Challenging the Doctrines of JPEG Steganography

Vojtěch Holub and Jessica Fridrich

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

What is this paper about?

This short paper focuses on steganography and steganalysis of grayscale JPEG images.

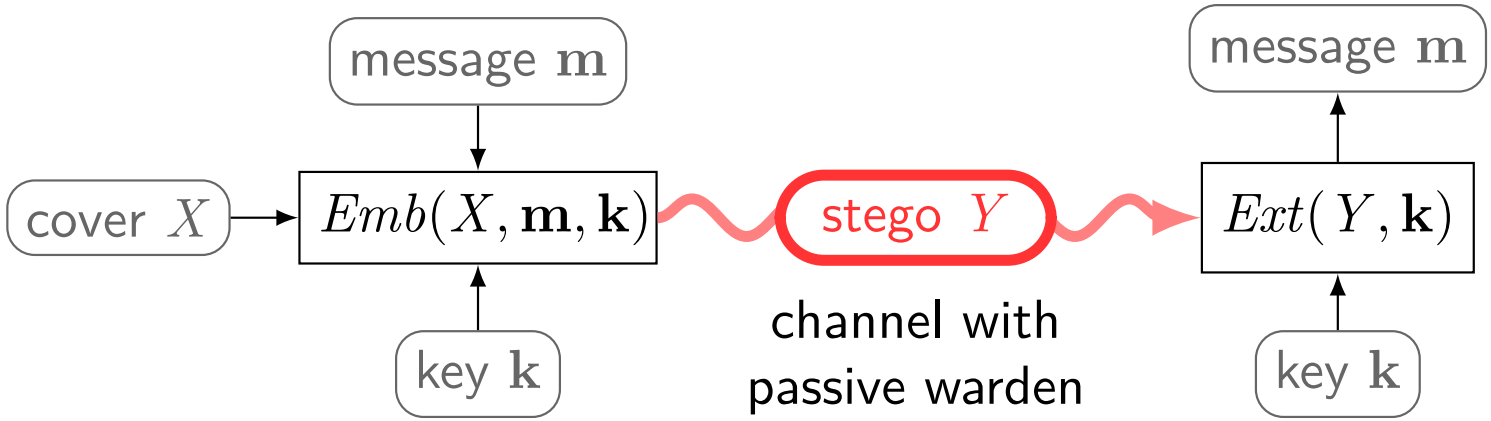
It gathers empirical evidence (scattered in other papers) to question the following long-lived doctrines:

1. Statistical impact of embedding should be evaluated in the embedding domain (steganography).
2. Most accurate detection is achieved in the embedding domain (steganalysis).

This paper does not give answers, it raises more questions.

Steganography and steganalysis

- **Steganography is the art of secret communication**



- **Steganographer's job**
  Modify a cover image to stego image so that it contains a secret message (by flipping LSBs, changing DCT coefficients, ...).
  Goal: make the embedding changes statistically undetectable.
- **Warden's job**: Distinguish between cover and stego images by building a detector. If cover source is known, the best detection is achieved using feature-based steganalysis and machine learning.

Modern JPEG steganography

- Modern steganography embeds messages while minimizing a distortion function.
- $\rho_{ij}(\mathbf{X}, Y_{ij})$ is the cost of changing cover DCT coefficient X_{ij} to stego coefficient Y_{ij}.
- Only additive distortions are taken into consideration – assumption that the embedding changes do not interact:

$$D(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho_{ij}(\mathbf{X}, Y_{ij})$$

- The actual embedding that minimizes the distortion function $D(\mathbf{X}, \mathbf{Y})$ is executed using Syndrome-Trellis Codes (STCs) (Filler, 2011).

Examples of distortion functions (1)

- Change rate

$$\rho_{ij} = \begin{cases} 0 & Y_{ij} = X_{ij} \\ 1 & Y_{ij} = X_{ij} \pm 1 \end{cases} \Longrightarrow D(\mathbf{X}, \mathbf{Y}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} [\, Y_{ij} \neq X_{ij}]$$

- Inverse of coefficient value

$$\rho_{ij} = \frac{1}{|X_{ij}|}$$

  - More frequent coefficient values (close to 0) are changed with smaller probability.
  - Less frequent (high values) are changed more often.
  - Avoids embedding in smooth image areas.
  - Part of UED distortion (Guo, 2012).

Examples of distortion functions (2)

- Side-informed steganography
  - Sender has uncompressed version of the JPEG image (precover $\mathbf{P}$).
  - $D_{ij} =$ raw DCT coefficients from uncompressed precover $\mathbf{P}$.
  - Sender rounds D_{ij} **up** or **down** to modulate its parity.
  - Embedding change probability depends on the quantization error $e_{ij} = |D_{ij} - X_{ij}|, \;\; e_{ij} \in [0, 0.5]$
    - when $e_{ij} \approx 0.5$, $\rho_{ij} \approx 0$, when $e_{ij} \approx 0$, ρ_{ij} is the largest
- Square loss distortion (side-informed):

$$\rho_{ij}^{(kl)} = \left(q_{kl}(1 - 2e_{ij})\right)^2$$

  - q_{kl} – quantization step for kl-th DCT mode
  - Since the cost is $\left\| |\mathbf{D} - \mathbf{X}| - |\mathbf{D} - \mathbf{Y}| \right\|_2$, one compute it in spatial domain (Parseval equality).

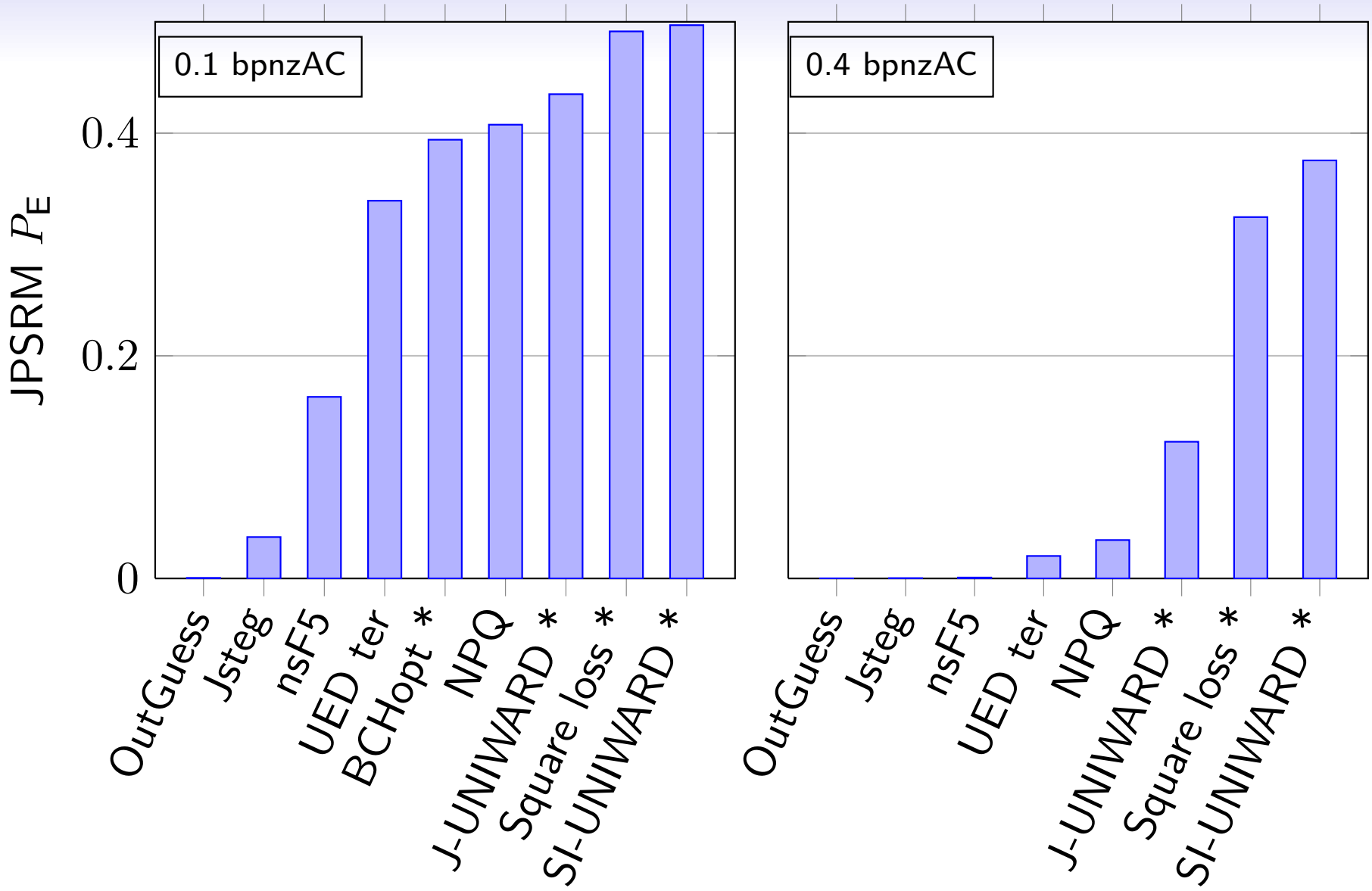
Published JPEG distortion functions

- nsF5 – Non-shrinkage F5 (Westfeld 2001 + improved coding)
- MOD – Model Optimized Distortion (Filler, 2011)
- UED – Uniform Embedding Distortion (Guo, 2012)
- NPQ – Normalized Perturbed Quantization (Huang, 2012) – side-informed
- UNIWARD family – UNIversal WAvelet Relative Distortion (Holub, 2013)
  - J-UNIWARD
  - SI-UNIWARD – side-informed

Minimizing spatial domain distortion

- Dependencies among DCT coefficients are very complex, a good model is unavailable.
- It is thus difficult to design a good distortion measure in the DCT domain.
- New trend is emerging: Determine the costs of changing a DCT coefficient in the spatial domain:
  - Unintentionally for side-informed BCHopt (Sachnev, 2009), EBS (Wang, 2012)
  - Purposely for any domain – UNIWARD family (J-, SI-) (Holub, 2013)

Experiments, BOSSbase 1.01, QF 75



Challenging the 1st doctrine

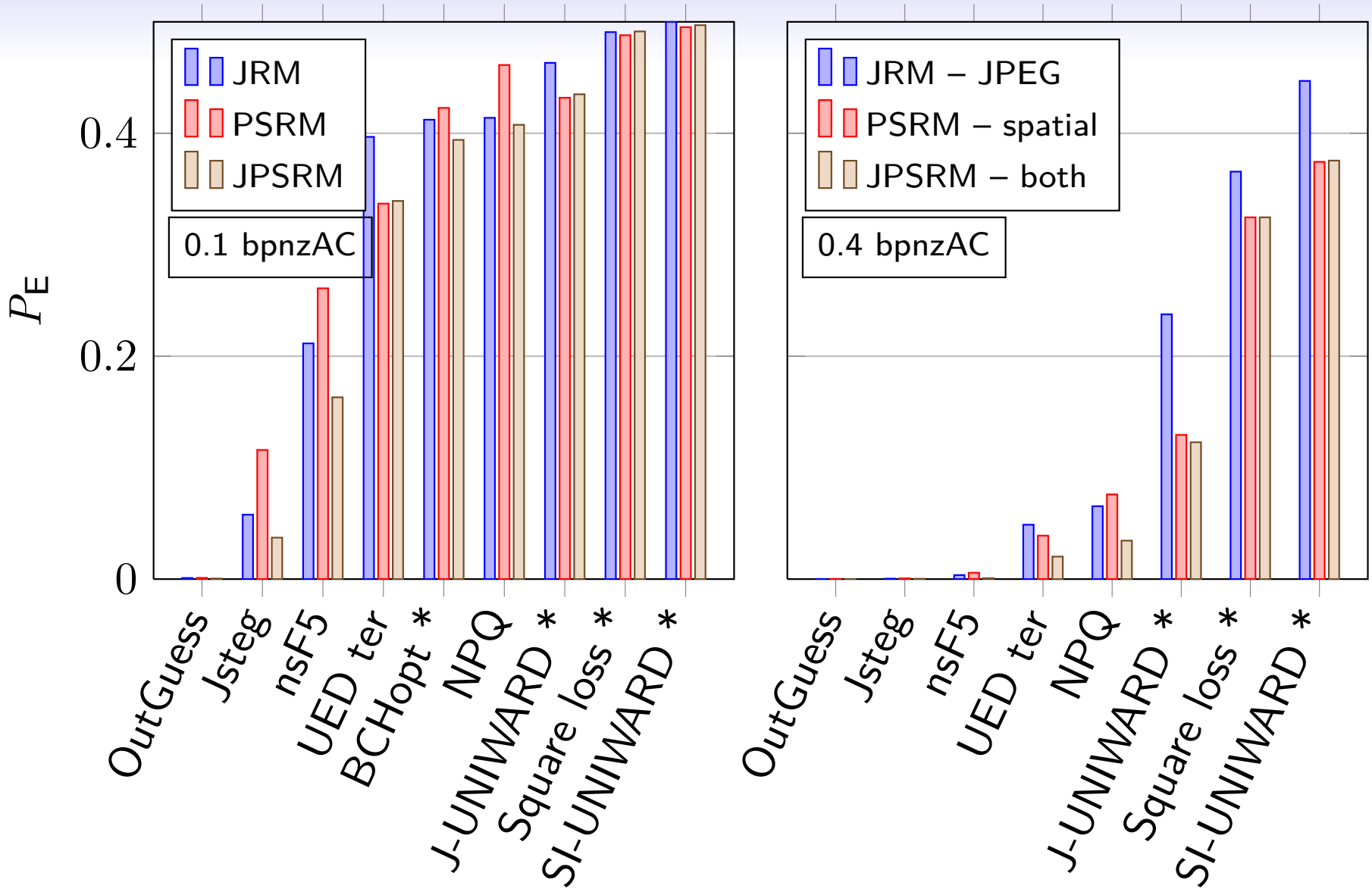
1st Doctrine: Statistical impact of embedding should be evaluated in the embedding domain

- The dependencies among DCT coefficients are too complex $\Longrightarrow$ compute distortion in spatial domain instead?
- Does minimization of a good spatial domain distortion lead to preservation of the unknown JPEG model? Or, do we simply have bad steganalysis features built from DCT coefficients?

Feature-based steganalysis

- Under the assumption that cover source is known, feature-based steganalysis with machine learning is currently the most accurate.
  1. Extract statistical features from a given image
  2. Detect using a trained classifier (machine learning)
- Best detection provided by rich media models:
  - For detection of spatial domain steganography: Projection Spatial Rich Model (PSRM - dim. 12870) (Holub, 2013)
  - For detection of JPEG steganography: JPEG Rich Model (JRM - dim. 22510) (Kodovský, 2012)
  - Merger of the two - JPSRM (dim 35380)
  - Classification using Ensemble classifier (Kodovský, 2012)

Experiments, BOSSbase 1.01, QF 75



Challenging the 2nd doctrine

2nd doctrine: The most accurate detection is achieved in the embedding domain

- Older algorithms (Jsteg, OutGuess, nsF5) are "faulty" and can be easily detected in JPEG domain.
- Newer algorithms (UNIWARDs, Square Loss) that minimize spatial domain distortion are much better detected in the spatial domain. Either they preserve dependencies of DCT coefficients well or our JPEG features are not sufficient.
- Adding JPEG features to steganalyze newer algorithms does not improve accuracy.

New steganalysis doctrine?

Most accurate detection is achieved in the domain where the distortion function is defined.