The optimal attack to histogram-based forensic detectors is simple(x)

Pedro Comesaña [#1], Fernando Pérez-González [#*2]

[#] *Signal Theory and Communications Department, University Vigo*
E. E. Telecomunicación, Campus-Lagoas Marcosende, Vigo 36310, Spain
[1] pcomesan@gts.uvigo.es
[2] fperez@gts.uvigo.es

[*] *Gradiant (Galician Research and Development Center in Advanced Telecommunications)*
Vigo 36310, Spain

***Abstract*—In the last years a number of counterforensics tools have been proposed. Although most of them are heuristic and designed *ad hoc*, lately a formal approach to this problem, rooted in transportation theory, has been pursued. This paper follows this path by designing optimal attacks against histogram-based detectors where the detection region is non-convex. The usefulness of our strategy is demonstrated by providing for the first time the optimal solution to the design of attacks against Benford's Law-based detectors, a problem that has deserved large practical interest by the forensic community. The performance of the proposed scheme is compared with that of the best existing counterforensic method against Benford-based detectors, showing the goodness (indeed, the optimality) of our approach.**

I. Introduction

Many fields of multimedia security, and signal processing in general, are perfect examples of a race between two parties with conflicting interests. This is the case, for example, of watermarking embedding and attacking, steganography and steganalysis, biometrics and spoofing, and shared-channel legal uses and jamming. In the last years, the general formulation of these scenarios, which is referred to as *Adversarial Signal Processing* [1], has received increasing attention.

Multimedia forensics, due to its nature, is also prone to this game between forensic detectors and attackers. Specifically, forensic detectors aim at determining whether a given content has undergone a certain processing, or even estimate the applied signal processing tools and their parameters; on the other hand, forensic attackers try to fool the detectors, by misleading their decisions/estimates.

The race between forensic detector and attacker can be even found in seminal works in this field. This is the case of the work by Popescu and Farid on resampling detection [2], where the authors do not only present their well-known detection method, but they also introduce an attack that would deceive resampling detectors. A number of other works can be found in the literature dealing with the design of attacks aimed at fooling JPEG detectors [3], [4], [5].

A common point of these works, as well as most other counterforensic methods in the literature, is their somewhat heuristic and *ad hoc* nature. In fact, most of these schemes are designed to deal with a particular forensic problem, and often to fool a specific detector, with a limited applicability to different scenarios. Furthermore, they typically show a clear lack of optimality: in fact, in most cases their optimality with respect to any meaningful criterion is not even discussed.

In this framework, Barni *et al.* [6] made an important contribution, introducing for the first time an attacking strategy that is general (i.e., non-targeted) and optimal. However, their approach is not consistent, in the sense that different target functions are considered at different stages of their scheme. Barni *et al.*'s method relies on transportation theory to address two different forensic problems: gamma-correction and histogram-stretching detection.

In a recent work, we have proposed another general counterforensic method [7] where a single target function is consistently optimized. Such method targets the so-called histogram-based forensic detectors, whose decisions are just based on the histogram of a function of the input signal.

Finally, in [8] Balado establishes connections between the counterforensics problem in [7], and permutation coding; through this coding paradigm, interesting mathematical links between counterforensics and steganography are also pointed out, thus enlarging the potential applicability of methods like the one proposed in this paper.

A. Forensic Benford's law-based detectors and counterforensics

Forensic detectors based on Most Significant Digits (MSD) have been devoted much attention in the last years. These detectors work with the relative frequency of the MSDs, and check the similarity of the resulting distribution with Benford's Law [9] (cf. Sect. II-D). Probably the main application of these methodology is the detection of JPEG compression (both single compression [10], and multiple compression [11], [12]).

Examples of adversarial signal processing can be also found in the literature to counteract these detectors. This is the case of [12], where the coefficients are sequentially modified depending on their absolute value; only transfers

WIFS'2014, December 3-5, 2014, Atlanta, Georgia, USA.

from those bins with a surplus of elements to those with a deficit are allowed, imposing a strong constraint on the attacker's strategy that will increase the distortion due to the histogram modification. On the other hand, in [13] an upper bound between each coefficient and its attacked version is established. Consequently, the latter scheme does not allow for producing arbitrary histograms; indeed, the output histogram might not be accurate for high frequencies. Finally, the attack proposed in [14] can produce any arbitrary histogram, based on the heuristic idea that those elements with the largest values should be modified the least possible, since, in general, their modification will imply a larger distortion than for the smaller values. According to the results reported in [14], the modification distortion achieved by such scheme is smaller than that in [12].

As it was already mentioned, all three schemes described in the last paragraph are heuristic. Following the framework described in [7], one of the aims of this work is to provide the optimal solution to the modification problem which produces a Benford's Law-compliant histogram.

The remaining of this paper is organized as follows: Sect. II provides the problem formulation, including a general framework description, a summary of the results reported in [7], and an introduction to Benford's domain. Sect. III presents the main results of this work, while Sect. IV reports the experimental results. Finally, Sect. V summarizes the main conclusions.

Concerning notation, bold fonts will denote vectors (e.g., $\mathbf{x}$), subindices will in general denote the component of a vector (e.g., x_i is the ith component of vector $\mathbf{x}$), and double subindices stand for the element of a matrix (e.g., $m_{i,j}$ is the (i,j)th element of matrix M). Furthermore, calligraphic capital letters are used for denoting sets (e.g., $\mathbf{x} \in \mathcal{X}$).

II. Problem formulation

First of all, we introduce the problem of attacking histogram-based forensic detectors; although the formulation of this problem was already posed in [7], it is also provided here for the sake of readability of the subsequent results. The main results presented in [7], and the constraints of that approach are also summarized; then, we explain how to remove some of those constraints by following a different optimization approach.

A. Framework description

Let $\mathbf{x}$ be a vector containing the samples of a discrete signal in its original space; $\mathbf{x}$ is assumed to belong to a set $\mathcal{X} \subset \mathbb{R}^N$, and it can be transformed through a function $f(\cdot)$, yielding $\mathbf{y} = f(\mathbf{x})$ ($\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^N$). We will assume $f(\cdot)$ to be a bijection (e.g., the identity function, the full-frame Discrete Fourier Transform (DFT), the block-Discrete Cosine Transform (DCT)).

The forensic detector $\phi_x : \mathcal{X} \mapsto \{0,1\}$ decides between two alternative hypotheses H_0 and H_1. For instance, H_1 can be "$\mathbf{x}$ was interpolated" and H_0 "$\mathbf{x}$ was not interpolated". Often the detector works in a transform domain, and consequently ϕ_x is expressed in terms of $\phi_y : \mathcal{Y} \mapsto \{0,1\}$, as $\phi_x(\mathbf{x}) = \phi_y(f(\mathbf{x}))$. Given ϕ_x, the *acceptance* and *rejection* regions for H_0 are defined in the original space as

$$\mathcal{R}_k^x \doteq \{\mathbf{x} \in \mathcal{X} : \phi_x(\mathbf{x}) = k\},\ k = 0, 1,$$

with a similar definition for $\mathcal{R}_k^y$, $k = 0, 1$.

Given a signal $\mathbf{x} \in \mathcal{R}_1^x$ and a distortion assessment function $g^x : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, a forensic attacker is interested in solving

$$\mathbf{x}^* = \arg \min_{\mathbf{x}' \in \mathcal{R}_0^x} g^x(\mathbf{x}, \mathbf{x}').$$

A typical choice for g^x is the squared Euclidean distance.

Based on the bijective nature of f, the previous problem definition is equivalent to

$$\mathbf{y}^* = \arg \min_{\mathbf{y}' \in \mathcal{R}_0^y} g^x(f^{-1}(\mathbf{y}), f^{-1}(\mathbf{y}')),$$

where $\mathbf{y}^* = f(\mathbf{x}^*)$.

Due to the bijectivity of f, there exists a distortion function $g^y : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ such that $g^x(f^{-1}(\mathbf{y}), f^{-1}(\mathbf{y}')) = g^y(\mathbf{y}, \mathbf{y}')$ (e.g., f is a wavelet transform and g^x is the Structural Similarity Index (SSIM) [15], or f is an orthonormal transform and g^x and g^y are Euclidean distances). Under this assumption, we can alternatively work in the transform domain, i.e.,

$$\mathbf{y}^* = \arg \min_{\mathbf{y}' \in \mathcal{R}_0^y} g^y(\mathbf{y}, \mathbf{y}'). \quad (1)$$

The value of the ith bin of the histogram of $\mathbf{y}$ is defined as

$$H(b_i, \mathbf{y}) \doteq \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}(y_j < b_i) \cdot \mathbf{1}(y_j \geq b_{i-1}), i = 1, \ldots, n_1,$$

and the corresponding value of the cumulative histogram as

$$H^c(b_i, \mathbf{y}) \doteq \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}(y_j < b_i), i = 1, \ldots, n_1,$$

where $\mathbf{1}(\cdot)$ is 1 if its Boolean argument is true, and 0 otherwise, and the histogram bins are delimited by the set of points $\mathcal{B} = \{b_0, b_1, \ldots, b_{n_1}\}$, where $b_0 < b_1 < \ldots < b_{n_1}$, $b_0 = -\infty$, $b_{n_1} = \infty$.[1] Therefore, the histogram of $\mathbf{y}$ is defined as $H(\mathcal{B}, \mathbf{y}) \doteq [H(b_1, \mathbf{y}),\ H(b_2, \mathbf{y}), \ldots, H(b_{n_1}, \mathbf{y})]$, while the set containing all the valid $H(\mathcal{B}, \mathbf{y})$ will be denoted by $\mathcal{H}$ (similarly, we will use $H^c(\mathcal{B}, \mathbf{y})$ and $\mathcal{H}^c$ for the cumulative histograms).

The test statistic ϕ_y is said to be *histogram-based* if there exists a function $\phi_H : \mathcal{H} \mapsto \{0,1\}$ such that $\phi_y(\mathbf{y}) = \phi_H(H(\mathcal{B}, \mathbf{y}))$ for all $\mathbf{y} \in \mathcal{Y}$. Therefore, given $\mathcal{B}$, for a histogram-based test we can define the equivalent acceptance and rejection sets as

$$\mathcal{R}_k^H \doteq \{H(\mathcal{B}, \mathbf{y}) : \phi_H(H(\mathcal{B}, \mathbf{y})) = k\},\ k = 0, 1.$$

We introduce $g^H : \mathcal{H} \times \mathcal{H} \mapsto \mathbb{R}$ to quantify the similarity between histograms:

$$g^H(H(\mathcal{B}, \mathbf{y}), H(\mathcal{B}, \mathbf{y}')) \doteq \min_{\mathbf{y}'' : H(\mathcal{B}, \mathbf{y}'') = H(\mathcal{B}, \mathbf{y}')} g^y(\mathbf{y}, \mathbf{y}''). \quad (2)$$

[1] This definition of the histogram and cumulative histogram is slightly different from that considered in [7], as in the current work the upper bound of the considered interval is open, and the lower bound closed.

B. Previous results

As discussed in [7], the solution to (1) is equivalent to solving

$$\mathbf{y}^* = \arg \min_{\mathbf{y}':H(\mathcal{B},\mathbf{y}')=H^\sharp(\mathcal{B},\mathbf{y}^\sharp)} g^y(\mathbf{y},\mathbf{y}'), \quad (3)$$

where

$$H^\sharp(\mathcal{B},\mathbf{y}^\sharp) = \arg \min_{H'(\mathcal{B},\mathbf{y}')\in\mathcal{R}_0^H} g^H(H(\mathcal{B},\mathbf{y}),H'(\mathcal{B},\mathbf{y}')). \quad (4)$$

In [7] we focused on the particular case where g^y is component-wise additive and dependent on the difference between the input vectors, i.e., $g^y(\mathbf{y},\mathbf{y}') = \sum_{i=1}^N g^{y_i}(y_i - y_i')$, where each g^{y_i} is convex, and $\mathcal{B} = \mathcal{Y}$. In such case, the solution to (1) given $H^\sharp(\mathcal{B},y^\sharp)$ is

$$y^*_{\pi_i} = y^\sharp_{\tau_i}, \quad i = 1,\ldots,N,$$

where π denotes an ordering permutation of $\mathbf{y}$, i.e., $y_{\pi_1} \leq y_{\pi_2} \leq \ldots y_{\pi_N}$, and τ a similarly defined ordering of $\mathbf{y}^\sharp$. Therefore, the optimization simply amounts to finding $H^\sharp(\mathcal{B},y^\sharp)$, that is, solving (4).

If the distortion function g^y is based on the Euclidean norm, then

$$g^H(H(\mathcal{B},\mathbf{y}),H(\mathcal{B},\mathbf{y}')) = \sum_{j=1}^N \left[(H^c)^{-1}\left(\tfrac{j}{N},\mathbf{y}\right) - (H^c)^{-1}\left(\tfrac{j}{N},\mathbf{y}'\right)\right]^2,$$

where one must take into account the straightforward relationship between H and H^c, and

$$(H^c)^{-1}(p,\mathbf{y}) \doteq \arg \min_{b_i,1\leq i\leq n_1:H^c(b_i,\mathbf{y})\geq p} b_i$$

is the inverse mapping of the cumulative histogram.

Consequently, the target function in (4) is convex and the optimization problem can be solved by off-the-shelf algorithms as long as $\mathcal{R}_0^H$ and the feasible region in (3) are also convex. The usefulness of the derived attack was shown in [7], where it was used for misleading a well-known double-JPEG compression detector [16] proposed by Pevny and Fridrich. Furthermore, the performance of the proposed attack is compared to that of other works in the literature specifically designed for double-JPEG detectors [17], showing the improvement achieved by the transportation-theoretic approach.

C. Need for the current approach

The main target of the current work is to provide the optimal solution to (4) for those cases where $\mathcal{R}_0^H$ and/or the target function are not convex, but g^y is still component-wise additive and dependent on the difference between the input vectors. There is a number of practical applications where this problem arises, that have received much attention in the literature, and whose optimal solution has not been found yet. These include:

- Benford's law-based detectors. This is a well-known problem in the literature (cf., the Introduction), but to the best of our knowledge the optimal solution (even for the Mean Square Error (MSE) distortion) is not known.
- Additive but non-convex distortion functions in the transform domain. This is the case, for example, of the measure proposed in [18], where the output of a steerable pyramid transform is fed to a normalizing function, which stands for the limited effect of any image on the visual system mechanisms. Similarly, in [19] the authors claim that there is no evidence that the relationship between the contrast and the perceived distortion is linear, but it can be probably better modeled by a sigmoid (monotonically increasing, but non-convex) function that takes into account the saturation effects typical of human senses.

D. Benford's domain definition

A set of numbers verify Benford's law if their MSD is distributed according to the following probability mass function (pmf)

$$P(d) = \log_{10}\left(1+\frac{1}{d}\right).$$

It is easy to check that the MSD of a non-null number x can be computed as

$$d(x) = \left\lfloor \frac{|x|}{10^{\lfloor \log_{10}(|x|)\rfloor}} \right\rfloor.$$

Definition: The coset representative of x in *Benford's domain* is defined as

$$c(x) \doteq \log_{10}(|x|) - \lfloor \log_{10}(|x|)\rfloor = \left[\log_{10}(|x|) - \frac{1}{2}\right] \mod \mathbb{Z} + \frac{1}{2},$$

where $x \mod \mathbb{Z}$ returns the quantization noise to the closest integer to x. □

With some abuse notation, we will denote by $c(\mathbf{x})$ the vector whose ith component is $c(x_i)$, $1 \leq i \leq N$.

In fact, $d(x)$ is a function of $c(x)$, as

$$d(x) = \left\lfloor 10^{c(x)} \right\rfloor.$$

Definition: Two vectors without null components $\mathbf{a}$ and $\mathbf{b}$ are said to be equivalent in Benford's domain iff for each component they share the coset representative in that domain, i.e., $c(a_i) = c(b_i)$, $1 \leq i \leq N$. □

Definition: Two vectors without null components $\mathbf{a}$ and $\mathbf{b}$ are said to have equivalent histograms in Benford's domain, denoted by $\mathbf{a} \sim \mathbf{b}$, iff $H(\mathcal{B},c(\mathbf{a})) = H(\mathcal{B},c(\mathbf{b}))$. □

Definition: Given an input vector $\mathbf{x}$, a forensic detector is said to be *histogram-based in Benford's domain* if its decision is exclusively based on $H(\mathcal{B},c(\mathbf{x}))$. □

Hereafter, we will focus on the analysis of those detectors.

In view of the previous definitions, it is clear that a vector $\mathbf{x}$ is said to follow Benford's law if and only if

$$H(b_i,c(\mathbf{x})) = N\left[\log_{10}\left(1+\frac{1}{i}\right)\right], \quad (5)$$

where $n_1 = 9$, and $b_i = \log_{10}(i+1)$, $1 \leq i \leq 8$. From (5), and noticing that a sufficient condition for Benford's law verification is the uniform distribution of the coset representatives in $[0,1)$, it is possible to introduce a generalization referred to

as *Strong Benford's law* [9]. A vector $\mathbf{x}$ is said to follow this strong law if and only if

$$H(b_i, c(\mathbf{x})) = N\left[\log_{10}\left(b_i^+\right) - \log_{10}\left(b_{i-1}^+\right)\right], \tag{6}$$

where $b^+ = \min(\max(b, 1), 10)$.

Therefore, the attack optimization problem will be formalized for histogram-based detectors in Benford's domain by replacing (2)-(4) by

$$g^{H_c}(H(\mathcal{B}, \mathbf{y}), H(\mathcal{B}, \mathbf{y}')) \doteq \min_{\mathbf{y}'': H(\mathcal{B}, c(\mathbf{y}'')) = H(\mathcal{B}, c(\mathbf{y}'))} g^y(\mathbf{y}, \mathbf{y}''),$$

$$\mathbf{y}^* = \arg \min_{\mathbf{y}': H(\mathcal{B}, c(\mathbf{y}')) = H_c^\sharp(\mathcal{B}, c(\mathbf{y}^\sharp))} g^y(\mathbf{y}, \mathbf{y}'), \tag{7}$$

$$H_c^\sharp(\mathcal{B}, \mathbf{y}^\sharp) = \arg \min_{H'(\mathcal{B}, \mathbf{y}') \in \mathcal{R}_0^{H_c}} g^{H_c}(H(\mathcal{B}, \mathbf{y}), H'(\mathcal{B}, \mathbf{y}')), \tag{8}$$

where $\mathcal{R}_0^{H_c}$ denotes the acceptance region for histogram-based detectors in Benford's domain. Due to the modulo-reduction inherent to the coset representative definition, it is obvious that the feasible region in (7) is non-convex. This represents a substantial difference with respect to the framework considered in [7].

Since $\mathbf{x}$ is the realization of a random vector $\mathbf{X}$, some tolerance on (5) and (6) is typically allowed by Benford's law (and strong Benford) detectors, and specifically by histogram-based ones, meaning that $\mathcal{R}_0^{H_c}$ would be a region around the histogram $N \cdot [P(1), P(2), \ldots, P(9)]$. However, for the sake of computational simplicity, and in order to avoid the dependence of the proposed attack with a particular detector, we will require (5) (or (6)) to be exactly verified (up to rounding effects), yielding a single-element set $\mathcal{R}_0^{H_c}$. From a practical point of view this means that the optimization in (8) is not longer performed, as its feasible set contains a single point that, consequently, is the desired solution. A similar approach was also followed in [14], where the target histogram was the average histogram of legal images. In any case, the reader should be aware that, since the approach proposed in the next section deals with the optimization of (7), it can be still used for the case where $\mathcal{R}_0^{H_c}$ has more than one element, requiring in such case to additionally perform the optimization in (8).

III. Main result

Given a histogram in Benford's domain $H_c^\sharp(\mathcal{B}, c(\mathbf{y}^\sharp))$, we want to solve (7), or equivalently,

$$\mathbf{y}^* = \arg \min_{\mathbf{y}': \mathbf{y}' \sim \mathbf{y}^\sharp} g^y(\mathbf{y}, \mathbf{y}'). \tag{9}$$

We assume, as it was already mentioned, g^y to be component-wise additive and dependent on the difference between the input vectors. In order to solve this problem we consider the cost matrix M, whose elements are

$$m_{i,j} = \min_{a: c(a) \in [b_{i-1}, b_i)} g^{y_j}(y_j - a), \tag{10}$$

i.e., the cost required to move the jth component of $\mathbf{y}$ to the ith histogram bin in Benford's domain. Then, solving (11) is equivalent to solving

$$\min_{a_{i,j} \in \mathcal{C}} \sum_{i=1}^{n_1} \sum_{j=1}^{N} a_{i,j} m_{i,j},$$

where $\mathcal{C}$ is defined as

$$\mathcal{C} \doteq \Big\{ a_{i,j} : a_{i,j} \in \{0,1\}, 1 \le i \le n_1, 1 \le j \le N;$$
$$\sum_{i=1}^{n_1} a_{i,j} = 1, 1 \le j \le N; \tag{11}$$
$$\sum_{j=1}^{N} a_{i,j} = H_c^\sharp(b_i, c(\mathbf{y}^\sharp)), 1 \le i \le n_i \Big\}. \tag{12}$$

Indeed, this is a linear optimization problem that can be solved by using the *simplex algorithm* [20].

Note that, although the definition of simplex problem does not enable one to consider $a_{i,j} \in \{0,1\}, 1 \le i \le n_1, 1 \le j \le N$, but only use $a_{i,j} \ge 0$ and the two equality constraints (i.e., (11) and (12)), properties of the simplex guarantee that the minimum is achieved (at least) at a vertex of the resulting polytope. By considering (11) and (12), this implies $a_{i,j} \in \{0,1\}$. Depending on the structure of M, the minimum could be simultaneously achieved at several vertices; if this were the case, then the convex combination of those vertices (i.e., the corresponding edges and/or faces of the polytope) would also provide the minimal value. In any case, the simplex algorithm will return one of those optimal vertices, and therefore the constraint $a_{i,j} \in \{0,1\}, 1 \le i \le n_1, 1 \le j \le N$ will be implicitly verified.

A. Cost matrix calculation

The only remaining point is the calculation of the cost matrix M. In order to do that, notice that (10) is equivalent to

$$m_{i,j} = \min_{b \in [b_{i-1}, b_i)} \min_{n \in \mathbb{N}} g^{y_j}(y_j - 10^{b+n}),$$

$1 \le i \le n_1, 1 \le j \le N$. Assuming, as it is typically the case, that $g^{y_j}(y)$ is continuous, monotonically increasing for $y > 0$, and monotonically decreasing for $y < 0$, the previous expression can be written as

$$m_{i,j} = \begin{cases} g^{y_j}(0), & \text{if } c(y_j) \in [b_{i-1}, b_i) \\ \min\Big(g^{y_j}(y_j - \text{sign}(y_j) 10^{b_i + \lfloor \log_{10}(|y_j|) - b_i \rfloor}), \\ \quad g^{y_j}(y_j - \text{sign}(y_j) 10^{b_{i-1} + \lceil \log_{10}(|y_j|) - b_{i-1} \rceil})\Big), \\ \text{otherwise} \end{cases}. \tag{13}$$

If $g^{y_j}(y)$ were not continuous, then one should consider the possibility for the discontinuities to be at histogram bin boundaries. Additionally, if $g^{y_j}(y)$ were not monotonically increasing for $y > 0$ and decreasing for $y < 0$, then (13) should be replaced by a more involved optimization procedure (e.g., splitting the optimization domain in intervals where the target function is monotonic). In the remaining part of this paper we will assume that $g^{y_j}(y) = y^2$ (i.e., the distortion criterion is the MSE), and, consequently, (13) can be used.

IV. Experimental results

In order to show the goodness of the counterforensic method proposed in the previous section, the following experimental setup is used:

- We consider the 1338 images in the UCID database [21] (in grayscale).
- All of them are JPEG compressed with a Quality Factor (QF) ranging from 10 to 100, with stepsize 10.
- The 8×8 DCT of the compressed images is computed.
- For each image, we separately consider each of the 63 AC 8×8 DCT frequencies.
- For each of those frequencies the non-null coefficients are modified in order to comply with Benford's Law or Strong Benford's Law (in both cases up to rounding effects), depending on the particular experiment. Null coefficients are not modified.
- The target function to be minimized is the distortion Mean Square Error (MSE).
- A further quantization of the DCT coefficients is not considered.

A possible practical limitation of this framework is the disregard of both the modification of null components and the quantization of DCT coefficients, which do not affect either the methodology or the conclusions. In any event, we remark that the comparison that we carry out below is fair in the sense that both methods have exactly the same constraints. Therefore, the results are illustrative enough of the power of the simplex method. We also notice that our scheme could be also easily adapted to deal with those further constraints. Indeed, both issues are solved by modifying the computation of $m_{i,j}$.

Concerning previous works in the literature, we establish our comparison with [14], since, as there reported, such scheme outperforms all previous proposals in the literature (i.e., [22], [13]). Pasquini *et al.*'s method sorts the DCT coefficients in decreasing order of magnitude. Then, each coefficient is orderly assigned to that closest bin whose maximum number of coefficients (according to the target histogram) has not been yet achieved. The output value for each coefficient is the closest point of the assigned bin to the original value.

A. Benford's Law ($n_1 = 9$)

In this case the target histogram bin boundaries are $b_i = \log_{10}(i+1)$, and the target number of coefficients in each bin $H(b_i, c(x)) = N \log_{10}\left(1 + \frac{1}{i}\right)$, $1 \leq i \leq 9$.

The obtained results verify that for each image, DCT frequency, and JPEG compression QF, the MSE necessary for generating a Benford's Law-compliant DCT histogram by following the scheme proposed in the previous section is smaller than that achieved with the scheme in [14]. Fig. 1 shows the histogram modification distortion Peak Signal-to-Noise Ratio (PSNR), where we consider the average MSE (over the entire image database) required to generate a compliant histogram, for both our simplex-based method and the method in [14], and different QFs. As one would expect, the MSE is decreasing with the QF, that is, the larger the QF (i.e., the lighter the compression), the easier it will be to generate a compliant histogram.

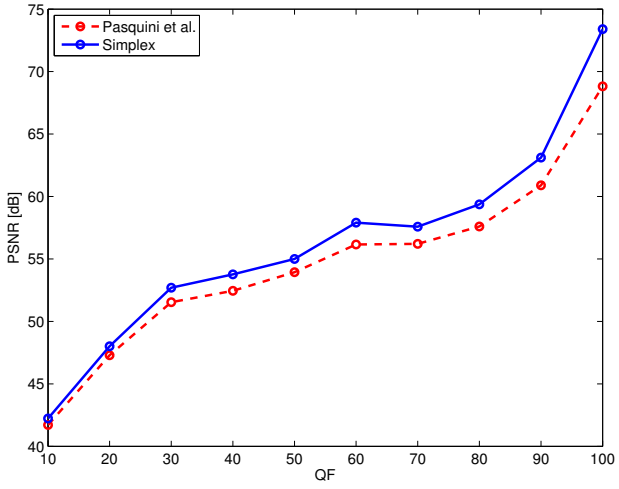

Fig. 1. Histogram modification distortion PSNR as a function of QF for Benford's Law.

On the other hand, Fig. 2 shows the MSE gain (in dB) achieved by our simplex-based strategy with respect to the method in [14], as a function of the DCT coefficient index when one averages the MSE of the database images, and the 10 values of the QF introduced above.

B. Strong Benford's Law ($n_1 = 20$)

In this case the target histogram bin boundaries are $b_i = \frac{i}{20}$, and the target number of coefficients in each bin $H(b_i, c(x)) = \frac{N}{20}$, $1 \leq i \leq 20$.

Similarly to Benford's Law, the plots in Fig. 3 show the PSNR as a function of the considered QF, while Fig. 2 illustrates the gain achieved by our proposal with respect to [14], and its dependence with the considered DCT frequency. The conclusions derived for Benford's Law in the previous section are also valid here: the higher the QF, the smaller the MSE. Furthermore, the MSE required for generating a compliant histogram with the simplex-based scheme proposed here is smaller than that required by the method in [14] for each image, each frequency, and each considered QF.

V. Conclusions

In this work we have revisited the problem of attacking histogram-based forensic tools. Specifically, we show that in a particular case of practical interest (that includes Benford's Law-based detectors), the modification distortion optimization problem can be reframed as a linear programming problem. Consequently, one can use well-known optimization tools, such as the simplex, to find the optimal solution. So far, the proposed solutions to this problem were heuristic and clearly suboptimal. This paper introduces for the first time the MSE optimal solution to this well-known problem.

Concerning future work, we will consider the use of different distortion measures, provided that the required constraints

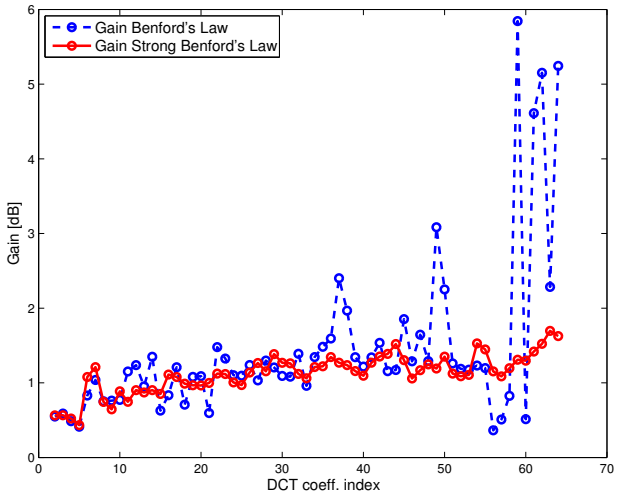

Fig. 2. MSE gain achieved by our proposal with respect to [14] as a function of DCT coefficient frequency for both Benford's Law and Strong Benford's Law (scanned in zigzag).

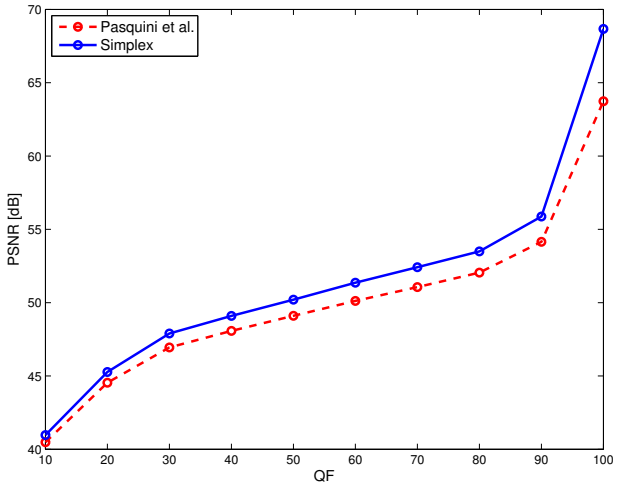

Fig. 3. Histogram modification distortion PSNR as a function of QF for Strong Benford's Law.

(i.e., component-wise additive, dependent on the difference between the input vectors, and monotonically increasing for positive values and monotonically decreasing for negative values) are verified.

ACKNOWLEDGMENTS

This work was partially funded by the EU 7th Framework Programme under project NIFTy (HOME/2012/ISEC/AG/INT/4000003892), the Spanish Government and the ERDF under project TACTICA, by the Spanish Government under project COMPASS (TEC2013-47020-C2-1-R), by the Galician Regional Government and the ERDF under projects "Consolidation of Research Units" (GRC2013/009), REdTEIC (R2014/037) and AtlantTIC.

REFERENCES

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 8682–8686.

[2] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, February 2005.

[3] M. C. Stamn, S. K. Tjoa, W. S. Lin, and K. J. R. Liu, "Anti-forensics of JPEG compression," in *Proc. IEEE ICASSP*, Dallas, TX, March 2010, pp. 1694–1697.

[4] G. Valenzise, M. Tagliasacchi, and S. Tubaro, "The cost of JPEG compression anti-forensics," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 1884–1887.

[5] W. Fan, K. Wang, F. Cayre, and Z. Xiong, "A variational approach to JPEG anti-forensics," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 3058–3062.

[6] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. of ACM Workshop on Multimedia and Security*, Coventry, UK, September 2012, pp. 97–104.

[7] P. Comesaña and F. Pérez-González, "Optimal counterforensics for histogram-based forensics," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 3048–3052.

[8] F. Balado, "The role of permutation coding in minimum-distortion perfect counterforensics," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 6281–6285.

[9] S. J. Miller, Ed., *The Theory and Applications of Benford's Law*. Princeton University Press, 2014.

[10] D. Fu, Y. Q. Shi, and W. Su, "A generalized Benford's law for JPEG coefficients and its applications in image forensics," in *Proc. SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents*, vol. 6505, San Jose, CA, January 2007, p. 1L.

[11] B. Li, Y. Q. Shi, and J. Huang, "Detecting double compressed JPEG images by using mode nased first digit features," in *Proc. IEEE MMSP*, Cairns, Australia, October 2008, pp. 730–735.

[12] S. Milani, M. Tagliasacchi, and S. Tubaro, "Discriminating multiple JPEG compression using first digit features," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 2253–2256.

[13] C. Pasquini and G. Boato, "JPEG compression anti-forensics based on first significant digit distribution," in *Proc. IEEE MMSP*, Pula, Italy, September-October 2013, pp. 500–505.

[14] C. Pasquini, P. Comesaña, F. Pérez-González, and G. Boato, "Transportation-theoretic image counterforensics to first significant digit histogram forensics," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 6529–6533.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[16] T. Pevny and J. Fridrich, "Detection of double-compression in JPEG images for applications in steganography," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, June 2008.

[17] P. Sutthiwan and Y. Q. Shi, "Anti-forensics of double JPEG compression detection," in *Lectures Notes in Computer Science. Proc. IWDW*, vol. 7128. Atlantic City, NJ: Springer, October 2011, pp. 411–424.

[18] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE ICIP*, vol. 2, Austin, TX, November 1994, pp. 982–986.

[19] A. D'Angelo, L. Zhaoping, and M. Barni, "A full-reference quality metric for geometrically distorted images," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 867–881, April 2010.

[20] G. B. Dantzig, *Linear Programming and Extensions*. Princeton University Press, 1998.

[21] G. Schaefer and M. Stich, ""UCID - An Uncompressed Colour Image Database"," in *Proc. of SPIE. Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, San Jose, CA, January 2004, pp. 472–480.

[22] S. Milani, M. Tagliasacchi, and S. Tubaro, "Antiforensics attack to Benford's law for the detection of double compressed images," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 2253–2256.