Further Study on the Security of S-UNIWARD

Tomáš Denemark, Jessica Fridrich, and Vojtěch Holub

Department of ECE, SUNY Binghamton, NY, USA

ABSTRACT

Recently, a new steganographic method was introduced that utilizes a universal distortion function called UNIWARD. The distortion between the cover and stego image is computed as a sum of relative changes of wavelet coefficients representing both images. As already pointed out in the original publication, the selection channel of the spatial version of UNIWARD (the version that hides messages in pixel values called S-UNIWARD) exhibits unusual properties – in highly textured and noisy regions the embedding probabilities form interleaved streaks of low and high embedding probability. While the authors of UNIWARD themselves hypothesized that such an artifact in the embedding probabilities may jeopardize its security, experiments with state-of-the-art rich models did not reveal any weaknesses. Using the fact that the cover embedding probabilities can be approximately estimated from the stego image, we introduce the novel concept of content-selective residuals and successfully attack S-UNIWARD. We also show that this attack, which is made possible by a faulty probabilistic selection channel, can be prevented by properly adjusting the stabilizing constant in the UNIWARD distortion function.

Keywords: Steganalysis, steganography, S-UNIWARD, content-selective residual, adaptive selection channel, distortion, security.

1. INTRODUCTION

Modern steganographic schemes for digital media constrain their embedding changes to regions, which are difficult to describe using statistical models, such as complex textures or noisy areas, making it hard for the Warden to detect the embedding. A practical way to implement content-adaptive steganography is to assign a cost of changing each cover element (e.g., pixel or DCT coefficient) and then embed a given message while minimizing the expected sum of costs over modified pixels. This can be achieved using syndrome-coding methods, such as the Syndrome-Trellis Codes (STCs).[1] Since STCs operate near the theoretical payload–distortion bound, advancing the security of steganography has thus been reduced to the problem of identifying distortion functions that would better capture the statistical impact of embedding changes. We note that for steganography in empirical covers,[2] such as digital media files, the security is typically evaluated empirically using classifiers for a given cover source.

The first stego method that followed the above embedding paradigm was HUGO.[3] Its distortion function was designed to minimize the statistical discrepancy introduced into the higher-order statistics of differences between adjacent pixels utilized by the SPAM feature vector.[4] In Ref. [5], an attempt was made to assign the costs based on a feedback received from a set of examples (images) from a particular cover source to minimize the margin between the classes of cover and stego features. The authors of Uniform Embedding Distortion (UED)[6] made the embedding costs inversely proportional to the frequency with which the cover values occur. The embedding costs of WOW[7] were designed to be large in such regions of the image in which the content cannot be easily modeled in any direction. This was achieved by employing a directional filterbank (wavelet basis) that assessed the content in multiple directions by computing directional noise residuals (wavelet coefficients in multiple subbands). The distortion function called UNIWARD (UNIversal WAvelet Relative Distortion) proposed in Ref. [8] was designed to be a universal measure for constructing steganographic methods in an arbitrary domain. It uses the same bank of directional filters as WOW but computes the cost of changing a given cover element simply as the sum of relative changes of wavelet coefficients over all subbands. UNIWARD was invented to allow an easy extension to other embedding domains and to give the distortion the proper non-additive form for the Gibbs construction[9] to embed while considering the interactions among nearby embedding changes.

E-mail: {tdenema1,fridrich,vholub1}@binghamton.edu; http://dde.binghamton.edu

Based on the results reported in the original publication, the spatial, JPEG, and side-informed JPEG versions of UNIWARD exhibited the highest level of security when tested empirically with rich media models[10, 11] on the BOSSbase 1.01 source.[12] The version of UNIWARD for embedding in the spatial (pixel) domain, however, exhibited one unusual behavior. In regions with complex content, the embedding probabilities, as dictated by the pixel costs, formed interleaved bands of high and low probabilities. While steganalysis with rich media models was apparently unable to capitalize on this embedding artifact, it remained an open problem whether such a probabilistic selection channel was a security weakness. The main contribution of this paper is to show that the artifacts in embedding probabilities can indeed be used to mount a simple but very powerful attack on S-UNIWARD that lowers the detection error below 2% for a wide range of payloads. The main idea is to divide the statistics computed from noise residuals into several subsets obtained from groups of pixels with low and high embedding probabilities estimated from the stego image. We further show that the artifacts in the embedding probabilities are caused by an improperly chosen stabilizing constant in the UNIWARD distortion function. By adjusting its value, not only the artifacts are suppressed but also the attack described in this paper becomes no more effective while the security w.r.t. rich models stays unchanged restoring thus the high security of schemes based on UNIWARD.

In the next section, we describe the UNIWARD distortion function, contrast its embedding probabilities with those of HUGO and WOW, and point out the existence of "streaking artifacts." In Section 3, we introduce the novel concept of a content-selective residual based on estimating the embedding change probabilities from the stego image. In Section 4, we first study the best settings for the parameters on which the newly proposed steganalysis features depend and then report the detection accuracy of S-UNIWARD across a range of payloads. The following Section 5 focuses on identifying the cause of the artifacts in the embedding probabilities and shows that increasing the value of the stabilizing constant in UNIWARD suppresses the artifacts. To test the efficiency of this fix, we subject the modified S-UNIWARD to the same tests as in Section 4.2 and conclude that the fixed S-UNIWARD is not susceptible to the proposed attack while its security w.r.t. steganalysis using the spatial rich model remains basically unchanged. The paper is concluded in Section 6, where we discuss the implications of our study for design of adaptive steganographic schemes and debate future research directions.

2. UNIWARD DISTORTION FUNCTION

Given a pair of b-bit grayscale cover and stego images, $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{X}, \mathbf{Y} \in \mathcal{I}^{M\times N}$, $\mathcal{I} = \{0, \ldots, 2^b - 1\}$, we let $W_{uv}^{(k)}(\mathbf{X})$ and $W_{uv}^{(k)}(\mathbf{Y})$, $k = 1, 2, 3$, $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}$ denote the uvth wavelet coefficient in the first-level undecimated Daubechies 8-tap wavelet decomposition. The index $k = 1, 2, 3$ corresponds to the LH, HL, and HH subbands, respectively. In UNIWARD, the distortion is computed as a sum of relative changes of all wavelet coefficients w.r.t. their cover values:

$$D(\mathbf{X}, \mathbf{Y}) = \sum_{k=1}^{3} \sum_{u=1}^{M} \sum_{v=1}^{N} \frac{|W_{uv}^{(k)}(\mathbf{X}) - W_{uv}^{(k)}(\mathbf{Y})|}{\sigma + |W_{uv}^{(k)}(\mathbf{X})|}, \tag{1}$$

where $\sigma > 0$ is a stabilizing constant introduced to avoid dividing by zero. In Ref. [8], the authors proposed $\sigma = 10 \times \mathrm{eps} \approx 2 \times 10^{-15}$, where eps is defined in Matlab as the difference between 1.0 and the next larger double-precision number.

Note that the ratio in (1) is smaller when a large cover wavelet coefficient is changed (where texture and edges appear), while embedding changes are discouraged in regions where $|W_{uv}^{(k)}(\mathbf{X})|$ is small for *at least one* k, which corresponds to a direction along which the content is modellable. Therefore, UNIWARD discourages changes even at clean edges and instead forces the embedding to use textured regions whenever possible. Such a selection channel adapts to the content more strongly than that of HUGO (see Figure 1).

The distortion function (1) is non-additive in the sense that the distortion introduced by changing a pair of nearby pixels is not equal to the sum of distortions when changing each pixel individually (and not changing the other pixel). All steganographic schemes proposed in Ref. [8] used the so-called additive approximation of

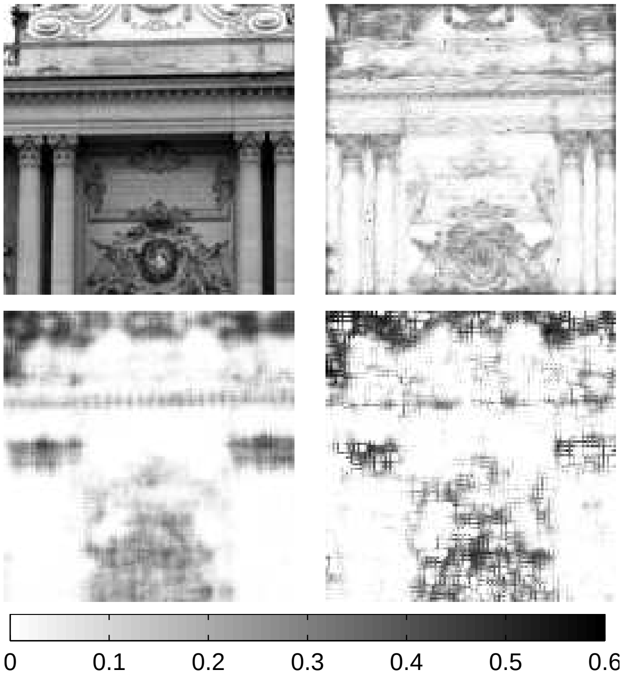

Figure 1. Embedding probabilities for relative payload $\alpha = 0.4$ bpp using HUGO (top right), WOW (bottom left), and S-UNIWARD (bottom right) for a 128×128 grayscale cover image (top left).

$D(\mathbf{X}, \mathbf{Y})$, which allowed a straightforward implementation using STCs.[1] The cost ρ_{ij} of modifying pixel ij from its cover value X_{ij} to Y_{ij}, and leaving all other cover elements unchanged, is:

$$\rho_{ij}(\mathbf{X}, Y_{ij}) \triangleq D(\mathbf{X}, \mathbf{X}_{\sim ij}Y_{ij}), \tag{2}$$

where $\mathbf{X}_{\sim ij}Y_{ij}$ is the cover image $\mathbf{X}$ with only its ijth element changed: $X_{ij} \to Y_{ij}$.[*] Note that since the support of 8-tap Daubechies wavelets is 8×8 pixels, computing each cost 2 involves summing up 3×16^2 pixels because there are three subbands. Due to the absolute values in (1), $\rho_{ij}(\mathbf{X}, X_{ij} + 1) = \rho_{ij}(\mathbf{X}, X_{ij} - 1)$, which allows S-UNIWARD to employ a ternary embedding operation and codes.

Since we describe an attack on the spatial version of UNIWARD called S-UNIWARD, we do not further explain how this distortion function can be used for embedding in the JPEG and side-informed JPEG domains. In fact, it seems that the attack described below is applicable only to S-UNIWARD and not its JPEG forms. More on this issue appears in Section 5.

To demonstrate the properties of the embedding function S-UNIWARD, in Figure 1 we show one cropped BOSSbase 1.01 image (top left) together with the embedding probabilities for HUGO (top right), WOW (bottom left), and S-UNIWARD (bottom right). While the directional costs of both WOW and S-UNIWARD force the changes solely into the textured areas, the spatial distribution of embedding probabilities for WOW and S-UNIWARD are quite different. Most notably, the embedding probabilities of S-UNIWARD exhibit interleaved streaks of large and small values. This is caused by the properties of the distortion function − it is a sum of *relative* wavelet coefficient changes, which makes the costs very (and perhaps overly) sensitive to content. It takes only one wavelet coefficient (among 3×16^2 coefficients affected by changing a single pixel $X_{ij} \to Y_{ij}$) to be close to zero to have a very large embedding cost ρ_{ij}. In contrast, since the embedding costs of WOW are obtained by adding reciprocal values of merely three "embedding suitabilities," which are themselves sums over many wavelet coefficients, one encounters a high embedding cost in WOW much less likely than in S-UNIWARD. The authors of UNIWARD acknowledged this and also pointed out: "While the streaks may increase the statistical detectability, steganalysis with rich media models showed no evidence for this."

[*] This notation was used in Ref. [9] and is also standard in the literature on random fields.[13]

3. CONTENT-SELECTIVE RESIDUALS

In this section, we describe a new concept of a content-selective residual and introduce a new set of steganalysis features to attack S-UNIWARD.

Since the embedding costs of changing a pixel by 1 or -1 are the same and equal to ρ_{ij}, the probabilities of changing a pixel by 1 or -1 are also the same and equal to:[14]

$$p_{ij} = \frac{\exp(-\lambda\rho_{ij})}{1+\exp(-\lambda\rho_{ij})}, \tag{3}$$

where $\lambda > 0$ is a constant determined by the payload constraint:

$$\alpha MN = \sum_{ij} h_3(p_{ij}), \tag{4}$$

where $0 \leq \alpha \leq 1$ is the relative payload expressed in bits per pixel (bpp) and $h_3(x) = -x\log_2 x - (1-x)\log_2(1-x) + x$ is the ternary entropy function expressed in bits. Given image $\mathbf{X}$ and payload α, the probability of changing pixel ij during embedding is thus $2p_{ij}(\mathbf{X},\alpha)$. Even though in general, for a stego image $\mathbf{Y}$ embedded with payload α, $p_{ij}(\mathbf{X},\alpha) \neq p_{ij}(\mathbf{Y},\alpha)$, these probabilities are "largely similar." This approximate knowledge of the probabilistic selection channel, together with the "streak" artifacts can be used to construct an attack.

First, we separate the image into disjoint classes of pixels – those that are likely to be changed and those that are less likely to be changed during embedding. Then, we compute noise residuals from each class – the so-called content-selective residuals (CSRs). Finally, the feature vector is formed by first and second-order statistics of CSRs. What is remarkable (as will be seen below) is that even though the proposed detection statistics would be rather inefficient in detecting S-UNIWARD if computed from *all* pixels, when divided into the classes, their detection power immensely increases as we draw strength from the faulty selection channel.

For our attack, we will use residuals of the first, second, and third order:

$$R^{(1)}_{ij} = X_{i,j+1} - X_{ij}, \tag{5}$$

$$R^{(2)}_{ij} = X_{i,j+1} - 2X_{ij} + X_{i,j-1}, \tag{6}$$

$$R^{(3)}_{ij} = -X_{i,j+2} + 3X_{i,j+1} - 3X_{ij} + X_{i,j-1}. \tag{7}$$

Note that these residuals are computed in the horizontal direction. To increase the robustness of the features and make them better populated, we merge these residuals with those computed from the transposed image or, alternatively, one can think of applying the "vertical versions" of these residuals to the original image.

3.1 Pixel classes

Let $0 < t_s < t_L < 1$ be two fixed thresholds and $0 < \overline{\alpha} < 1$ a fixed relative payload. We will say that pixel X_{ij} is of 'type s' (**s**mall probability of embedding change) when $p_{ij}(\mathbf{X},\overline{\alpha}) < t_s$ and of 'type L' (**L**arge probability of an embedding change) when $p_{ij}(\mathbf{X},\overline{\alpha}) > t_L$, otherwise it has no type.

Note that each residual sample of order d, $R^{(d)}_{ij}$, involves $d+1$ adjacent pixels, Thus, we will collect the first-order statistics (histograms) of residuals separately for 2^{d+1} classes of $d+1$ neighboring pixels defined by their type. For example, for $d = 1$, a horizontally adjacent pixel pair $(X_{ij}, X_{i,j+1})$ can be of class $[s\ s]$, $[s\ L]$, $[L\ s]$, and $[L\ L]$ depending on the corresponding pixel classes. In general, for the residual of order d, $d+1$ neighboring pixels will be divided into 2^{d+1} classes, $\mathcal{C}^{(d)}_c$, $c \in \{1,\ldots,2^{d+1}\}$. We denote the set of residual values $R^{(d)}_{ij}$ collected only from class $\mathcal{C}^{(d)}_c$, $d \in \{1,2,3\}$, $c \in \{1,\ldots,2^{d+1}\}$ as $R^{(d)}_{ij}(\mathcal{C}^{(d)}_c)$.

When collecting the second-order statistics of residuals, we will work with classes consisting of $d+2$ neighboring pixels as a pair of adjacent residuals, $(R^{(d)}_{ij}, R^{(d)}_{i,j+1})$, involves $d+2$ pixels.

Residual order d	1	2	3
Histogram	$3(2T_h+1)$	$6(2T_h+1)$	$10(2T_h+1)$
Acronym	1st 1D	2nd 1D	3rd 1D
2D co-occurrence	$6(2T_c+1)^2$	$10(2T_c+1)^2$	-
Acronym	1st 2D	2nd 2D	-

Table 1. The feature dimension and the acronym for each content-selective residual type and its representation.

3.2 Histograms

To curb the residuals' range and allow a compact representation, the residuals of all orders $d \in \{1,2,3\}$ are truncated to the range $[-T_h, T_h]$, $R_{ij}^{(d)} \leftarrow \mathrm{trunc}_{T_h}(R_{ij}^{(d)})$, where

$$\mathrm{trunc}_T(x) = \begin{cases} x & \text{when } -T \le x \le T \\ -T & \text{when } x < -T \\ T & \text{when } T \le x. \end{cases} \tag{8}$$

For a fixed residual order d and class $\mathcal{C}_c^{(d)}$, we denote with $h_c^{(d)}(l)$, $l \in \{-T_h, \dots, T_h\}$, the histogram of $R_{ij}^{(d)}(\mathcal{C}_c^{(d)})$ over all ij. The histograms can be further compacted using the directional symmetries of natural images. In particular, for a class $\mathcal{C}$, let $\overleftarrow{\mathcal{C}}$ denote the mirror image of $\mathcal{C}$. For example, $[s\ L] = \overleftarrow{[L\ s]}$, $[s\ L\ s\ L] = \overleftarrow{[L\ s\ L\ s]}$, etc. By inspecting the definition of the residuals (5)–(7), it is easy to see that for d odd, $R_{ij}^{(d)}(\mathcal{C}) = -R_{ij}^{(d)}(\overleftarrow{\mathcal{C}})$ and $R_{ij}^{(d)}(\mathcal{C}) = R_{ij}^{(d)}(\overleftarrow{\mathcal{C}})$ for $d = 2$. Given the fact that the residuals themselves are distributed symmetrically around zero, we will add the histograms of classes that are mirror images of each other. This reduces the number of classes we need to consider from $2^2 = 4$ to 3 for the first-order residuals, from $2^3 = 8$ to 6 for the second-order residuals, and from $2^4 = 16$ to 10 for the third-order residuals.

To summarize, for residual order $d = 1, 2, 3$, the feature vector is formed by 3, 6, and 10 histograms, each holding $2T_h + 1$ values. Thus, the first part of our feature vector formed by histograms has the dimensionality of $19 \times (2T_h + 1)$.

3.3 Co-occurrences

The steganalysis features formed by histograms of residuals will be further supplemented with two-dimensional (2D) co-occurrence matrices (sample joint probability distributions). To keep the dimensionality of the co-occurrences low, we truncate with a lower value of the truncation threshold, T_c, than for the residuals, $T_c < T_h$. We also use 2D co-occurrences only with residuals of the first and second order.

We can apply the same symmetrization to the 2D co-occurrences as we did with the histograms with one small change. When merging the co-occurrences for two mirror-image classes, we need to transpose one of the matrices. This can be easily seen on the example of a co-occurrence for two second-order residuals $R_{ij}^{(2)}$ and $R_{i,j+1}^{(2)}$. For the class $\mathcal{C} = [s\ L\ s\ L]$, $R_{ij}^{(2)}$ is computed from a pixel triple $[s\ L\ s]$ while $R_{i,j+1}^{(2)}$ is computed from $[L\ s\ L]$. For the mirror class, $\overleftarrow{\mathcal{C}} = [L\ s\ L\ s]$, the pixel triples for each residual exchange. Thus, the total number of classes for residuals of order $d = 1$ is down from $2^3 = 8$ to 6 and from $2^4 = 16$ to 10 for $d = 2$. The total feature dimensionality of co-occurrences from both residuals is thus $6 \times (2T_c + 1)^2 + 10 \times (2T_c + 1)^2$.

To summarize, the feature vector is formed by 19 histograms of residuals with orders 1,2, and 3 and 16 2D co-occurrence matrices for residuals of order 1 and 2. The dimensionality broken down per residual order and representation type (histogram vs. co-occurrence) is shown in Table 1. In this table we also introduce acronyms for each type of features for easier referencing.

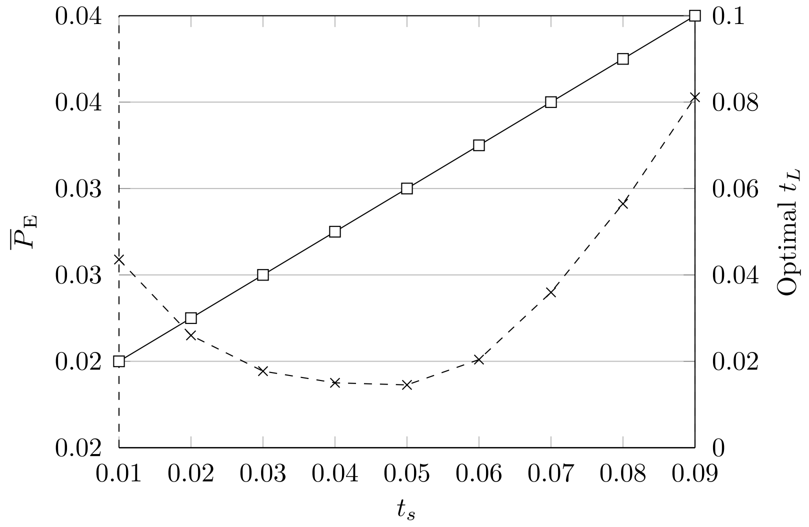

Figure 2. The value of t_L (dashed line) that minimizes the detection error $\overline{P}_{\mathrm{E}}$ and its value (solid line) as a function of t_s. The chart was computed for the 1st 1D CSR with $T_h = 10$ and $\alpha = \overline{\alpha} = 0.4$ bpp. The optimal values of the parameters are $t_s = 0.05$, $t_L = 0.06$.

4. ATTACKING S-UNIWARD

All our experiments were run on the standard database BOSSbase 1.01.[15] This source contains 10,000 images acquired by seven digital cameras in the RAW format (CR2 or DNG) and subsequently processed by converting them to 8-bit grayscale, resizing, and central-cropping to 512 × 512 pixels. The script for this processing is also available from the BOSS competition web site.

We report the detection performance using the average detection error $\overline{P}_{\mathrm{E}} = (P_{\mathrm{FA}} + P_{\mathrm{MD}})/2$ estimated by training on a random half of the BOSSbase database and testing on the remaining half and averaging over ten database splits. The symbols $P_{\mathrm{FA}}, P_{\mathrm{MD}}$ stand for the probability of false alarm and missed detection. The classifier was the ensemble 2.0[†] with auto settings for the optimal subspace dimensionality and the number of base learners.

4.1 Parameter setting

The attack on UNIWARD described in the previous section depends on several parameters that need to be set. First, we determine the proper values of the thresholds t_s and t_L for the embedding change probabilities controlling the pixel classes as well as the truncation thresholds T_h and T_c for histograms and co-occurrences. Additionally, because the steganalyst will in general not know the true payload α, we need to use a fixed value of the relative payload, $\overline{\alpha}$, for which the embedding change probabilities p_{ij} will be computed. The purpose of the first set of experiments in this subsection is to determine the values of t_s, t_L, T_h, T_c, and $\overline{\alpha}$.

For the first experiment, we fixed $T_h = 10$, $\overline{\alpha} = 0.4$ and computed the best values of t_s and t_L based only on the 1st 1D CSR (see Figure 2). The optimal values of these thresholds necessarily depend on the cover source and will need to be adjusted for a different source of images than the BOSSbase.

Next, with the thresholds t_s and t_L fixed at $t_s = 0.05$ and $t_L = 0.06$, we investigated the effect of the parameter $\overline{\alpha}$. Figure 3 shows $\overline{P}_{\mathrm{E}}$ obtained using the 1st 1D CSR as a function of the payload α for three different values of $\overline{\alpha}$. Note that, in general, when the real payload α is unknown, it is better to use a larger testing payload $\overline{\alpha}$. Also notice the unusual non-monotone dependence of $\overline{P}_{\mathrm{E}}$ on payload. This is due to the fact that with higher embedding payload, the differences between the interleaved streaks are smaller (as S-UNIWARD loses its content adaptivity), which weakens the proposed attack for larger payloads. As expected, for small payloads, the detection error eventually increases as it becomes more difficult to detect a small number of embedding changes.

[†] `http://dde.binghamton.edu/download/ensemble/`

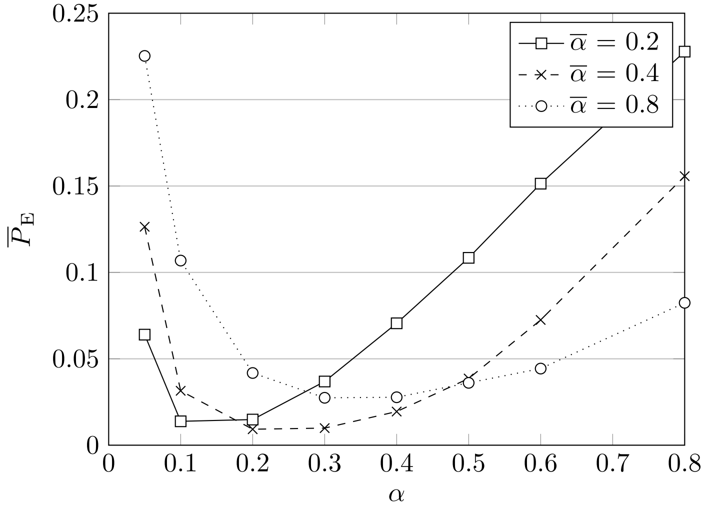

Figure 3. Detection error $\overline{P}_{\mathrm{E}}$ as a function of α for three different values of $\overline{\alpha}$ for the 1st 1D CSR with $T_h = 10$.

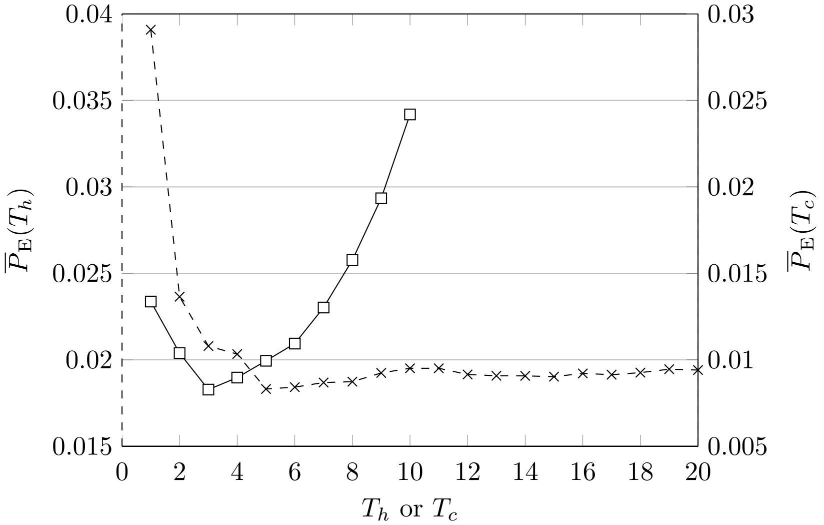

Figure 4. Detection error $\overline{P}_{\mathrm{E}}$ as a function of T_h and T_c for the 1st 1D and 1st 2D CSR, respectively. The rest of the parameters were set to $\alpha = \overline{\alpha} = 0.4$, $t_s = 0.05$, $t_L = 0.06$.

To see the effect of the truncation thresholds T_h and T_c, in Figure 4 we show the detection error as a function of the thresholds for α and $\overline{\alpha}$ fixed at 0.4. Based on this experiment, we fixed $T_h = 10$ and $T_c = 3$ for all CSRs in this paper. Notice that it is possible to achieve a low detection error with the nine-dimensional (!) 1st 1D CSR with $T_h = 1$ even for small payloads α. This indicates a serious security weakness of S-UNIWARD caused by the strong artifacts in the selection channel.

The main reason why the proposed attack works so well for S-UNIWARD are the interleaved streaks of high and low embedding probabilities in p_{ij}. They allowed us to split pixels, which have similar statistical properties, into classes that do not change with embedding much (e.g., the class $[s\ s]$) and classes that do get affected by embedding – the classes $[s\ L]$, $[L\ s]$, and $[L\ L]$. For example, for the 1st 1D CSR with $T_h = 10$ and $\alpha = \overline{\alpha} = 0.4$, the detection performance of the class $[s\ s]$ by itself is poor – $\overline{P}_{\mathrm{E}} = 0.4757$. The union of classes $[s\ L]$, $[L\ s]$ detects better at $\overline{P}_{\mathrm{E}} = 0.3405$ because it is more affected by embedding. However, when adding the class $[s\ s]$ to the union of $[s\ L]$ and $[L\ s]$, the detection error suddenly drops to 0.0734. This is because the statistic collected from pixel class $[s\ s]$ serves as a powerful *reference*. Further adding the third class $[L\ L]$, which by itself detects at $\overline{P}_{\mathrm{E}} = 0.3376$, makes the detection error drop to 0.0197 for the complete 63-dimensional 1st 1D CSR.

	Dim	Content-selective	Dim	Non-selective
1st 1D	63	0.01841	21	0.47645
1st 2D	294	0.00852	49	0.46403
2nd 1D	126	0.00943	21	0.46578
2nd 2D	490	0.00687	49	0.42767
3rd 1D	210	0.00740	21	0.43075
Combined	1183	0.00470	161	0.41380

Table 2. Detection error $\overline{P}_{\mathrm{E}}$ for all CSRs and their union in their content-selective and non-selective versions for $\alpha = 0.4$ bpp.

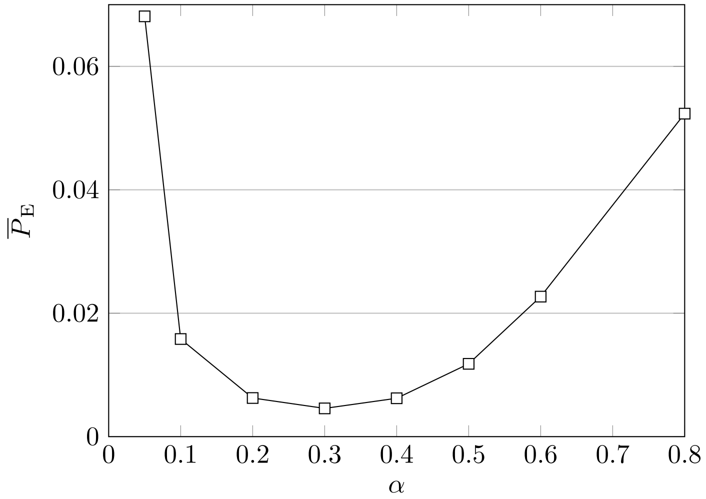

Figure 5. Detection error $\overline{P}_{\mathrm{E}}$ as a function of α for the CSR features.

Note that the CSRs will not work well for HUGO or WOW because their embedding change probabilities are more correlated with the content. The differences of pixel pairs that change have a different distribution than those of pairs that do not change. Thus, one cannot utilize a reference for detection.

4.2 Attacking S-UNIWARD

Based on the experiments in Section 4.1, we select for the attack the combination of three histograms and two 2D co-occurrences for three residuals listed in Table 1. The parameters are fixed to $t_s = 0.05$, $t_L = 0.06$, $\overline{\alpha} = 0.4$, $T_h = 10$, and $T_c = 3$.

The detection power of each individual CSR and their combination is shown in the third column of Table 2. Notice that the detection error strongly correlates with feature dimensionality. Each content-selective residual individually can detect S-UNIWARD quite reliably. While merging the features does help lower the error, the improvement is not dramatic. To prove that the detection power resides in the faulty selection channel, we supply in the fifth column of the table the detection errors when steganalyzing S-UNIWARD with histograms and 2D co-occurrences computed from all residuals without dividing them into classes (non-selective residuals). Since these features have no knowledge of the probability map, they do not exploit the faulty selection channel, and their detection is poor. This motivates the fix of S-UNIWARD proposed in the next section. It focuses on removing the artifacts from the embedding change probabilities.

As our final experiment, we computed the detection error of the 1183-dimensional combined CSR features across different payloads. The results shown in Figure 5 confirm that the CSR features can reliably detect S-UNIWARD for a wide range of stego payloads. The unusual increase of the detection error for large payloads is due to two effects, both related to the fact that the algorithm has to change a larger amount of pixels. Because for large payloads the embedding change probabilities will be generally larger, the thresholds t_s and t_L are no longer

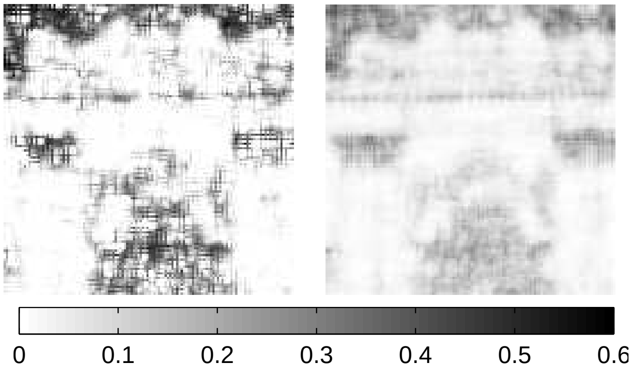

Figure 6. Embedding probabilities for relative payload $\alpha = 0.4$ bpp for the image shown in the upper left corner of Figure 1 for S-UNIWARD with $\sigma = 10 \times \text{eps} \approx 2 \times 10^{-15}$ (left) and $\sigma = 1$ (right).

optimal. Second, the artifacts in the embedding change probabilities become less prominent as S-UNIWARD loses its adaptivity.

5. FIXING S-UNIWARD

As discussed in Section 1, one of the reasons for the interleaved streaks in the embedding change probabilities is the fact that the distortion is a sum of relative changes of wavelet coefficients. The stabilizing constant σ in the definition of the UNIWARD embedding distortion is too small ($\sigma = 10 \times \text{eps}$, eps as in Matlab), causing an unstable behavior and making UNIWARD "overly sensitive" to content. We observed that with increasing value of σ, the interleaved streaks became much less pronounced (see Figure 6), which will likely diminish the strength of the CSR features. Increasing the value of σ could, however, negatively impact the security of S-UNIWARD w.r.t. rich models.

To investigate this issue, we carried out the following experiment. For a wide range of σ and fixed $\alpha = \overline{\alpha} = 0.4$, we computed the detection error when steganalyzing S-UNIWARD using the CSR features, the SRM, and their merger. We note that for estimating the embedding change probabilities p_{ij} for the CSRs, we *always* used S-UNIWARD with $\sigma = 10 \times \text{eps}$ no matter what value of σ was used for creating the stego images. We did this for two reasons. First, the original low value of σ allows us to better estimate the location of the streaks. Second, by using a fixed value of σ, we could use the same thresholds t_s and t_L and avoided having to optimize them for each σ.

Figure 7 shows that the faulty selection channel undermines the security of S-UNIWARD until $\sigma \approx 2^{-4}$, while at $\sigma \approx 2^{-3}$ the CSR features cease to detect the embedding. At the same time, the security w.r.t. the SRM remains essentially unchanged with increasing σ until $\sigma \approx 1$, after which the error quickly drops. The merger of the CSR features with the SRM exhibits the highest detection error at $\sigma \approx 1$, which we recommend as the proper value of the stabilizing constant for S-UNIWARD.

To show that the CSR features are no longer capable of utilizing the faulty selection channel, we repeat the experiment from Section 4 and produce a analogue of Table 2. Table 3 shows detection error of CSR and the residuals computed from all pixels (without dividing into classes). Although the individual CSRs do have a small edge over their non-selective counterparts, when merged together the detection error of both becomes quite similar. The CSRs no longer gain much from any leftover artifacts in the embedding change probabilities.

Finally, in Figure 8 we show the detection error of S-UNIWARD implemented with $\sigma = 1$ as a function of relative payload, α, when steganalyzing with CSR features, SRM, and their union. The CSR features have a rather weak performance, the error trend is no longer pathological (the detection error decreases with increasing payload), and merging the CSR features with the SRM does not lead to any significant improvement in the detection.

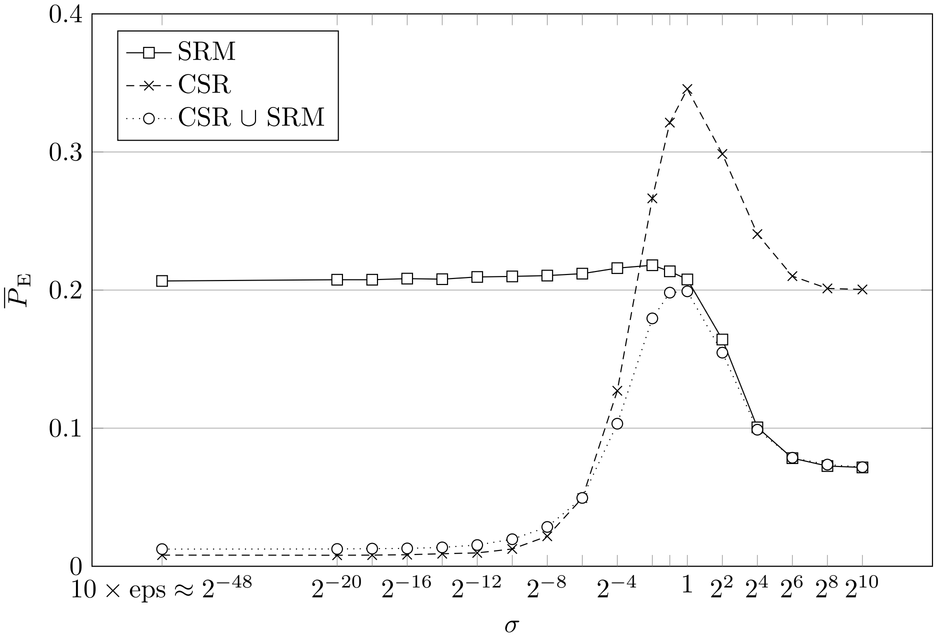

Figure 7. Detection error $\overline{P}_{\mathrm{E}}$ of S-UNIWARD implemented with a range of values for σ when steganalyzing with the CSR features, the SRM, and their union. The value of σ that seems to provide the best overall security is $\sigma \approx 1$.

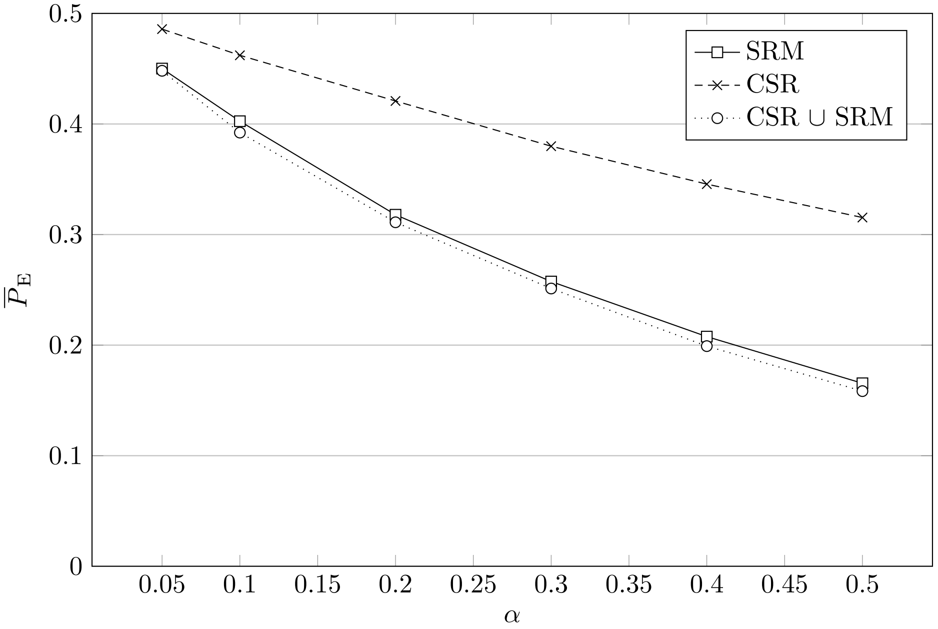

Figure 8. Detection error $\overline{P}_{\mathrm{E}}$ of S-UNIWARD with $\sigma = 1$ as a function of α when using the CSR features, the SRM, and their union.

	Dim	Content-selective	Dim	Non-selective
1st 1D	63	0.45030	21	0.46973
1st 2D	294	0.39877	49	0.45705
2nd 1D	126	0.42544	21	0.46079
2nd 2D	490	0.36998	49	0.38992
3rd 1D	210	0.40055	21	0.43796
Combined	1183	0.35310	161	0.37790

Table 3. Detection error $\overline{P}_{\mathrm{E}}$ for all five residual types and their union in their content-selective and non-selective versions for relative payload $\alpha = 0.4$ bpp and $\sigma = 1$.

The experiments in this section confirm that a proper adjustment of the parameter σ is an effective measure against CSR features that utilize artifacts in the probabilistic selection channel without affecting the security of S-UNIWARD w.r.t. SRM. This situation appears somewhat reminiscent of the flaw (and its quick fix) found in HUGO by Gökhan Gül[16] during the BOSS competition[15] and later explained in Ref [17].

We acknowledge that more extensive experiments are needed to see the impact of changing σ on the security of JPEG and side-informed JPEG versions of UNIWARD. However, the CSR features are unlikely to be effective against JPEG implementations of UNIWARD because the distortion caused by a change in a DCT coefficient affects a block of 8×8 pixels and, consequently, any potential artifacts average out. Experiments with the SRM only (not shown in this paper) indicate that the JPEG versions of UNIWARD exhibit the highest security when selecting $\sigma \approx 2^{-6}$.[18]

6. CONCLUSIONS

The basic premise of content adaptive steganography is that a higher embedding security can be achieved by restricting the embedding changes to regions in the image that contain complex content because the Warden will be unable to detect the traces of embedding in such regions. A potential problem of adaptive schemes lies in the fact that the selection channel is dictated by the content, which means that it is also approximately available to the Warden, who can adjust her detector accordingly. During the BOSS competition, numerous participants attempted to use the approximate knowledge of HUGO's selection channel to improve their attack, but, to the best knowledge of the authors of this article, no one succeeded. The first case when a knowledge of the selection channel was exploited to improve the detection of steganography was described by Schöttle et al. [19]. The authors showed that the performance of the weighted stego-image detector[20] can be improved for a crippled steganography method that follows the so-called naive content-adaptive embedding paradigm. Böhme et al. [21] have proposed to capture the interaction between the Warden and the steganographer using an adaptive embedding via the Game Theory. In particular, they showed on a toy example that it is more advantageous for the sender to embed according to a strategy corresponding to the Nash equilibrium than to minimize the KL divergence between cover and stego distributions under an omnipotent Warden.

In this paper, we show for the first time for a non-trivial steganographic method that an approximate knowledge of the probabilistic selection channel can be exploited for detection. In particular, we utilize artifacts in the selection channel caused by an improper value of a stabilizing constant in the spatial-domain steganographic algorithm called S-UNIWARD as originally described in [8]. The embedding probabilities of S-UNIWARD exhibit interleaved streaks of high and low embedding probabilities, which allowed us to compute the statistics of noise residuals across groups of pixels that, with a high probability, do not change and those groups that do change – the so-called content-adaptive residuals. The former serve as a reference for the latter, enabling a rather accurate detection even for small embedding payloads with as few as nine features.

We also describe a way to correct for the faulty selection channel by properly adjusting the value of the stabilizing constant of UNIWARD. Tests on the union of the content-adaptive residuals and the spatial rich model show that once the artifacts in the embedding probabilities are suppressed the proposed attack is no longer effective.

The lesson to be learned from this work is that one needs to be cautious when designing content-adaptive steganography. The selection channel needs to be correlated with the content to prevent the attacker from splitting pixels with similar statistical properties into those that are likely to be modified and those that are unlikely to be modified during embedding.

7. ACKNOWLEDGMENTS

The work on this paper was supported by Air Force Office of Scientific Research under the research grant number FA9950-12-1-0124. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of AFOSR or the U.S. Government.

REFERENCES

1. Filler, T., Judas, J., and Fridrich, J., "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security* **6**, 920–935 (September 2011).
2. Böhme, R., [*Advanced Statistical Steganalysis*], Springer-Verlag, Berlin Heidelberg (2010).
3. Pevný, T., Filler, T., and Bas, P., "Using high-dimensional image models to perform highly undetectable steganography," in [*Information Hiding, 12th International Conference*], Böhme, R. and Safavi-Naini, R., eds., Lecture Notes in Computer Science **6387**, 161–177, Springer-Verlag, New York, Calgary, Canada (June 28–30, 2010).
4. Pevný, T., Bas, P., and Fridrich, J., "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security* **5**, 215–224 (June 2010).
5. Filler, T. and Fridrich, J., "Design of adaptive steganographic schemes for digital images," in [*Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics III*], Alattar, A., Memon, N. D., Delp, E. J., and Dittmann, J., eds., **7880**, OF 1–14 (January 23–26, 2011).
6. Guo, L., Ni, J., and Shi, Y.-Q., "An efficient JPEG steganographic scheme using uniform embedding," in [*Fourth IEEE International Workshop on Information Forensics and Security*], (December 2–5, 2012).
7. Holub, V. and Fridrich, J., "Designing steganographic distortion using directional filters," in [*Fourth IEEE International Workshop on Information Forensics and Security*], (December 2–5, 2012).
8. Holub, V. and Fridrich, J., "Digital image steganography using universal distortion," in [*1st ACM IH&MMSec. Workshop*], Puech, W., Chaumont, M., Dittmann, J., and Campisi, P., eds. (June 17–19, 2013).
9. Filler, T. and Fridrich, J., "Gibbs construction in steganography," *IEEE Transactions on Information Forensics and Security* **5**(4), 705–720 (2010).
10. Fridrich, J. and Kodovský, J., "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security* **7**, 868–882 (June 2011).
11. Kodovský, J. and Fridrich, J., "Steganalysis of JPEG images using rich models," in [*Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012*], Alattar, A., Memon, N. D., and Delp, E. J., eds., **8303**, 0A 1–13 (January 23–26, 2012).
12. Filler, T., Pevný, T., and Bas, P., "BOSS (Break Our Steganography System)." `http://www.agents.cz/boss` (July 2010).
13. Winkler, G., [*Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction (Stochastic Modelling and Applied Probability)*], Springer-Verlag, Berlin Heidelberg, 2nd ed. (2003).
14. Fridrich, J. and Filler, T., "Practical methods for minimizing embedding impact in steganography," in [*Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX*], Delp, E. J. and Wong, P. W., eds., **6505**, 02–03 (January 29–February 1, 2007).
15. Bas, P., Filler, T., and Pevný, T., "Break our steganographic system – the ins and outs of organizing BOSS," in [*Information Hiding, 13th International Conference*], Filler, T., Pevný, T., Ker, A., and Craver, S., eds., Lecture Notes in Computer Science **6958**, 59–70 (May 18–20, 2011).

16. Gül, G. and Kurugollu, F., “A new methodology in steganalysis : Breaking highly undetactable steganograpy (HUGO),” in [*Information Hiding, 13th International Conference*], Filler, T., Pevný, T., Ker, A., and Craver, S., eds., Lecture Notes in Computer Science, 71–84 (May 18–20, 2011).
17. Kodovský, J., Fridrich, J., and Holub, V., “On dangers of overtraining steganography to incomplete cover model,” in [*Proceedings of the 13th ACM Multimedia & Security Workshop*], Dittmann, J., Craver, S., and Heitzenrater, C., eds., 69–76 (September 29–30, 2011).
18. Holub, V. and Fridrich, J., “Universal distortion design for steganography in an arbitrary domain,” *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop* (2013). Under review.
19. Schöttle, P., Korff, S., and Böhme, R., “Weighted stego-image steganalysis for naive content-adaptive embedding,” in [*Fourth IEEE International Workshop on Information Forensics and Security*], (December 2–5, 2012).
20. Ker, A. D. and Böhme, R., “Revisiting weighted stego-image steganalysis,” in [*Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*], Delp, E. J., Wong, P. W., Dittmann, J., and Memon, N. D., eds., **6819**, 5 1–17 (January 27–31, 2008).
21. Böhme, R. and Schöttle, P., “A game-theoretic approach to content-adaptive steganography,” in [*Information Hiding, 14th International Conference*], Kirchner, M. and Ghosal, D., eds., Lecture Notes in Computer Science **7692**, 125–141 (May 15–18, 2012).