of the steganalyst, that quantifies the lack of knowledge about which type of object, cover or stego, the steganalyst faces. This definition is strongly tied to steganographic security. We give information-theoretic intuitions of both definitions and extend them to practical content-adaptive embedding schemes with the introduction of the *recovery rate* (cf. Definition 3.7). The recovery rate measures the amount of embedding positions, which the steganalyst can recover from a stego object (cf. Section 3.1) with the SSI available to her.

We argue that the common usage of SSI in the development of side-informed embedding is as if it was perfect SSI. To show that this rarely (or never) is the case, we develop a targeted attack against several widely used adaptivity criteria. We modify a powerful variant of WS steganalysis for the detection of initial sequential embedding to construct a version that reliably detects naïve adaptive embedding, i.e., the strategy that places all the embedding changes in the best suitable positions (cf. Section 3.2.2.2). Our experiments suggest a superior performance of our WS variant. This variant detects all four tested adaptivity criteria with very high accuracy (cf. Section 3.3.6).

In general, we conclude that steganographers should avoid naïve adaptive embedding and take the recovery rate of their adaptivity criteria into account. New proposals of embedding schemes should thus include a detailed examination of the recovery rate to preclude trivial targeted attacks. Our experiments furthermore suggest that, at least for

our four examined adaptivity criteria, the recovery rate is approximately constant for each criterion tested. Thus, the recovery rate can be seen as a property of the adaptivity criterion.

6.1.2 Game-Theoretical Modeling of Steganography

Inspired by the results of our targeted attack against naïve adaptive embedding and the obvious assumption that any rational steganographer would change her strategy against an informed steganalyst, we conclude that this contest is best framed with the help of game theory. Although there are occasional attempts to frame steganography with game theory, they are mostly tailored to very specific scenarios [17, 50, 69].

We present a game-theoretic framework that captures all relevant properties of a steganographic set-up with the availability of side information (cf. Section 4.4). We give rigorous definitions of heterogeneous cover sources (cf. Definition 4.14), suitability (cf. Definition 4.15) and an adaptivity criterion (cf. Definition 4.16) in our framework. We then identify the canonical embedding and detection strategies and their information-theoretic optimal counterparts.

We then continue to instantiate our framework with specific heterogeneous cover sources and determine the game-theoretical solutions in each instantiation (cf. Sections 5.1 and 5.2). In total five different instantiations resemble different models about the knowledge and power of the steganographer and steganalyst, respectively. The results of all instantiations share that the steganalyst will take all possible embedding positions into account. The classical embedding strategies of naïve adaptive and random uniform embedding are only viable in degenerate corner cases, i.e., perfectly uncertain embedding positions or homogeneous cover sources.

Furthermore, based on instantiations with only two available embedding positions, we show that our game-theoretic optimal embedding strategies differ from the information-theoretic optimal ones. Information-theoretic optimal embedding has traditionally been the “holy grail” in steganographic research, due to the information-theoretic security definition. Our results indicate that a steganographer who follows a game-theoretic optimal embedding strategy may rather think of information-theoretic optimal embedding as an ordinary wine glass.

Also, other researchers that followed us in studying game-theoretical models in steganography agree with us, in that “[...] the KL divergence is no longer an appropriate measure of security and Alice’s optimal embedding strategy should be determined from a framework based on the game theory.” [14, p. 902804-12]

Finally, we argue why the concept of the so-called *equalizer strategies* can induce important aspects for practical embedding schemes and highlight what has to be considered in the design of new side-informed embedding functions. If perfect SSI or perfect uncertainty can be guaranteed, we do not need a game-theoretic model and the classical embedding strategies are sufficient. If we can guarantee neither, we propose a new paradigm for secure steganography, called *equalizer embedding strategy* (cf. Definition 5.5). This way, we translate our game-theoretical optimal embedding strategies to a practical embedding function. With an equalizer embedding strategy,

the level of uncertainty per position, measured by an imperfect adaptivity criterion, and the probability to change this position for embedding leverage each other out and create a uniform local advantage on the side of the steganalyst.

6.2 Outlook and Future Research

There are several ways in how to foster the results of this thesis in future research.

Considering the targeted attack on naïve adaptive embedding, one could elaborate the knowledge on how the embedding operation changes single positions to come up with a more accurate estimation. We restricted our attacks to the simplest form of recalculating the value of the adaptivity criterion for the positions in the stego object.

Another direction would be to take imperfect recovery into account and develop a version of our initial attack strategy that does tolerate some “gaps” in the order of the recoverability. If we can achieve this, slightly randomized versions of naïve adaptive embedding could be detected with a higher accuracy.

Our game-theoretical framework has established a new direction in steganography research that is now widely accepted as one of “today’s most interesting research challenges” [55, p. 8] and “[game theory is] an alternative and appealing possibility to formally capture the sender’s and Warden’s ignorance” [28, p. 361].

But, there is still a lot of space for advances in the game-theoretic analysis of steganography in the future. For all instantiations, we were only able to solve the equilibrium strategies for reduced versions of the cover source that emit cover objects of length 2. So, one of the main aspects of future work here is, as stated in a position paper by the leading researchers in steganography in last year’s most important conference on information hiding and multimedia security (IH &MMSec 2013):

“Find equilibria for practical covers, and transfer insights of game-theoretic solutions from current toy models to the real world.” [58, p. 54]

For real-world embedding strategies, first and foremost, the utilization of an *equalizer embedding strategy* would be of great interest. It will be interesting to see how such embedding strategies perform against machine learning-based steganalysis. It is commonly assumed, although it has never been proven, that this blind steganalysis implicitly recognizes adaptive embedding and anticipates it, given the right features. The exact functionality of machine learning-based steganalysis is not well understood and often treated as some kind of *black box* in the steganography literature.

We finally mention that game theory research discusses plenty of extensions to the basic games which could be of great interest to the steganography community. Most prevalent are repeated games, as it would be very interesting to see if the “square-root law of steganographic capacity” [57] could be confirmed with means of game theory. Further approaches of interest are, amongst others, blocking games [37] and search games [75].

Bibliography

- [1] AMIRI, E. & TARDOS, G. (2009). High rate fingerprinting codes and the fingerprinting capacity. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, 336–345, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- [2] ANDERSON, R.J. (1996). Stretching the limits of steganography. In R.J. Anderson, ed., *Information Hiding (1st International Workshop)*, vol. 1174 of *Lecture Notes in Computer Science*, 39–48, Springer-Verlag, Berlin Heidelberg.
- [3] ANDERSON, R.J. & PETITCOLAS, F.A.P. (1998). On the limits of steganography. *IEEE Journal on Selected Areas in Communications*, **16**, 474–481.
- [4] BARNI, M. & TONDI, B. (2013). The source identification game: An information-theoretic perspective. *IEEE Transactions on Information Forensics and Security*, **8**, 450–463.
- [5] BAS, P., FILLER, T. & PEVNÝ, T. (2011). Break our steganographic system — the ins and outs of organizing BOSS. In T. Filler, T. Pevný, S. Craver & A. Ker, eds., *Information Hiding (13th International Workshop)*, vol. 6958 of *Lecture Notes in Computer Science*, 59–70, Springer-Verlag, Berlin Heidelberg.
- [6] BÖHME, R. (2008). Weighted stego-image steganalysis for JPEG covers. In K. Solanki, ed., *Information Hiding (10th International Workshop)*, vol. 5284 of *Lecture Notes in Computer Science*, 178–194, Springer-Verlag, Berlin Heidelberg.
- [7] BÖHME, R. (2009). An epistemological approach to steganography. In S. Katzenbeisser & A.R. Sadeghi, eds., *Information Hiding (11th International Workshop)*, vol. 5806 of *Lecture Notes in Computer Science*, 15–30, Springer, Berlin Heidelberg.
- [8] BÖHME, R. (2010). *Advanced Statistical Steganalysis*. Springer, Berlin Heidelberg.
- [9] BÖHME, R. & KIRCHNER, M. (2013). Counter-forensics: Attacking image forensics. In H.T. Sencar & N.D. Memon, eds., *Digital Image Forensics*, 327–366, Springer Berlin Heidelberg.
- [10] BÖHME, R. & WESTFELD, A. (2004). Exploiting preserved statistics for steganalysis. In J. Fridrich, ed., *Information Hiding (6th International Workshop)*, vol. 3200 of *Lecture Notes in Computer Science*, 82–96, Springer Berlin Heidelberg.
- [11] CACHIN, C. (2004). An information-theoretic model for steganography. *Information and Computation*, **192**, 41–56.
- [12] CHIA, P.H. & CHUANG, J. (2011). Colonel blotto in the phishing war. In *Decision and Game Theory for Security*, 201–218, Springer Berlin Heidelberg.
- [13] DALVI, N., DOMINGOS, P., MAUSAM, SANGHAI, S. & VERMA, D. (2004). Adversarial classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 99–108, ACM, New York, NY, USA.
- [14] DENEMARK, T. & FRIDRICH, J. (2014). Detection of content adaptive LSB matching: a game theory approach. In *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics*, vol. 9028, 902804–902804–12.

- [15] DENEMARK, T., FRIDRICH, J. & HOLUB, V. (2014). Further study on the security of SUNIWARD. *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics*, **9028**, 2–6.
- [16] DUMITRESCU, S., WU, X. & WANG, Z. (2003). Detection of LSB steganography via sample pair analysis. *IEEE Transactions on Signal Processing*, **51**, 1995–2007.
- [17] ETTINGER, M. (1998). Steganalysis and game equilibria. In D. Aucsmith, ed., *Information Hiding (2nd International Workshop)*, vol. 1525 of *Lecture Notes in Computer Science*, 319–328, Springer, Berlin Heidelberg.
- [18] FILLATRE, L. (2012). Adaptive steganalysis of least significant bit replacement in grayscale natural images. *IEEE Transactions on Signal Processing*, **60**, 556–569.
- [19] FILLER, T. & FRIDRICH, J. (2009). Complete characterization of perfectly secure stego-systems with mutually independent embedding operation. In *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1429–1432, IEEE Computer Society, Washington, DC, USA.
- [20] FILLER, T., KER, A.D. & FRIDRICH, J. (2009). The square root law of steganographic capacity for markov covers. In E.J. Delp III, J. Dittmann, N.D. Memon & P.W. Wong, eds., *Media Forensics and Security*, vol. 7254, 725408–725408, SPIE.
- [21] FILLER, T., JUDAS, J. & FRIDRICH, J. (2010). Minimizing embedding impact in steganography using trellis-coded quantization. In N.D. Memon, J. Dittmann, A.M. Alattar & E.J. Delp III, eds., *Media Forensics and Security II*, vol. 7541, 754105, SPIE.
- [22] FILLER, T., JUDAS, J. & FRIDRICH, J. (2011). Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, **6**, 920–935.
- [23] FRANZ, E. (2003). Steganography preserving statistical properties. In F. Petitcolas, ed., *Information Hiding (5th International Workshop)*, vol. 2578 of *Lecture Notes in Computer Science*, 278–294, Springer Berlin Heidelberg.
- [24] FRANZ, E. & PFITZMANN, A. (2000). Steganography secure against cover–stego-attacks. In A. Pfitzmann, ed., *Information Hiding (3rd International Workshop)*, vol. 1768 of *Lecture Notes in Computer Science*, 29–46, Springer Berlin Heidelberg.
- [25] FRANZ, E. & SCHNEIDEWIND, A. (2004). Adaptive steganography based on dithering. In *Proceedings of the 2004 workshop on Multimedia and security, MM & Sec '04*, 56–62, ACM, New York, NY, USA.
- [26] FRIDRICH, J. (2006). Minimizing the embedding impact in steganography. In *Proceedings of ACM Multimedia and Security Workshop (MM&SEC)*, 2–10, ACM, New York, NY, USA.
- [27] FRIDRICH, J. (2009). *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, New York, NY, USA, 1st edn.
- [28] FRIDRICH, J. (2013). Effect of cover quantization on steganographic fisher information. *IEEE Transactions on Information Forensics and Security*, **8**, 361–373.
- [29] FRIDRICH, J. (2013). On the role of side information in steganography in empirical covers. In *IS&T/SPIE Electronic Imaging*, 86650I–86650I, International Society for Optics and Photonics.

-
- [30] FRIDRICH, J. & DU, R. (2000). Secure steganographic methods for palette images. In A. Pfitzmann, ed., *Information Hiding*, vol. 1768 of *Lecture Notes in Computer Science*, 47–60, Springer.
- [31] FRIDRICH, J. & GOLJAN, M. (2004). On estimation of secret message length in LSB steganography in spatial domain. In E.J.D. III & P.W. Wong, eds., *Security, Steganography, and Watermarking of Multimedia Contents VI*, vol. 5306, 23–34, SPIE.
- [32] FRIDRICH, J., GOLJAN, M. & DU, R. (2001). Detecting LSB steganography in color and gray-scale images. *IEEE Multimedia*, **8**, 22–28.
- [33] FRIDRICH, J., GOLJAN, M. & SOUKAL, D. (2004). Perturbed quantization steganography with wet paper codes. In *Proceedings of ACM Multimedia and Security Workshop (MM&SEC)*, 4–15, ACM, New York, NY, USA.
- [34] FRIDRICH, J., GOLJAN, M. & SOUKAL, D. (2004). Searching for the stego-key. In *Electronic Imaging 2004*, 70–82, International Society for Optics and Photonics.
- [35] FRIDRICH, J., KODOVSKÝ, J., HOLUB, V. & GOLJAN, M. (2011). Breaking HUGO—the process discovery. In *Information Hiding (13th International Workshop)*, 85–101, Springer Berlin Heidelberg.
- [36] GLOE, T. & BÖHME, R. (2010). The ‘Dresden image database’ for benchmarking digital image forensics. In *ACM Symposium on Applied Computing*, 1584–1590, ACM.
- [37] GUEYE, A. (2011). *A Game Theoretical Approach to Communication Security*. Ph.D. thesis, University of California, Berkeley, Electrical Engineering and Computer Sciences.
- [38] GUO, L., NI, J. & SHI, Y.Q. (2014). Uniform embedding for efficient JPEG steganography. *IEEE Transactions on Information Forensics and Security*, **9**, 814–825.
- [39] HARSANYI, J.C. (1967). Games with incomplete information played by “Bayesian” players, i-iii part i. the basic model. *Management science*, **14**, 159–182.
- [40] HEMPSTALK, K. (2006). Hiding behind corners: Using edges in images for better steganography. In *Proc. Computing Women’s Congress, Hamilton, New Zealand*.
- [41] HOLUB, V. & FRIDRICH, J. (2012). Designing steganographic distortion using directional filters. In *4th IEEE International Workshop on Information Forensics and Security (WIFS 2012)*, 234–239, IEEE.
- [42] HOPPER, N., LANGFORD, J. & VON AHN, L. (2002). Provably secure steganography. In M. Yung, ed., *Advances in Cryptology – CRYPTO 2002*, vol. 2442 of *Lecture Notes in Computer Science*, 119–123, Springer, Berlin Heidelberg.
- [43] HUSSAIN, M. & HUSSAIN, M. (2011). Embedding data in edge boundaries with high PSNR. In *7th International Conference on Emerging Technologies*, 1–6.
- [44] INUSAH, S. & KOZUBOWSKI, T.J. (2006). A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, **136**, 1090–1102.
- [45] JOHNSON, B., SCHÖTTLE, P. & BÖHME, R. (2012). Where to hide the bits? In J. Grossklags & J. Walrand, eds., *GameSec 2012*, no. 7638 in *Lecture Notes in Computer Science*, 1–17, Springer, Berlin Heidelberg.

- [46] JOHNSON, B., SCHÖTTLE, P., LASZKA, A., GROSSKLAGS, J. & BÖHME, R. (2013). Bitspotting: Detecting optimal adaptive steganography. In *Proceedings of the 12th International Workshop on Digital-Forensics and Watermarking (IWDW)*.
- [47] KATZENBEISSER, S. & PETITCOLAS, F.A.P. (2002). Defining security in steganographic systems. In E.J. Delp III & P.W. Wong, eds., *Security and Watermarking of Multimedia Contents IV*, vol. 4675, 50–56, SPIE.
- [48] KER, A. (2005). Steganalysis of LSB matching in grayscale images. *IEEE Signal Processing Letters*, **12**, 441–444.
- [49] KER, A.D. (2005). Improved detection of LSB steganography in grayscale images. In J. Fridrich, ed., *Information Hiding (7th International Workshop)*, vol. 3200 of *Lecture Notes in Computer Science*, 97–115, Springer Berlin Heidelberg.
- [50] KER, A.D. (2007). Batch steganography and the threshold game. In E.J. Delp III & P.W. Wong, eds., *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505, 650504, SPIE.
- [51] KER, A.D. (2007). A weighted stego image detector for sequential LSB replacement. In *IAS '07: Proceedings of the Third International Symposium on Information Assurance and Security*, 453–456, IEEE Computer Society, Washington, DC, USA.
- [52] KER, A.D. (2009). Estimating the information theoretic optimal stego noise. In *IWDW '09: Proceedings of the 8th International Workshop on Digital Watermarking*, 184–198, Springer-Verlag, Berlin, Heidelberg.
- [53] KER, A.D. (2010). The square root law in stegosystems with imperfect information. In R. Böhme, P. Fong & R. Safavi-Naini, eds., *Information Hiding (12th International Workshop)*, vol. 6387 of *Lecture Notes in Computer Science*, 145–160, Springer Berlin Heidelberg.
- [54] KER, A.D. (2011). A curiosity regarding steganographic capacity of pathologically nonstationary sources. In N.D. Memon, J. Dittmann, A.M. Alattar & E.J. Delp III, eds., *Media Watermarking, Security, and Forensics III*, vol. 7880, 78800E, SPIE.
- [55] KER, A.D. (2014). Know your steganographic enemy. *Communications of the ACM*, **57**, 8–8.
- [56] KER, A.D. & BÖHME, R. (2008). Revisiting weighted stego-image steganalysis. In E.J.D. III, P.W. Wong, J. Dittmann & N.D. Memon, eds., *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819, 681905, SPIE.
- [57] KER, A.D., PEVNÝ, T., KODOVSKÝ, J. & FRIDRICH, J. (2008). The square root law of steganographic capacity. In *MM&Sec '08: Proceedings of the 10th ACM workshop on Multimedia and security*, 107–116, ACM, New York, NY, USA.
- [58] KER, A.D., BAS, P., BÖHME, R., COGRANNE, R., CRAVER, S., FILLER, T., FRIDRICH, J. & PEVNÝ, T. (2013). Moving steganography and steganalysis from the laboratory into the real world. In *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '13*, 45–58, ACM, New York, NY, USA.
- [59] KERCKHOFFS, A. (1883). La cryptographie militaire. *Journal des sciences militaires*, **IX**, 5–38.

-
- [60] KOCHER, P., JAFFE, J. & JUN, B. (1999). Differential power analysis. In *Advances in Cryptology – CRYPTO’99*, 388–397, Springer.
- [61] KOCHER, P.C. (1996). Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Advances in Cryptology – CRYPTO’96*, 104–113, Springer.
- [62] LAM, E. & GOODMAN, J. (2000). A mathematical analysis of the DCT coefficient distributions for images. *IEEE Transactions on Image Processing*, **9**, 1661–1666.
- [63] LASZKA, A., JOHNSON, B., SCHÖTTLE, P., GROSSKLAGS, J. & BÖHME, R. (2013). Managing the weakest link: A game-theoretic approach for the mitigation of insider threats. In *Proceedings of the 18th European Symposium on Research in Computer Security (ESORICS)*, 273–290.
- [64] LEYTON-BROWN, K. & SHOHAM, Y. (2008). *Essentials of Game Theory: A Concise, Multidisciplinary Introduction*. No. 3 in Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool.
- [65] LOU, D.C., WU, N.I., WANG, C.M., LIN, Z.H. & TSAI, C.S. (2010). A novel adaptive steganography based on local complexity and human vision sensitivity. *Journal of Systems and Software*, **83**, 1236 – 1248, {SPLC} 2008.
- [66] MAILLÉ, P., REICHL, P. & TUFFIN, B. (2011). Interplay between security providers, consumers, and attackers: a weighted congestion game approach. In *Decision and Game Theory for Security*, 67–86, Springer Berlin Heidelberg.
- [67] NASH, J. (1951). Non-cooperative games. *The Annals of Mathematics*, **54**, 286–295.
- [68] NEYMAN, J. & PEARSON, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A (Mathematical or Physical Character)*, **231**, 289–337.
- [69] ORSDEMIR, A., ALTUN, O., SHARMA, G. & BOCKO, M. (2008). Steganalysis-aware steganography: Statistical indistinguishability despite high distortion. In E.J. Delp III, P.W. Wong, J. Dittmann & N.D. Memon, eds., *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819, 681915, SPIE.
- [70] PÉREZ-FREIRE, L., COMESAÑA, P. & PÉREZ-GONZÁLEZ, F. (2005). Information-theoretic analysis of security in side-informed data hiding. In M. Barni, J. Herrera-Joancomartí, S. Katzenbeisser & F. Pérez-González, eds., *Information Hiding (7th International Workshop)*, vol. 3727 of *Lecture Notes in Computer Science*, 131–145, Springer Berlin Heidelberg.
- [71] PEVNÝ, T., FILLER, T. & BAS, P. (2010). Using high-dimensional image models to perform highly undetectable steganography. In R. Böhme, P. Fong & R. Safavi-Naini, eds., *Information Hiding (12th International Workshop)*, vol. 6387 of *Lecture Notes in Computer Science*, 161–177, Springer, Berlin Heidelberg.
- [72] PITA, J., JAIN, M., ORDÓNEZ, F., PORTWAY, C., TAMBE, M., WESTERN, C., PARUCHURI, P. & KRAUS, S. (2009). Using game theory for los angeles airport security. *AI Magazine*, **30**, 43.
- [73] PRAMITHA, K., SURESH, L. & SHUNMUGANATHAN, K. (2011). Image steganography using mod-4 embedding algorithm based on image contrast. In *International*

- Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN)*, 364–369.
- [74] PRUZHANSKY, V. (2011). Some interesting properties of maximin strategies. *International Journal of Game Theory*, **40**, 351–365.
- [75] ROBERTS, D.M. & GITTINS, J.C. (1978). The search for an intelligent evader: Strategies for searcher and evader in the two-region problem. *Naval Research Logistics Quarterly*, **25**, 95–106.
- [76] SALLEE, P. (2004). Model-based steganography. In T. Kalker, I. Cox & Y. Ro, eds., *Digital Watermarking*, vol. 2939 of *Lecture Notes in Computer Science*, 254–260, Springer Berlin Heidelberg.
- [77] SCHÖTTLE, P. & BÖHME, R. (2012). A game-theoretic approach to content-adaptive steganography. In M. Kirchner & D. Ghosal, eds., *Information Hiding (14th International Workshop)*, vol. 7692 of *Lecture Notes in Computer Science*, 125–141, Springer, Berlin Heidelberg.
- [78] SCHÖTTLE, P. & BÖHME, R. (2016). Game theory and adaptive steganography. *IEEE Transactions on Information Forensics and Security*, **11**, 760–773.
- [79] SCHÖTTLE, P., KORFF, S. & BÖHME, R. (2012). Weighted stego-image steganalysis for naive content-adaptive embedding. In *4th IEEE International Workshop on Information Forensics and Security (WIFS 2012)*, 193–198, IEEE.
- [80] SCHÖTTLE, P., JOHNSON, B., LASZKA, A., GROSSKLAGS, J. & BÖHME, R. (2013). A game-theoretic analysis of content-adaptive steganography with independent embedding. In *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*.
- [81] SEDIGHI, V. & FRIDRICH, J. (2015). Effect of imprecise knowledge of the selection channel on steganalysis. In *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, 33–42.
- [82] SHARP, T. (2001). An implementation of key-based digital signal steganography. In I.S. Moskowitz, ed., *Information Hiding (3rd International Workshop)*, vol. 2137 of *Lecture Notes in Computer Science*, 13–26, Springer Berlin Heidelberg.
- [83] SIMMONS, G.J. (1983). The prisoners’ problem and the subliminal channel. In *Advances in Cryptology – CRYPTO ’83*, 51–67, Plenum Press.
- [84] SINGH, K., SINGH, L., SINGH, A. & DEVI, K. (2007). Hiding secret message in edges of the image. In *International Conference on Information and Communication Technology*, 238–241.
- [85] VOLOSHYNOVSKIY, S., HERRIGEL, A., BAUMGAERTNER, N. & PUN, T. (2000). A stochastic approach to content adaptive digital image watermarking. In A. Pfitzmann, ed., *Information Hiding (3rd International Workshop)*, vol. 1768 of *Lecture Notes in Computer Science*, 211–236, Springer Berlin Heidelberg.
- [86] VON NEUMANN, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, **100**, 295–320.
- [87] VON NEUMANN, J. & MORGENSTERN, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.
- [88] WANG, Y. & MOULIN, P. (2008). Perfectly secure steganography: Capacity, error

- exponents, and code constructions. *IEEE Transactions on Information Theory*, **54**, 2706–2722.
- [89] WESTFELD, A. & PFITZMANN, A. (2000). Attacks on steganographic systems. In A. Pfitzmann, ed., *Information Hiding (3rd International Workshop)*, vol. 1768 of *Lecture Notes in Computer Science*, 6–76, Springer Berlin Heidelberg.
- [90] WU, D.C. & TSAI, W.H. (2003). A steganographic method for images by pixel-value differencing. *Pattern Recognition Letters*, **24**, 1613–1626.
- [91] WU, H.C., WU, N.I., TSAI, C.S. & HWANG, M.S. (2005). Image steganographic scheme based on pixel-value differencing and LSB replacement methods. *IEE Proceedings - Vision, Image and Signal Processing*, **152**, 611 – 615.
- [92] YANG, C.H., WENG, C.Y., WANG, S.J. & SUN, H.M. (2008). Adaptive data hiding in edge areas of images with spatial LSB domain systems. *IEEE Transactions on Information Forensics and Security*, **3**, 488–497.
- [93] YU, X., TAN, T. & WANG, Y. (2005). Extended optimization method of LSB steganalysis. In *IEEE International Conference on Image Processing*, vol. 2, II – 1102–5.

Appendix A

Information-Theoretic Derivations

A.1 Derivation of Definition 3.2

The definition of *perfect uncertainty* follows from notions of the entropy as follows:

Let $\mathbf{x} \sim \mathbf{X}^{(q)}$ for $q \in \{0, 1\}$ and $q \sim Q$. Then, perfect uncertainty about Q given \mathbf{x} can be described with the entropy of the conditional distribution $H(Q|\mathbf{x}) = H(Q)$.

It follows:

$$H(Q|\mathbf{X} = \mathbf{x}) = H(Q) \Leftrightarrow \Pr(Q = 0|\mathbf{X} = \mathbf{x}) = \Pr(Q = 1|\mathbf{X} = \mathbf{x}), \quad (\text{A.1})$$

and with $j \in \{0, 1\}$

$$\Pr(Q = j|\mathbf{x}) = \frac{\Pr(\mathbf{x}|Q = j) \Pr(Q = j)}{\Pr(\mathbf{x})}, \quad (\text{A.2})$$

and

$$\Pr(Q = 0) = \Pr(Q = 1) = \frac{1}{2} \text{ (equal priors)}, \quad (\text{A.3})$$

that

$$\Pr(\mathbf{x}|Q = 0) = \Pr(\mathbf{x}|Q = 1) \Leftrightarrow \mathcal{P}_0(\mathbf{x}) = \mathcal{P}_1(\mathbf{x}). \quad (\text{A.4})$$

A.2 Derivation of Remark 3.9

Proof. If Equation (3.1) holds for all cover realizations, it holds that

$$\forall \mathbf{x} : \mathcal{P}_0(\mathbf{x}) = \mathcal{P}_1(\mathbf{x}) \Leftrightarrow \log \frac{\mathcal{P}_0(\mathbf{x})}{\mathcal{P}_1(\mathbf{x})} = \log 1 = 0, \quad (\text{A.5})$$

and thus:

$$D_{KL}(\mathcal{P}_0||\mathcal{P}_1) = 0. \quad (\text{A.6})$$

If Equation (3.1) does not hold for one realization of the cover source \mathbf{x} and w. l. o. g. $\mathcal{P}_0(\mathbf{x}) > \mathcal{P}_1(\mathbf{x})$, it follows that $D_{KL}(\mathcal{P}_0||\mathcal{P}_1) > 0$. \square

Appendix B

Game Theory in Related Fields

We are not the first to consider game theory for modeling situations with rational defenders and rational attackers. In this section we present game-theoretic approaches in two fields closely connected to steganography, namely multimedia forensics and watermarking. Although there is no direct counterpart to the analysis of adaptive steganography, we also decided to include a game-theoretic approach from the area of adversarial classification, as interesting parallels exist and the applicability of results in adversarial classification for universal steganalysis in practice seems worth exploring.

B.1 Multimedia Forensics

In 2013 Barni and Tondi [4] present a meta-game, modeling the source identification problem in multimedia forensics. They build on the hypothesis testing framework presented in [9] to cast the problem of a forensic analyst (FA) who has to decide if a given sequence (image), possibly altered by an adversary (AD), was generated by the source \mathbf{X} or by the source \mathbf{Y} . The set-up is similar to the one in steganography, as the early works in this area did not consider the presence of an adversary, aiming to impede the forensic analysis. The authors set out “to derive the ultimate achievable performance of the forensic analysis in the presence of an adversary” [4, p. 450].

The set-up is as follows: There are two sources, \mathbf{X} and \mathbf{Y} , distributed according to \mathcal{P}_X and \mathcal{P}_Y , respectively. Both FA and AD know both distributions and the goal of the FA is to distinguish sequences generated by \mathbf{X} from sequences generated by other sources. The aim of the AD is, given a sequence $\mathbf{y}^n = (y_1, \dots, y_n)$ generated by \mathbf{Y} , to transform this sequence into another sequence $\mathbf{z}^n = (z_1, \dots, z_n)$ in such a way that the FA believes it was drawn from \mathbf{X} . The situation is modeled as a strategic two-player zero-sum game. The strategy space of the FA is to choose the acceptance region Λ_0 for sequences generated by \mathbf{X} , subject to a prescribed false positive rate P_{fp} . The strategies of the AD are all functions $f(\cdot)$ that map a sequence \mathbf{y}^n to another sequence \mathbf{z}^n , subject to a maximum allowed average per-letter distortion D . The payoff function (for the FA) is defined as the false negative error probability:

$$u(\Lambda_0, f) = -P_{fn} = - \sum_{\mathbf{y}^n: f(\mathbf{y}^n) \in \Lambda_0} \mathcal{P}_Y(\mathbf{y}^n), \quad (\text{B.1})$$

where Λ_0 is the acceptance region. As this general game proves intractable, the authors switch to the asymptotic case, where the length of the sequence n tends to infinity. In this situation they identify an asymptotic Nash equilibrium, in which the optimal strategy of the FA does neither depend on the strategy chosen by the AD

nor the probability density function \mathcal{P}_Y . The optimal strategy of the AD is a simple minimization problem, as the optimal strategy of the FA is universally optimal and thus Λ_0 is fixed. When computing the payoff in equilibrium, the authors show that, asymptotically, two sources \mathcal{P}_X and \mathcal{P}_Y are either perfectly indistinguishable or not. In the first case, P_{fn} tends to 1, in the second case, P_{fn} tends to 0 exponentially fast.

B.2 Digital Watermarking

In digital watermarking, together with steganography, the most popular field in information hiding, game theory is directly used to construct the so-called Tardos-Codes [1]. Tardos-Codes are *fingerprinting codes* that are used in digital watermarking to mark each copy \mathbf{X}^v of a digital medium $\mathbf{X} = (X_1, \dots, X_n)$ in order to detect *collusion attacks*. A collusion attack is an attack, where several malicious users (pirates) perform an attack by comparing their respective copies of the watermarked medium \mathbf{X}^v , thus detecting some of the positions where the watermark was embedded, by simply comparing the values at all the positions \mathbf{X}_j^v for $j \in \{1, \dots, n\}$. A fingerprinting code is called ε -secure against t pirates if for any set T of pirates with $|T| \leq t$, the probability that either none of the pirates is caught or some user is falsely accused is at most ε . For this, a fingerprinting code consists of an accusation algorithm and a randomized procedure that generates codewords \mathbf{X}^v over an finite alphabet Σ^n for users $v \in U$.

The Tardos-Codes are then generated using the notions of a t -channel and a *bias-based code generation*. A t -channel is a randomized procedure that produces an output bit f from an input $x \in \{0, 1\}^t$. It is determined by the function $S : x \in \{0, 1\}^t \rightarrow [0, 1]$, defined as $S(z) = Pr[f = 1 | x = z]$. The pirates are said to use the channel S to produce their forged codeword F .

A *bias-based code generation* is a process consisting of two phases, determined by a probability distribution D on the interval $[0, 1]$ (the *bias distribution*). First, the *bias vector* $\mathbf{P} = (P_1, \dots, P_n)$ is selected, by selecting individual biases $P_j \in [0, 1]$ independently and according to the distribution D . In the second phase the bits of each codeword \mathbf{X}^v are selected, where it holds for each it that $Pr[X_j^v = 1] = P_j$.

For $p \in [0, 1]$ and a channel S , the distribution $B_{p,S}$ on the binary vector $x \in \{0, 1\}^t$ and the binary variable $f \in \{0, 1\}$ is defined by choosing individual digits x_i of x independently for $i \in \{1, \dots, t\}$ with an identical distribution of expectation p and finally obtaining f from x through the channel S . $I_{p,S}$ is defined as the mutual information (c.f. Definition 2.8 on p. 19) in this distribution. The game that the authors choose to proof that these codes are optimal consists of the pirates choosing the channel S and the distributor of the fingerprinting codes who chooses the probability $p \in [0, 1]$. Then, the pirates have to pay the distributor the amount of $I_{p,S}$ as payoff. The authors show that the pirates are always better off using a pure strategy and the equilibrium that is based on (continuous) minmax strategies gives the optimal bias distribution. In the equilibrium it holds that:

$$\min_S \max_p I_{p,S} = \max_D \min_S E_{p \in D} [I_{p,S}], \quad (\text{B.2})$$

where $E_{p \in D}[\cdot]$ represents the expectation as p is distributed according to D .

Finally, the authors show that with the help of these values a fingerprinting code can be constructed that has optimal capacity and is ε -secure.

B.3 Adversarial Classification

A recent theme at the intersection of machine learning and security is adversarial classification [13], a subfield of knowledge discovery and data mining (KDD). Dalvi et al. propose a game theoretic framework for the following situation: data is actively manipulated by an adversary seeking to increase the false negative rate of a binary classifier. Here, the classifier is assumed to be a data mining algorithm and possible domains of interest are spam classification, surveillance, counter-terrorism and intrusion detection. The authors exemplify their game in the spam detection domain, where a classifier (CL) has to classify a given instance (email) $\mathbf{x} = (x_1, \dots, x_n)$ as either malicious (i.e. spam) or innocent (i.e. regular email). They assume that innocent instances are generated i.i.d. from a distribution $\mathcal{P}(\mathbf{X}|-)$ and malicious ones likewise from a distribution $\mathcal{P}(\mathbf{X}|+)$. Each instance \mathbf{x} consists of n features or attributes x_i . Furthermore, the authors assume that there are two sets, the training set \mathcal{S} and the test set \mathcal{T} . The CL wants to learn a function $y_C = C(\mathbf{x})$ from \mathcal{S} to correctly predict the type of instances in \mathcal{T} and the adversary (AD) wants to modify sequences \mathbf{x} to \mathbf{x}' so that CL misclassifies them. Then, the authors define costs V_i for measuring the different features x_i and a utility for correctly ($U_C(+|-), U_C(-|+)$) and falsely ($U_C(-|-), U_C(+|+)$) classifying an instance \mathbf{x} , defining the strategy space of the CL. The AD has a cost $W_i(x_i, x'_i)$ for changing the i -th feature and similar utilities $U_A(\pm|\mp)$.

As this is a non-zero-sum game and the number of actions is doubly exponential in the number of features n , the authors conclude that computing a Nash equilibrium will be intractable, although they prove that it exists.

The authors continue in deriving an optimal naïve Bayes classifier, which performs an LR test on \mathcal{S} to optimally classify instances from \mathcal{T} . Then, the AD assumes this kind of classifier on the side of the CL and optimizes his strategy. The optimal strategy of the AD is formulated as a constrained optimization problem and the solution is given as a binary linear program. Now, the CL is allowed to adapt its strategy to the optimal strategy by the AD. Under the assumption that the AD has not tampered the training set \mathcal{S} an optimal response to the AD's optimal strategy is derived. The authors present efficient ways of implementing both strategies and then test the classifier against a naïve Bayes classifier on two spam datasets, incurring different costs for false positives, i.e., classifying a non-spam email as spam. Their new classifier clearly outperforms the naïve Bayes classifier. In the end, the authors try to cast their game as a repeated game and allow the AD to adapt his strategy to the new optimal strategy by the CL. By this, unsurprisingly, the payoff alternates, depending which of the player is allowed to adapt its strategy. As a result from this, the authors conclude that in a repeated game, AD and CL will never reach an equilibrium.

Appendix C

Omitted Proofs

C.1 Proof of Lemma 5.11

Proof. For the sequence 00 it holds:

$$\Pr[00|s] = (1 - f(0))(1 - f(1))(1 - \bar{a}_0)\bar{a}_0 + f(0)(1 - f(1))\bar{a}_0^2 \quad (\text{C.1})$$

$$+ (1 - f(0))f(1)(1 - \bar{a}_0)^2 + f(0)f(1)\bar{a}_0(1 - \bar{a}_0) \\ > (1 - f(0))(1 - f(1)) [\bar{a}_0^2 + 2(1 - \bar{a}_0)\bar{a}_0 + (1 - \bar{a}_0)^2] \quad (\text{C.2})$$

$$= (1 - f(0))(1 - f(1))(\bar{a}_0 + 1 - \bar{a}_0)^2 \quad (\text{C.3})$$

$$= (1 - f(0))(1 - f(1)) = \Pr[00|c] \quad (\text{C.4})$$

since $f(0) > 1 - f(0)$ and $f(1) > 1 - f(1)$.

For the sequence 11 it holds:

$$\Pr[11|s] = f(0)f(1)(1 - \bar{a}_0)\bar{a}_0 + (1 - f(0))f(1)\bar{a}_0^2 \quad (\text{C.5})$$

$$+ f(0)(1 - f(1))(1 - \bar{a}_0)^2 + (1 - f(0))(1 - f(1))\bar{a}_0(1 - \bar{a}_0) \\ < f(0)f(1) [\bar{a}_0^2 + 2(1 - \bar{a}_0)\bar{a}_0 + (1 - \bar{a}_0)^2] \quad (\text{C.6})$$

$$= f(0)f(1)(\bar{a}_0 + 1 - \bar{a}_0)^2 \quad (\text{C.7})$$

$$= f(0)f(1) = \Pr[11|c] \quad (\text{C.8})$$

since $1 - f(0) < f(0)$ and $1 - f(1) < f(1)$. □

C.2 Proof of Lemma 5.12

Proof. For the sequence 01 it holds:

$$\Pr[01|c] > \Pr[01|s] \quad (\text{C.9})$$

$$0 > 4\bar{a}_0^2\tilde{f}(0)\tilde{f}(1) + 2\bar{a}_0 \left[(1 - f(0))\tilde{f}(1) + \tilde{f}(0)(1 - f(1)) \right] \\ + 2(f(0) - 1)\tilde{f}(1) \quad (\text{C.10})$$

The above inequality holds when $(\bar{a}_0)_1 < \bar{a}_0 < (\bar{a}_0)_2$, where

$$(\bar{a}_0)_{1,2} = \frac{(f(0) - 1)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1)}{4\tilde{f}(0)\tilde{f}(1)} \\ \mp \frac{\sqrt{\left[(1 - f(0))\tilde{f}(1) + \tilde{f}(0)(1 - f(1)) \right]^2 - 8\tilde{f}(0)\tilde{f}(1)(f(0) - 1)\tilde{f}(1)}}{4\tilde{f}(0)\tilde{f}(1)}. \quad (\text{C.11})$$

since $(f(0) - 1)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1) < 0$, we have that $(\bar{a}_0)_1 < 0$. Therefore, Eve classifies realization 01 as cover when

$$\begin{aligned} \bar{a}_0 < (\bar{a}_0)_2 &= \frac{(f(0) - 1)\tilde{f}(1) + \tilde{f}(0)(f(1) - 1)}{4\tilde{f}(0)\tilde{f}(1)} \\ &+ \frac{\sqrt{\left[(1 - f(0))\tilde{f}(1) + \tilde{f}(0)(1 - f(1))\right]^2 - 8\tilde{f}(0)\tilde{f}(1)^2(f(0) - 1)}}{4\tilde{f}(0)\tilde{f}(1)}. \end{aligned} \quad (\text{C.12})$$

For the sequence 10 it holds:

$$\Pr[10|c] > \Pr[10|s] \quad (\text{C.13})$$

$$0 > 4\bar{a}_0^2\tilde{f}(0)\tilde{f}(1) + 2\bar{a}_0 \left(-f(0)\tilde{f}(1) - \tilde{f}(0)f(1) \right) + 2f(0)\tilde{f}(1) \quad (\text{C.14})$$

The above inequality holds when $(\bar{a}_0)_1 < \bar{a}_0 < (\bar{a}_0)_2$, where

$$(\bar{a}_0)_{1,2} = \frac{f(0)\tilde{f}(1) + \tilde{f}(0)f(1)}{4\tilde{f}(0)\tilde{f}(1)} \mp \frac{\sqrt{\left[f(0)\tilde{f}(1) + \tilde{f}(0)f(1)\right]^2 - 8\tilde{f}(0)\tilde{f}(1)f(0)\tilde{f}(1)}}{4\tilde{f}(0)\tilde{f}(1)}. \quad (\text{C.15})$$

$$\frac{f(0)\tilde{f}(1) + \tilde{f}(0)f(1)}{4\tilde{f}(0)\tilde{f}(1)} = \frac{f(0)}{4\tilde{f}(0)} + \frac{f(1)}{4\tilde{f}(1)} \quad (\text{C.16})$$

$$= \frac{f(0)}{4f(0) - 2} + \frac{f(1)}{4f(1) - 2} \quad (\text{C.17})$$

$$= \frac{1}{4} \frac{4f(0) - 2 + 2}{4f(1) - 2} + \frac{1}{4} \frac{4f(1) - 2 + 2}{4f(2) - 2} \quad (\text{C.18})$$

$$= \frac{1}{4} \left(1 + \frac{2}{4f(1) - 2} + 1 + \frac{2}{4f(2) - 2} \right) \quad (\text{C.19})$$

$$> \frac{1}{4} \left(1 + \frac{2}{4 - 2} + 1 + \frac{2}{4 - 2} \right) \quad (\text{C.20})$$

$$= \frac{1}{4} (1 + 1 + 1 + 1) \quad (\text{C.21})$$

$$= 1 \quad (\text{C.22})$$

since $\frac{f(0)\tilde{f}(1) + \tilde{f}(0)f(1)}{4\tilde{f}(0)\tilde{f}(1)} > 1$, we have that $(\bar{a}_0)_2 > 1$. Therefore, Eve classifies realization 01 as cover when

$$\bar{a}_0 > (\bar{a}_0)_1 = \frac{f(0)\tilde{f}(1) + \tilde{f}(0)f(1) - \sqrt{\left[f(0)\tilde{f}(1) + \tilde{f}(0)f(1)\right]^2 - 8f(0)\tilde{f}(0)\tilde{f}(1)^2}}{4\tilde{f}(0)\tilde{f}(1)}. \quad (\text{C.23})$$

□

C.3 Proof of Lemma 5.13

Proof. The lemma can be expressed as: Eve never classifies both realization 01 and 10 as cover.

We have that Eve's optimal decision rule is

$$w_1 = \log \frac{f(0)(1 - f(0) + \bar{a}_0 \tilde{f}(0))}{(1 - f(0))(f(0) - \bar{a}_0 \tilde{f}(0))} \quad (\text{C.24})$$

$$w_2 = \log \frac{f(1)(1 - f(1) + \tilde{f}(1) - \bar{a}_0 \tilde{f}(1))}{(1 - f(1))(f(1) - \tilde{f}(1) + \bar{a}_0 \tilde{f}(1))} \quad (\text{C.25})$$

$$\tau = \log \left[\frac{1 - f(0) + \bar{a}_0 \tilde{f}(0)}{1 - f(0)} \frac{1 - f(1) + \tilde{f}(1) - \bar{a}_0 \tilde{f}(1)}{1 - f(1)} \right]. \quad (\text{C.26})$$

Now, assume that, for some \bar{a}_0 , the claim of the lemma does not hold. Then,

$$w_1 > \tau \quad (\text{C.27})$$

$$f(0)(1 - f(1)) > (f(0) - \bar{a}_0 \tilde{f}(0))(1 - f(1) + \tilde{f}(1) - \bar{a}_0 \tilde{f}(1)) \quad (\text{C.28})$$

$$\begin{aligned} 0 &> f(0)\tilde{f}(1) - \bar{a}_0 f(0)\tilde{f}(1) - \bar{a}_0 \tilde{f}(0) + \bar{a}_0 \tilde{f}(0)f(1) \\ &\quad - \bar{a}_0 \tilde{f}(0)\tilde{f}(1) + \bar{a}_0^2 \tilde{f}(0)\tilde{f}(1) \end{aligned} \quad (\text{C.29})$$

and

$$w_2 > \tau \quad (\text{C.30})$$

$$f(1)(1 - f(0)) > (1 - f(0) + \bar{a}_0 \tilde{f}(0))(f(1) - \tilde{f}(1) + \bar{a}_0 \tilde{f}(1)) \quad (\text{C.31})$$

$$\begin{aligned} 0 &> \bar{a}_0 \tilde{f}(1) - \tilde{f}(1) + f(0)\tilde{f}(1) - \bar{a}_0 f(0)\tilde{f}(1) + \bar{a}_0 \tilde{f}(0)f(1) \\ &\quad - \bar{a}_0 \tilde{f}(0)\tilde{f}(1) + \bar{a}_0^2 \tilde{f}(0)\tilde{f}(1). \end{aligned} \quad (\text{C.32})$$

By adding Equation C.29 and C.32 together, we have that

$$\begin{aligned} 0 &> 2\bar{a}_0^2 \tilde{f}(0)\tilde{f}(1) + \bar{a}_0(-2f(0)\tilde{f}(1) + 2\tilde{f}(0)f(1) - 2\tilde{f}(0)\tilde{f}(1) + \tilde{f}(1) \\ &\quad - \tilde{f}(0)) + \tilde{f}(0)\tilde{f}(1) \end{aligned} \quad (\text{C.33})$$

$$= 2\bar{a}_0^2 \tilde{f}(0)\tilde{f}(1) - 2\bar{a}_0 \tilde{f}(0)\tilde{f}(1) + \tilde{f}(0)\tilde{f}(1) \quad (\text{C.34})$$

$$= (2\bar{a}_0^2 - 2\bar{a}_0 + 1)\tilde{f}(0)\tilde{f}(1) \quad (\text{C.35})$$

$$= (\bar{a}_0^2 + (\bar{a}_0 - 1)^2)\tilde{f}(0)\tilde{f}(1) > 0, \quad (\text{C.36})$$

which is a contradiction. Therefore, the claim of the lemma has to hold. \square

C.4 Proof of Lemma 5.14

Proof. For $\bar{a}_0 \in [0, \tau_1]$ it holds:

Taking the derivative of the specific pay-off function from Theorem 5.7 with respect to \bar{a}_0 yields:

$$\text{payoff}|_{e=[scsc]} = -\Pr[00|s](\bar{a}_0) + \Pr[01|s](\bar{a}_0) - \Pr[10|s](\bar{a}_0) + \Pr[11|s](\bar{a}_0) \quad (\text{C.37})$$

$$\begin{aligned} &= \bar{a}_0^2 \left[-f(0)(1-f(1)) + (1-f(0))(1-f(1)) - (1-f(0))f(1) + f(0)f(1) \right. \\ &\quad + f(0)f(1) - (1-f(0))f(1) + (1-f(0))(1-f(1)) - f(0)(1-f(1)) \\ &\quad - (1-f(0))(1-f(1)) + f(0)(1-f(1)) - f(0)f(1) + (1-f(0))f(1) \\ &\quad \left. + (1-f(0))f(1) - f(0)f(1) + f(0)(1-f(1)) - (1-f(0))(1-f(1)) \right] \\ &+ \bar{a}_0 \left[- (1-f(0))(1-f(1)) + 2(1-f(0))f(1) - f(0)f(1) \right. \\ &\quad + (1-f(0))f(1) - 2(1-f(0))(1-f(1)) + f(0)(1-f(1)) \\ &\quad - f(0)(1-f(1)) + 2f(0)f(1) - (1-f(0))f(1) \\ &\quad \left. + f(0)f(1) - 2f(0)(1-f(1)) + (1-f(0))(1-f(1)) \right] \\ &\quad - f(0)f(1) + f(0)(1-f(1)) + (1-f(0))(1-f(1)) - (1-f(0))f(1) \end{aligned}$$

$$= 2\bar{a}_0 \left[(1-f(0))\tilde{f}(1) + f(0)\tilde{f}(1) \right] - \tilde{f}(1) \quad (\text{C.38})$$

$$= 2\bar{a}_0 \tilde{f}(1) - \tilde{f}(1) . \quad (\text{C.39})$$

$$\frac{\partial \text{payoff}}{\partial \bar{a}_0} \Big|_{e=[scsc]} = 2\tilde{f}(1) > 0 \quad (\text{C.40})$$

For $\bar{a}_0 \in [\tau_1, 1]$ it holds:

Taking the derivative of the specific pay-off function from Theorem 5.7 with respect to \bar{a}_0 yields:

$$\text{pay-off}|_{e=[sscc]} = \Pr[10|s](\bar{a}_0) + \Pr[11|s](\bar{a}_0) - \Pr[01|s](\bar{a}_0) - \Pr[00|s](\bar{a}_0) \quad (\text{C.41})$$

$$\begin{aligned} &= \bar{a}_0^2 \left[(1-f(0))(1-f(1)) - f(0)(1-f(1)) + f(0)f(1) - (1-f(0))f(1) \right. \\ &\quad + (1-f(0))f(1) - f(0)f(1) + f(0)(1-f(1)) - (1-f(0))(1-f(1)) \\ &\quad - f(0)f(1) + (1-f(0))f(1) - (1-f(0))(1-f(1)) + f(0)(1-f(1)) \\ &\quad \left. - f(0)(1-f(1)) + (1-f(0))(1-f(1)) - (1-f(0))f(1) + f(0)f(1) \right] \\ &+ \bar{a}_0 \left[f(0)(1-f(1)) - 2f(0)f(1) + (1-f(0))f(1) \right. \\ &\quad + f(0)f(1) - 2f(0)(1-f(1)) + (1-f(0))(1-f(1)) \\ &\quad - (1-f(0))f(1) + 2(1-f(0))(1-f(1)) - f(0)(1-f(1)) \\ &\quad \left. - (1-f(0))(1-f(1)) + 2(1-f(0))f(1) - f(0)f(1) \right] \\ &+ f(0)f(1) + f(0)(1-f(1)) - (1-f(0))(1-f(1)) - (1-f(0))f(1) \\ &= \bar{a}_0[2 - 4f(0)] + \tilde{f}(0) . \end{aligned} \quad (\text{C.42})$$

$$\frac{\partial \text{pay-off}}{\partial \bar{a}_0} \Big|_{e=[sscc]} = -2\tilde{f}(0) < 0 \quad (\text{C.43})$$

□

C.5 Proof of Lemma 5.15

Proof. Taking the derivative of the specific pay-off function from Theorem 5.7 with respect to \bar{a}_0 yields:

$$\begin{aligned} \text{payoff} \Big|_{e=[sssc]} &= -\Pr[00|s](\bar{a}_0) - \Pr[01|s](\bar{a}_0) - \Pr[10|s](\bar{a}_0) + \Pr[11|s](\bar{a}_0) \quad (\text{C.44}) \\ &= \bar{a}_0^2 \left[-f(0)(1-f(1)) + (1-f(0))(1-f(1)) - (1-f(0))f(1) + f(0)f(1) \right. \\ &\quad - f(0)f(1) + (1-f(0))f(1) - (1-f(0))(1-f(1)) + f(0)(1-f(1)) \\ &\quad - (1-f(0))(1-f(1)) + f(0)(1-f(1)) - f(0)f(1) + (1-f(0))f(1) \\ &\quad \left. + (1-f(0))f(1) - f(0)f(1) + f(0)(1-f(1)) - (1-f(0))(1-f(1)) \right] \\ &\quad + \bar{a}_0 \left[-(1-f(0))(1-f(1)) + 2(1-f(0))f(1) - f(0)f(1) \right. \\ &\quad - (1-f(0))f(1) + 2(1-f(0))(1-f(1)) - f(0)(1-f(1)) \\ &\quad - f(0)(1-f(1)) + 2f(0)f(1) - (1-f(0))f(1) \\ &\quad \left. + f(0)f(1) - 2f(0)(1-f(1)) + (1-f(0))(1-f(1)) \right] \\ &\quad - f(0)f(1) + f(0)(1-f(1)) - (1-f(0))(1-f(1)) - (1-f(0))f(1) \\ &= 2\bar{a}_0^2 \left[f(0)(1-f(1)) + (1-f(0))f(1) - f(0)f(1) - (1-f(0))(1-f(1)) \right] \\ &\quad + 2\bar{a}_0 \left[(1-f(0))(1-f(1)) + f(0)f(1) - 2f(0)(1-f(1)) \right] \\ &\quad - f(0)\tilde{f}(1) - 1 + f(0) \\ &= 8\bar{a}_0^2 \left[-\tilde{f}(0)\tilde{f}(1) \right] + 4\bar{a}_0 \left[f(0)\tilde{f}(1) - \tilde{f}(0)(1-f(1)) \right] \quad (\text{C.45}) \\ &\quad - f(0)\tilde{f}(1) - 1 + f(0). \end{aligned}$$

$$\frac{\partial \text{payoff}}{\partial \bar{a}_0} \Big|_{e=[sssc]} = -16\bar{a}_0\tilde{f}(0)\tilde{f}(1) + 4\left(f(0)\tilde{f}(1) - \tilde{f}(0)(1-f(1))\right) \quad (\text{C.46})$$

The second derivative follows immediately. □

Appendix D

Curriculum Vitae

