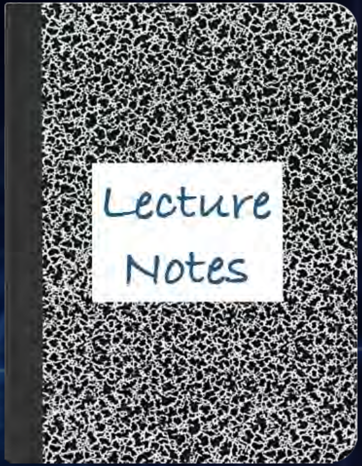
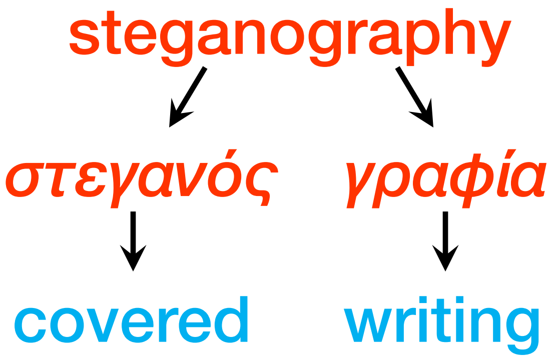


CS 419: Computer Security

Week 13: Hiding Communication Steganography

Paul Krzyzanowski



steganography

στεγανός γραφία

covered writing

The art of secret writing.

— *Oxford English Dictionary*

Steganography

Hiding the existence of a message

Undetectability: An observer should not realize that hidden data is present

Defining characteristics

1. **The cover object appears untouched**

   The modifications should be invisible and statistically difficult to detect

2. **The hidden data is the payload**

   You embed arbitrary messages to transmit secretly

3. **Adversary model: detection is the threat**

   If an analyst suspects the content contains hidden data, the mission fails

Watermarking

Embedding information into content

Persistence, not secrecy: The data should survive transformations

Defining characteristics

1. **The watermark does not need to be hidden**

   Some might be invisible, but secrecy is not required.

2. **The watermark is tied to the object, not a message payload**

   It could include ownership, license information, or tracking data

3. **Adversary model: removal is the threat**

   The watermark should survive cropping, compression, resampling, screenshotting, etc.

Watermarking vs. Steganography

Both techniques embed a message in data. Often used interchangeably

Goal of steganography: secrecy ⇒ conceal a message

- Intruder cannot detect there's a message in the content
- Primarily used for 1:1 communication

Examples: invisible ink, embedding secret text into an image

Goal of watermarking: identification ⇒ preserve a message

- Embed authorship or authenticity information into content
- Presence is intended: Doesn't have to be invisible
- The goal is to detect the watermark and preserve it
- Primarily used for 1:many communication

Examples: show that a photo belongs to Getty images, embed a logo in a TV show

Watermarking: Sample Applications

Encoding identifiable information (called a watermark), into content like images or video, to claim ownership or verify authenticity

Applications

- Copyright/creation affirmation
  - Embed information about owner
  - Label AI-created content
- Copy protection rules
  - Embed rights management information
  - But you need a trusted player
- Content authentication
  - Detect changes to the content

Fragile vs. Robust Watermarking

Fragile watermarks: designed to break if the content is modified

- **This is good for authentication and tamper detection**
- **Examples: currency, passports, entry tickets**
  - These will no longer be valid if tampered – users should not want to remove them

Robust watermarks: designed to survive transformation

- **This is good for tracking content authorship and ownership**
- **Examples: photos, videos, audio, documents**
  - Users may try to remove these

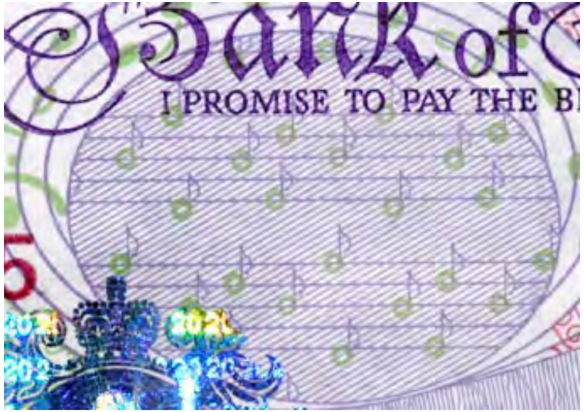
Classic techniques in watermarking

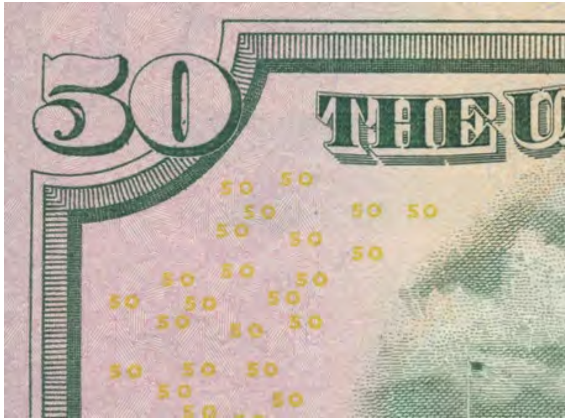
Change thickness of paper while wet via a pattern in the paper mold

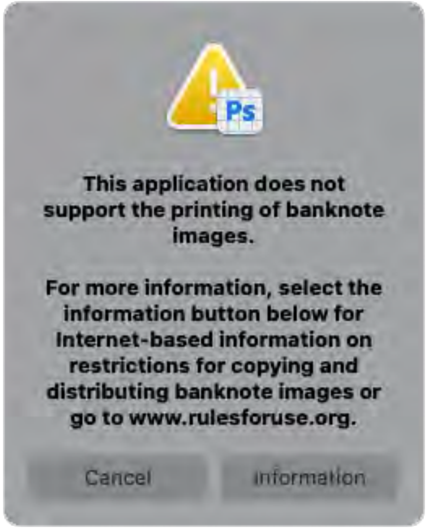
- First used in 1282: identify paper maker or trade guild that made the paper
- The dry paper could be rolled again to create even thickness but varying density
- Later used in banknotes to enable detection of authentic banknotes
  - First used in 1661 issue of the Stockholms Banco

Watermarking: EURion constellation (Omron rings)

- Series of five small images are repeated throughout the banknote
- Software recognizes the pattern to prevent scanning
- Used by the Armenia, Australia, Canada, China, EU, India, Japan, Mexico, Switzerland, Thailand, UK, U.S, … Zimbabwe






This application does not support the printing of banknote images.

For more information, select the information button below for Internet-based information on restrictions for copying and distributing banknote images or go to www.rulesforuse.org.

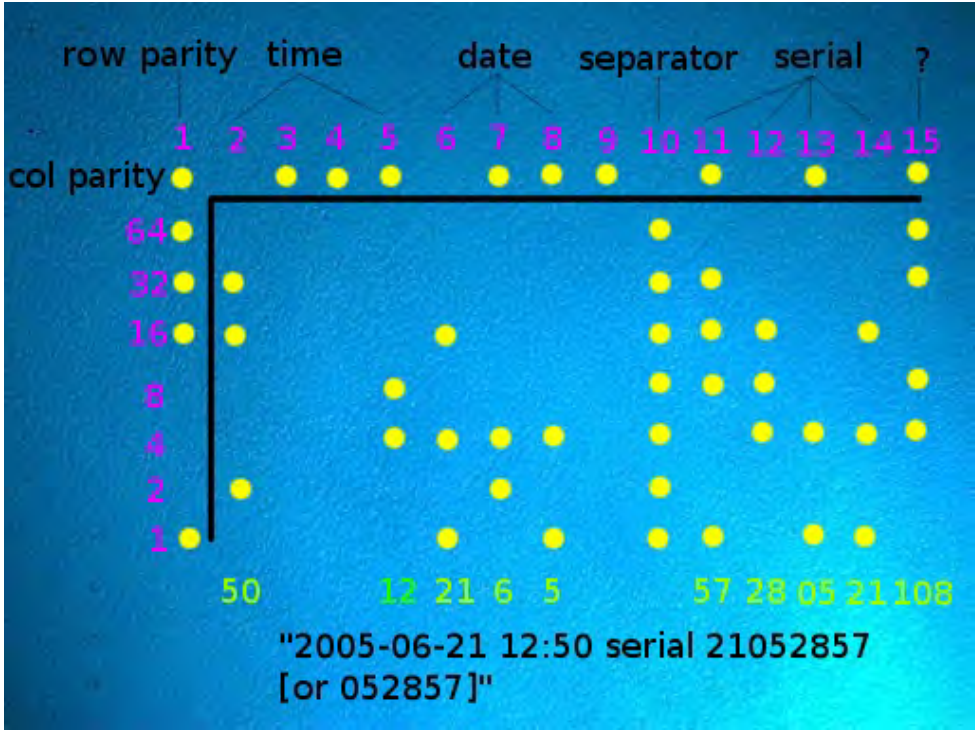
Machine ID codes in laser printers

Watermark added by the manufacturer to identify the printer used to create printed content

See http://www.eff.org/Privacy/printers/

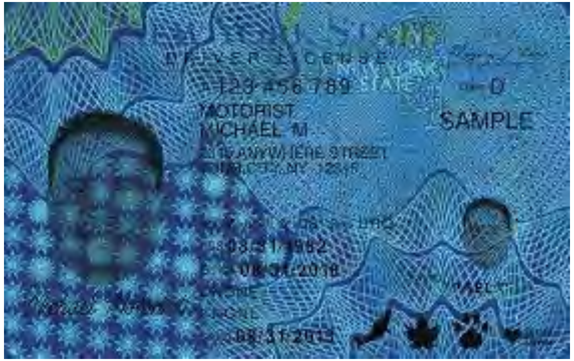
Machine ID codes in laser printers

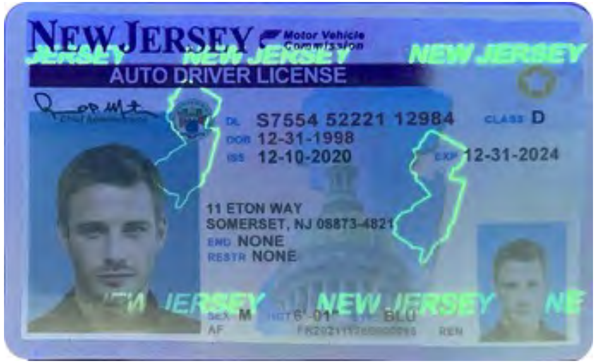
Machine ID codes in laser printers

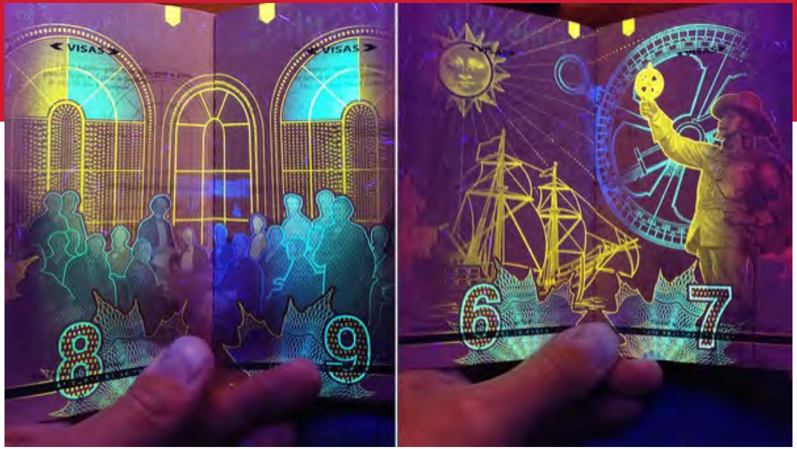


Designed by Xerox to identify counterfeit currency and help track down counterfeiters

UV Watermarking







Passports (Canada↑, Hungary↓)

Also, currency, hand stamps for amusement park/club re-entry

Fragile vs. Robust Watermarking

Fragile watermarks: designed to break if the content is modified

- **This is good for authentication and tamper detection**
- **Examples: currency, passports, entry tickets**
  - These should no longer be valid if tampered

Robust watermarks: designed to survive transformation

- **This is good for tracking content authorship and ownership**
- **Examples: photos, videos, audio, documents**

Classic techniques in steganography

- **Invisible ink (1st century AD - WW II)**
  - Used from antiquity through World War II. Lemon juice, vinegar, milk, cobalt chloride, and sympathetic inks.
- **Messages hidden on the body**
  - Herodotus describes shaving a slave’s head, writing a message on the scalp, and waiting for the hair to grow back.
- **Wax covered tablets – ancient Greece**
  - Herodotus described messages carved into wood underneath a wax writing surface
- **Altered printed characters**
  - Letters slightly overwritten with pencil or ink to signal a code.
  - Similar variants include smudges, underlines, or bent type from manual presses.

https://www.giac.org/paper/gsec/3494/steganography-age-terrorism/102620

Classic techniques in steganography

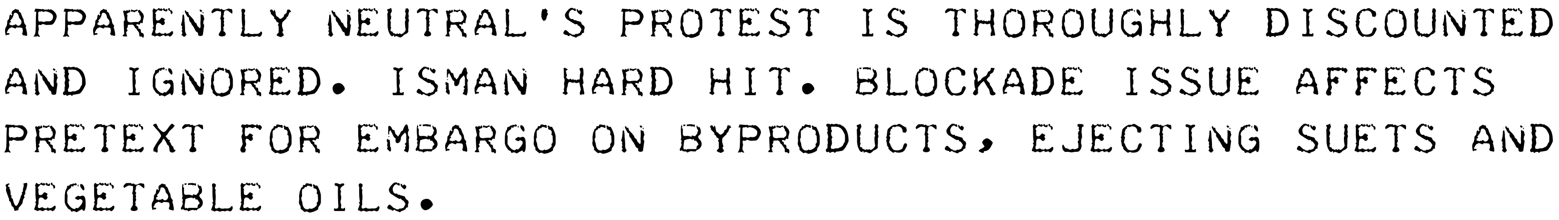
- **Pin-prick codes in printed text**
  - Micro-perforations over selected letters in typewritten text, books, or newspapers
- **Microdots (early 20th century) – first used in 1941**
  - Photographically reduced images the size of a typographic period; widely used in WWII espionage.
- **Newspaper clippings, knitting instructions, XOXO signatures, report cards, …**
  - Charles Dickens in *A Tale of Two Cities,* describes a fictitious account how knitting patterns were used during the French Revolution to pass information on who was going to die
  - WW II war censors eventually prohibited or tampered flower deliveries, radio song requests, weather reports, children's drawings sent in the mail, …

Null Cipher (concealment cipher)

- **Hide message in a large amount of irrelevant data**
- **Agreed technique for extracting content**
  - First letter of each word, Nth letter of each word
  - Some specific pattern to define which words or letters are significant (e.g., 4-5-5-4 words)

Null Cipher (concealment cipher)

Sent by a German spy in WWI:



APPARENTLY NEUTRAL'S PROTEST IS THOROUGHLY DISCOUNTED AND IGNORED. ISMAN HARD HIT. BLOCKADE ISSUE AFFECTS PRETEXT FOR EMBARGO ON BYPRODUCTS, EJECTING SUETS AND VEGETABLE OILS.

Reference: David Kahn, *The Codebreakers*, p. 521

Null Cipher (concealment cipher)

The 2nd letter of each word contains the message

APPARENTLY NEUTRAL'S PROTEST IS THOROUGHLY DISCOUNTED
AND IGNORED. ISMAN HARD HIT. BLOCKADE ISSUE AFFECTS
PRETEXT FOR EMBARGO ON BYPRODUCTS, EJECTING SUETS AND
VEGETABLE OILS.

PERSHING SAILS FROM NY JUNE I

By WWII, not used by spies but by regular people trying to beat the censor.

Reference: David Kahn, *The Codebreakers*, p. 521

BBC News 27 April 2006

Judge creates own Da Vinci code

The judge who presided over the failed Da Vinci Code plagiarism case at London's High Court hid his own secret code in his written judgement.

Seemingly random italicised letters were included in the 71-page judgement given by Mr Justice Peter Smith, which apparently spell out a message.

Mr Justice Smith said he would confirm the code if someone broke it.

"I can't discuss the judgement, but I don't see why a judgement should not be a matter of fun," he said.

Italicised letters in the first few pages spell out **"Smithy Code"**, while the following pages also contain marked out letters.

http://news.bbc.co.uk/go/pr/fr/-/1/hi/entertainment/4949488.stm

Motivation

Steganography received little attention in computing until recently

- **Industry's desire to protect copyrighted digital work**
  - Detect counterfeit, unauthorized presentation, embed key, embed author ID
  - This is mostly *watermarking* (more on that later)
- **Covert way to distribute malware – bypass detection**
  - E.g., embed in a JPEG file, which would raise no suspicion when downloaded
- **Covert way to exfiltrate data**
  - Upload harmless-looking content that contains embedded data
  - **Network steganography**: embed the secret data within unused data fields, headers (example, communicate with C2 servers via TXT fields in DNS messages

Steganography ≠ Copy protection ≠ Cryptography

Malware delivery & exfiltration via steganography

Trick content-inspecting firewalls by disguising malware and/or data

- **Data Exfiltration**
  - Russian hackers hid malware in a trojanized update from SolarWinds, a popular IT management platform – this included a backdoor
    - Successfully breached Cisco, Intel, Microsoft, and U.S. government agencies
    - Used network steganography to hide command-and-control communications inside ordinary-looking DNS traffic
    - Later, exfiltrated stolen information as XML files via HTTPS to disguise content

Malware delivery & exfiltration via steganography

Malware Infiltration

- 2019: A nation-state actor also used steganography to hide Windows DLLs (dynamic linked libraries) inside of WAV files to install a cryptomining app
- 2020: Attackers embedded skimming malware in SVG graphics in an attack on Dutch eCommerce platform Sansec – JavaScript would parse it out.
- 2020-2023: Turla APT: JPEG steganography for C2 instructions
  - Used PNG and JPEG images posted on public websites to conceal secondary-stage payloads & command-and-control information
- 2023: PolyglotDuke: C2 Instructions embedded in PNG social media images
  - Pro-Russian group distributed images on X and other platforms that contained hidden instructions identifying targets for DDoS campaigns.

Steganography has now become common in low-end commodity malware, not just nation-state operations

Code hidden in photo, files stolen: Upstate man stole GE technology to try to help China

syracuse.com

Anne Hayes • April 1, 2022

Schenectady, N.Y – A Schenectady County man who hid data in the code of a digital photograph of a sunset was convicted Thursday of conspiracy to commit economic espionage against General Electric in order to benefit the Chinese government.

Xiaoqing Zheng, 59, was originally accused of stealing GE trade secrets regarding turbine technology and planning to give the information to contacts in China, according to federal court documents.

Although the jury convicted Zheng, a U.S. citizen, of conspiracy to commit economic espionage, they could not reach a unanimous decision regarding the charge of economic espionage, according to a news release from the U.S. Attorney's Office of the Northern District of New York.

…

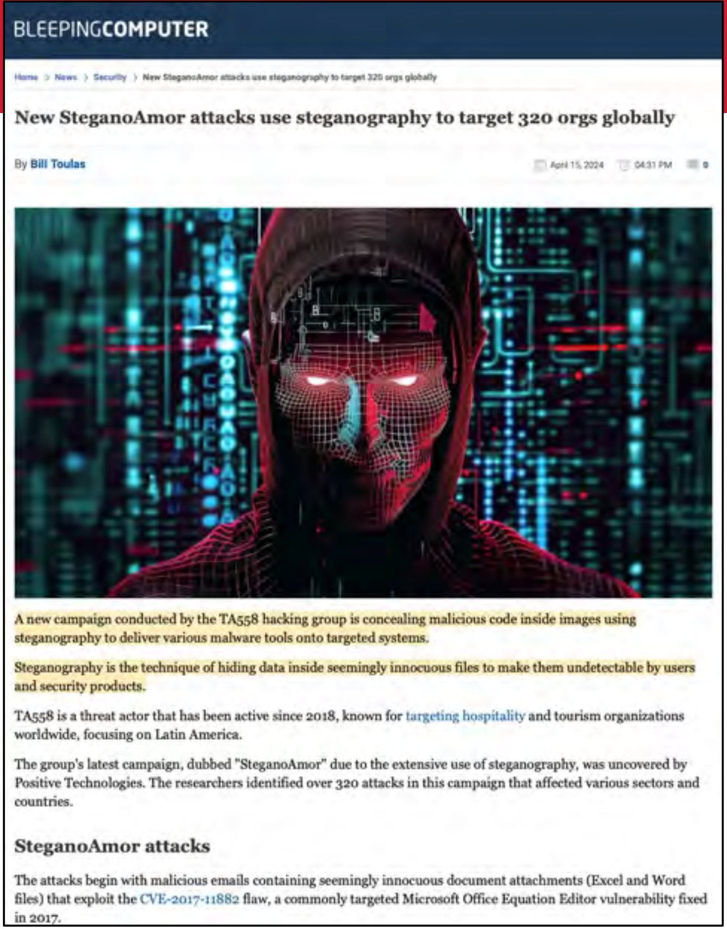
In 2018, Zheng used a means of hiding data within the code of another file to conceal 40 files in the code of a digital photograph of a sunset. He then emailed the photograph file to his personal email account, according to court documents.

URhttps://www.syracuse.com/crime/2022/04/code-hidden-in-photo-files-stolen-upstate-man-stole-ge-technology-to-try-to-help-china.htmlL

SteganoAmor

2024 attack campaign

- **Attacker sends emails with "harmless" content exploiting a vulnerability in the Microsoft Office Equation Editor**
  - Memory corruption vulnerability: stack buffer overflow overwrites the return address to execute the attacker's code
  - The victim has to open the file to activate this
- **When the attachment is opened, it sends a request to a URL to download an RTF document. When opened, it runs a VBS script that fetches payload embedded in images in URLs**



BLEEPINGCOMPUTER

Home › News › Security › New SteganoAmor attacks use steganography to target 320 orgs globally

New SteganoAmor attacks use steganography to target 320 orgs globally

By Bill Toulas April 15, 2024 04:31 PM 0

A new campaign conducted by the TA558 hacking group is concealing malicious code inside images using steganography to deliver various malware tools onto targeted systems.

Steganography is the technique of hiding data inside seemingly innocuous files to make them undetectable by users and security products.

TA558 is a threat actor that has been active since 2018, known for targeting hospitality and tourism organizations worldwide, focusing on Latin America.

The group's latest campaign, dubbed "SteganoAmor" due to the extensive use of steganography, was uncovered by Positive Technologies. The researchers identified over 320 attacks in this campaign that affected various sectors and countries.

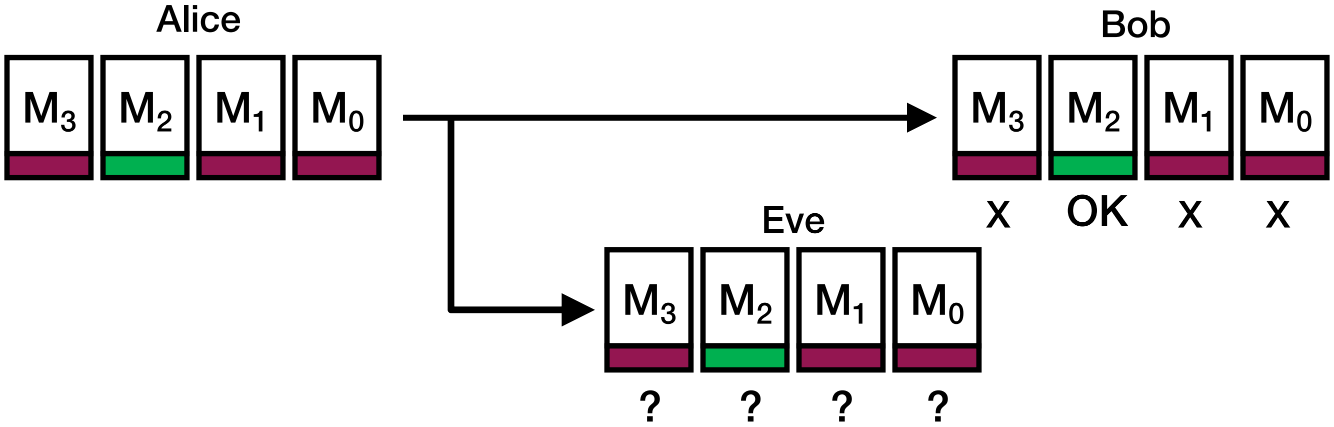
SteganoAmor attacks

The attacks begin with malicious emails containing seemingly innocuous document attachments (Excel and Word files) that exploit the CVE-2017-11882 flaw, a commonly targeted Microsoft Office Equation Editor vulnerability fixed in 2017.

https://global.ptsecurity.com/analytics/pt-esc-threat-intelligence/steganoamor-campaign-ta558-mass-attacking-companies-and-public-institutions-all-around-the-world

Chaffing & Winnowing

- **Separate good messages from the bad ones**
  - Easy for someone who has the key, difficult for someone who does not
- **Stream of un-encoded messages with signatures or MACs**
  - Some signatures are bogus
  - Need to have the key to test



Steganography in images

Spatial domain: LSB steganography

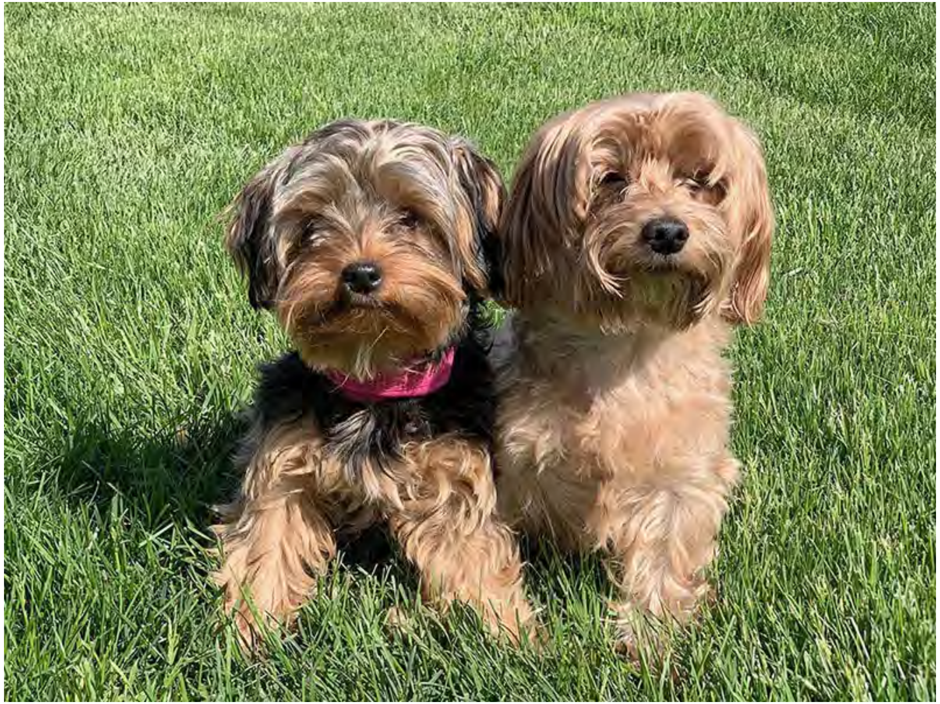
- Replace low-order bits of selected pixels with the message
- Option: use a pseudo-random schedule to identify which pixels to use

Frequency domain – work on transformed coefficients, not raw pixels

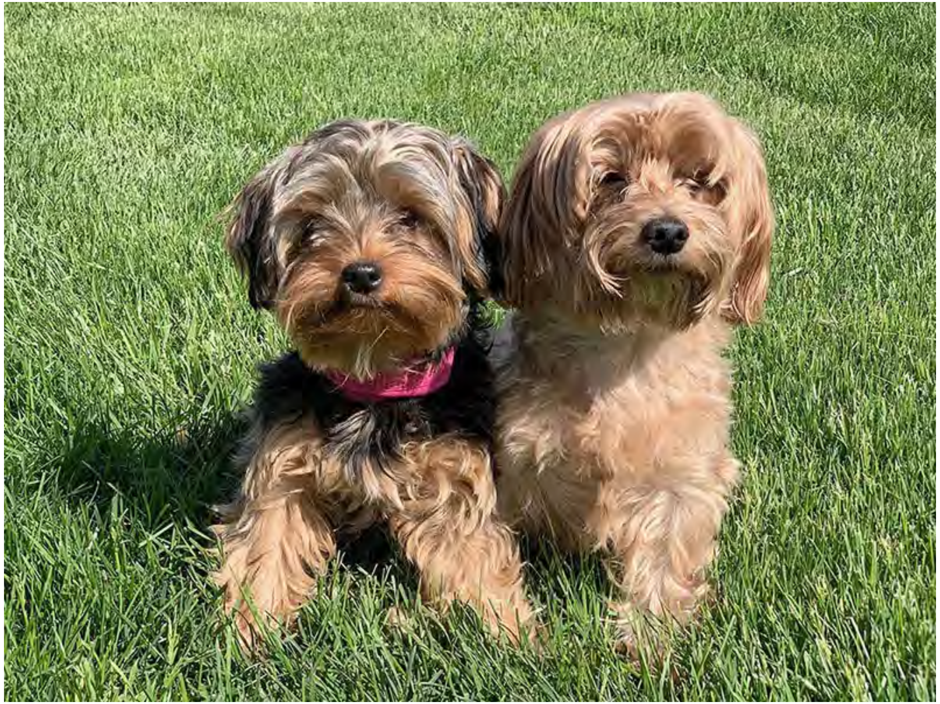
- Like JPEG compression: embed signal in mid- to high- frequency bands
- Alter the least perceptible parts of the image to avoid detection
  - But watch out: these are the same bits targeted by lossy image compression software (such as jpeg)

Metadata

- Add information the end of a PNG image's metadata or EXIF header
- This doesn’t count as steganography because it’s not really hidden but casual observers won’t see it



Just the picture



With the U.S. Declaration of Independence embedded

Differences

There are differences – but you don’t notice them in the photo

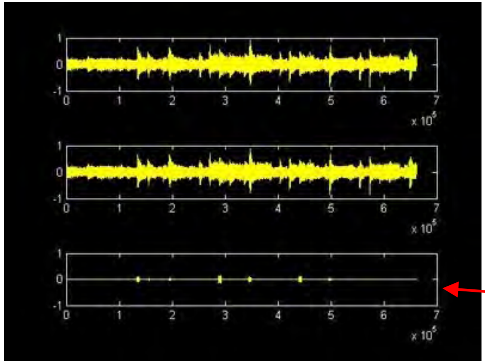
The choice of *cover medium* (the content the casual user sees) is crucial:

Media with high noise (photos, music) works better because small changes are less noticable.

Audio

Perceptual coding

- Inject the signal into areas that will not be detected by humans
- LSB steganography can also be used
- May be obliterated by compression



Amazon MP3 audio

Identifies where the song was purchased, not the user

Difference

Echomodulation Audio Watermarking

Adds low-amplitude echoes into the audio.
Too subtle for humans to perceive, but a decoder can detect them
Goal is robustness.

Spread Spectrum Audio Watermarking

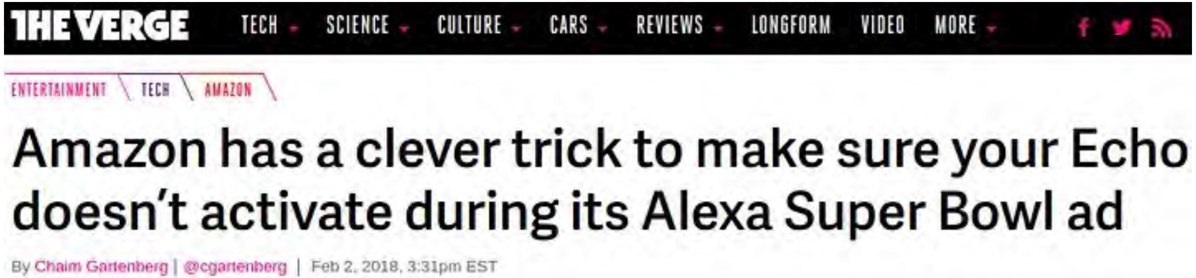
Used by Universal Music Group on Spotify audio – but largely dying out now. Still present on some music.

Intrasonics

Intrasonics (now Ipsos) – embeds watermarks into ads so clients can track source & medium used to transmit ad using echomodulation.

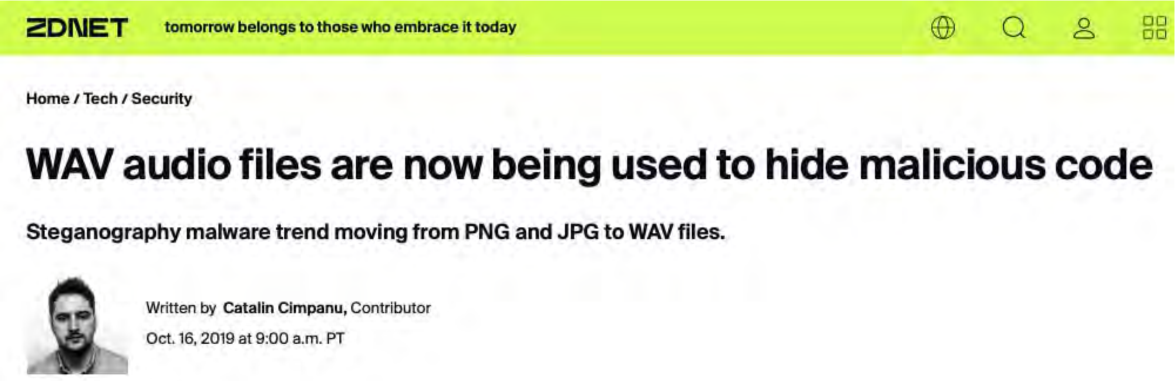
Audio

Amazon used inaudible signaling to prevent Echo devices from activating during its ads



Amazon has a clever trick to make sure your Echo doesn't activate during its Alexa Super Bowl ad

By Chaim Gartenberg | @cgartenberg | Feb 2, 2018, 3:31pm EST



Home / Tech / Security

WAV audio files are now being used to hide malicious code

Steganography malware trend moving from PNG and JPG to WAV files.

Written by Catalin Cimpanu, Contributor

Oct. 16, 2019 at 9:00 a.m. PT

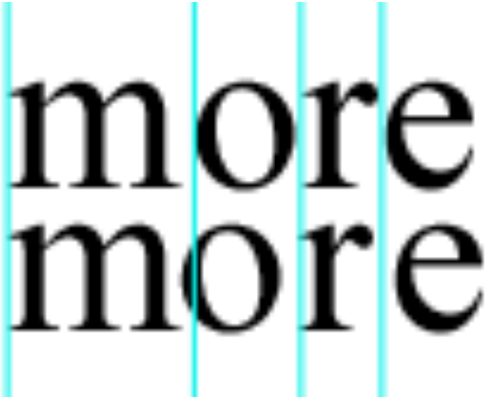
Video steganography & watermarking

Video = images + soundtrack

- **Coding still frames - spatial or frequency**
- **Modify motion vectors**
- **Caption/subtitle data**
- **Audio channel**
- **Visible watermarking**
  - used by most networks (logo at bottom-right)
    This isn't steganography!

Text

- Text lines shifted up/down
  (40 lines text ⇒ 2^{40} codes)
- Word space coding
- Character encoding — minor changes to shapes of characters



Works only on "images" of text e.g., PDF, postscript

Text-based steganography

> “Apparently, during the 1980’s, British Prime Minister Margaret Thatcher became so irritated at press leaks of cabinet documents that she had the word processors programmed to encode their identity in the word spacing of documents, so that disloyal ministers could be traced.”

– *Ross Anderson*
Stretching the Limits of Steganography

Text – non-visual

- **Embed zero-width non-printing characters**
  - Zero-width space or zero-width non-joiner (used to prevent ligature use)
- **White text on white background**
- **Overlapping objects**
- **PDF hidden pages**
- **HTML – invisible text designed to be picked up by search engines**
  - Non-rendered text (CSS element to not display text)
  - White on white
  - Obscured by other objects

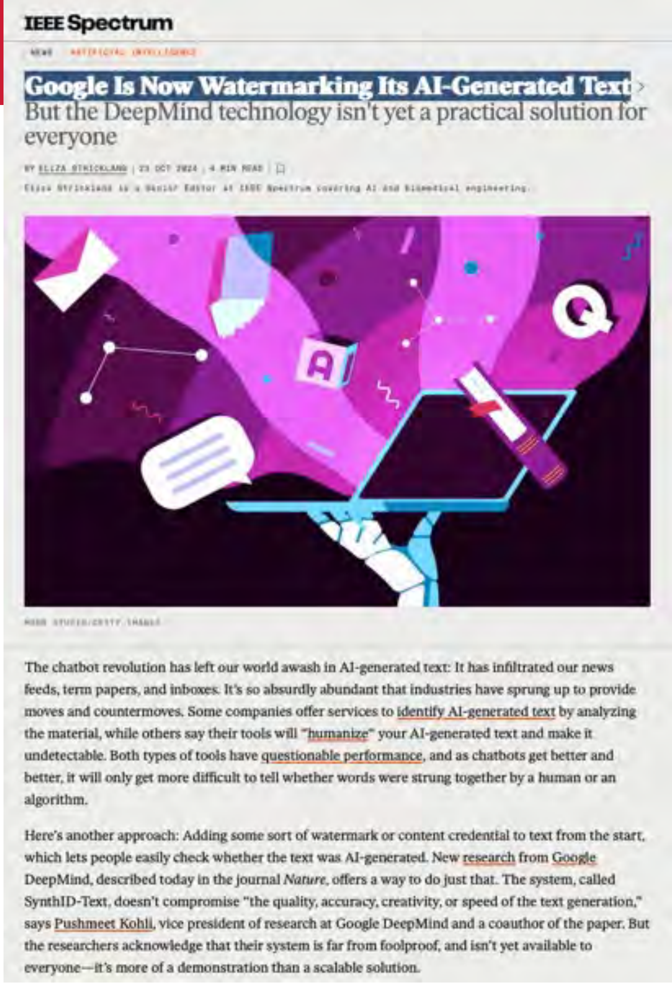
Watermarking AI-Generated Text

Google SynthID: images & text

Goal: Identify LLM-generated content, realizing that text can be easily altered to remove any watermarks

- **Images: makes tiny, structured modifications to an image's internal feature representation during generation.**
- **Text: Alters some words that a chatbot outputs to "introduce a statistical signature into the generated text"**
  - SynthID-Text randomly assigns scores to candidate words that may be generated by the LLM and has the LLM output words with higher scores
  - A detector can then calculate the over
- **Acknowledged to be "far from foolproof"**
  - Users can make significant edits or ask another chatbot to summarize the text

https://spectrum.ieee.org/watermark



IEEE Spectrum

NEWS ARTIFICIAL INTELLIGENCE

Google Is Now Watermarking Its AI-Generated Text

But the DeepMind technology isn't yet a practical solution for everyone

BY ELIZA STRICKLAND | 23 OCT 2024 | 4 MIN READ

Eliza Strickland is a Senior Editor at IEEE Spectrum covering AI and biomedical engineering.

MOOR STUDIO/GETTY IMAGES

The chatbot revolution has left our world awash in AI-generated text: It has infiltrated our news feeds, term papers, and inboxes. It's so absurdly abundant that industries have sprung up to provide moves and countermoves. Some companies offer services to identify AI-generated text by analyzing the material, while others say their tools will "humanize" your AI-generated text and make it undetectable. Both types of tools have questionable performance, and as chatbots get better and better, it will only get more difficult to tell whether words were strung together by a human or an algorithm.

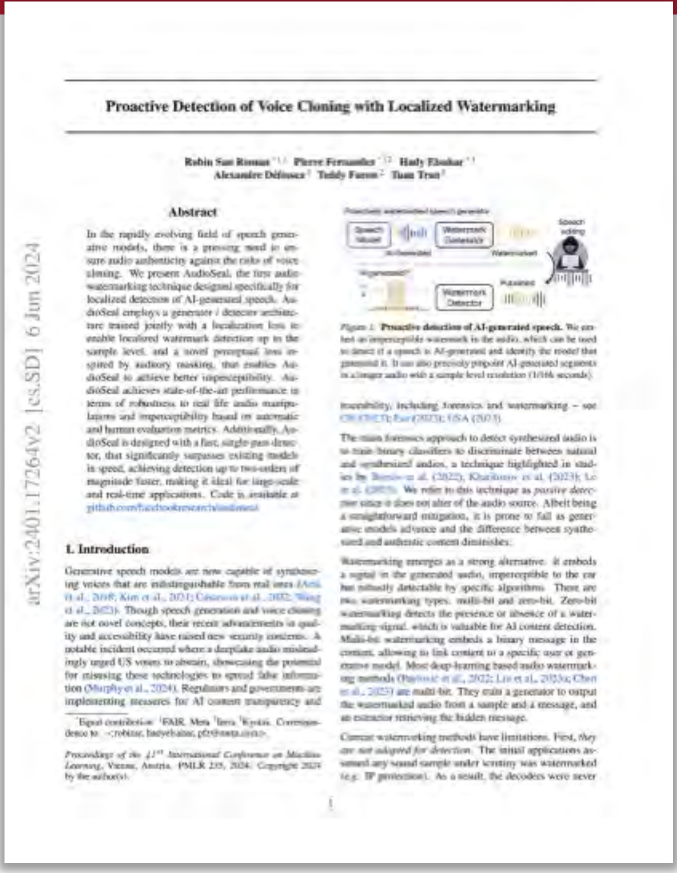
Here's another approach: Adding some sort of watermark or content credential to text from the start, which lets people easily check whether the text was AI-generated. New research from Google DeepMind, described today in the journal *Nature*, offers a way to do just that. The system, called SynthID-Text, doesn't compromise "the quality, accuracy, creativity, or speed of the text generation," says Pushmeet Kohli, vice president of research at Google DeepMind and a coauthor of the paper. But the researchers acknowledge that their system is far from foolproof, and isn't yet available to everyone—it's more of a demonstration than a scalable solution.

Watermarking AI-Generated Voice

Meta AudioSeal

Goal: detect AI-generated speech to mitigate risks like voice cloning & misinformation

- **Localized detection**
  - Embeds watermark into the audio in a way that it could be detected from even small sampled segments
- **Imperceptible**
  - Balances watermark strength with psychoacoustic principles
- **Robust**
  - Detectable after common manipulations such as bandpass filtering, speed changes, compression (e.g., MP3 encoding)
  - Uses psychoacoustic models to identify regions of audio that are both perceptually significant and resilient to compression
- **Efficient**
  - Single-pass detection with no need for synchronization



arXiv:2401.17264v2 [cs.SD] 6 Jun 2024

Proactive Detection of Voice Cloning with Localized Watermarking

Robin San Roman [*1,3] **Pierre Fernandez** [*1,2] **Hady Elsahar** [*1]
Alexandre Défossez [3] **Teddy Furon** [2] **Tuan Tran** [1]

Abstract

In the rapidly evolving field of speech generative models, there is a pressing need to ensure audio authenticity against the risks of voice cloning. We present AudioSeal, the first audio watermarking technique designed specifically for localized detection of AI-generated speech. AudioSeal employs a generator / detector architecture trained jointly with a localization loss to enable localized watermark detection up to the sample level, and a novel perceptual loss inspired by auditory masking, that enables AudioSeal to achieve better imperceptibility. AudioSeal achieves state-of-the-art performance in terms of robustness to real life audio manipulations and imperceptibility based on automatic and human evaluation metrics. Additionally, AudioSeal is designed with a fast, single-pass detector, that significantly surpasses existing models in speed, achieving detection up to two-orders of magnitude faster, making it ideal for large-scale and real-time applications. Code is available at github.com/facebookresearch/audioseal.

1. Introduction

Generative speech models are now capable of synthesizing voices that are indistinguishable from real ones (Arik et al., 2018; Kim et al., 2021; Casanova et al., 2022; Wang et al., 2023). Though speech generation and voice cloning are not novel concepts, their recent advancements in quality and accessibility have raised new security concerns. A notable incident occurred where a deepfake audio misleadingly urged US voters to abstain, showcasing the potential for misusing these technologies to spread false information (Murphy et al., 2024). Regulators and governments are implementing measures for AI content transparency and traceability, including forensics and watermarking – see Chi (2023); Eur (2023); USA (2023).

The main forensics approach to detect synthesized audio is to train binary classifiers to discriminate between natural and synthesized audios, a technique highlighted in studies by Borsos et al. (2023); Kharitonov et al. (2023); Le et al. (2023). We refer to this technique as *passive detection* since it does not alter of the audio source. Albeit being a straightforward mitigation, it is prone to fail as generative models advance and the difference between synthesized and authentic content diminishes.

Watermarking emerges as a strong alternative. It embeds a signal in the generated audio, imperceptible to the ear but robustly detectable by specific algorithms. There are two watermarking types: multi-bit and zero-bit. Zero-bit watermarking detects the presence or absence of a watermarking signal, which is valuable for AI content detection. Multi-bit watermarking embeds a binary message in the content, allowing to link content to a specific user or generative model. Most deep-learning based audio watermarking methods (Pavlović et al., 2022; Liu et al., 2023a; Chen et al., 2023) are multi-bit. They train a generator to output the watermarked audio from a sample and a message, and an extractor retrieving the hidden message.

Current watermarking methods have limitations. First, *they are not adapted for detection*. The initial applications assumed any sound sample under scrutiny was watermarked (e.g. IP protection). As a result, the decoders were never

Figure 1. **Proactive detection of AI-generated speech.** We embed an imperceptible watermark in the audio, which can be used to detect if a speech is AI-generated and identify the model that generated it. It can also precisely pinpoint AI-generated segments in a longer audio with a sample level resolution (1/16k seconds).

[*] Equal contribution [1] FAIR, Meta [2] Inria [3] Kyutai. Correspondence to: <robinsr, hadyelsahar, pfz@meta.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1

https://arxiv.org/abs/2401.17264

C2PA Standard (c2pa.org)

- **Coalition for Content Provenance and Authenticity**
  - Alliance between Adobe, Arm, Intel, Microsoft and Truepic
  - Standards for certifying the source and history of media content
- **Requires a C2PA-enabled capture device**
  - Content is hashed & signed to create a tamper-evident Content Credentials record
  - Can include attribution information
  - Any changes (crops, additions, removals, AI mods) are recorded

Leica M11-D
Leica M11-P

Canon EOS R1

Canon EOS
R5 Mark II

Fujifilm X-T50

Nikon Z6 III

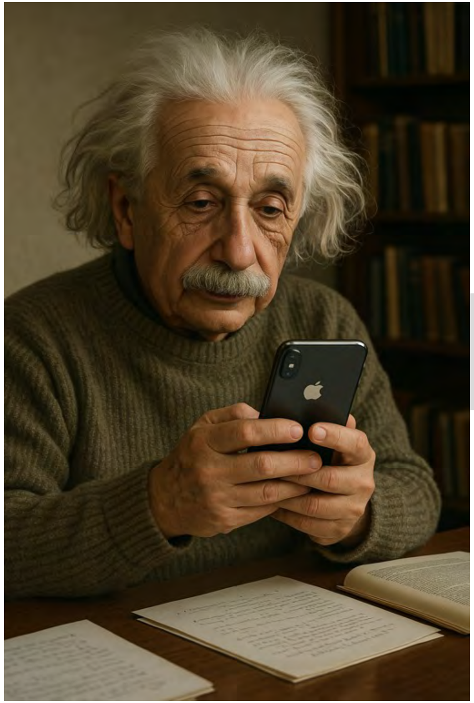
Fujifilm GFX100S II

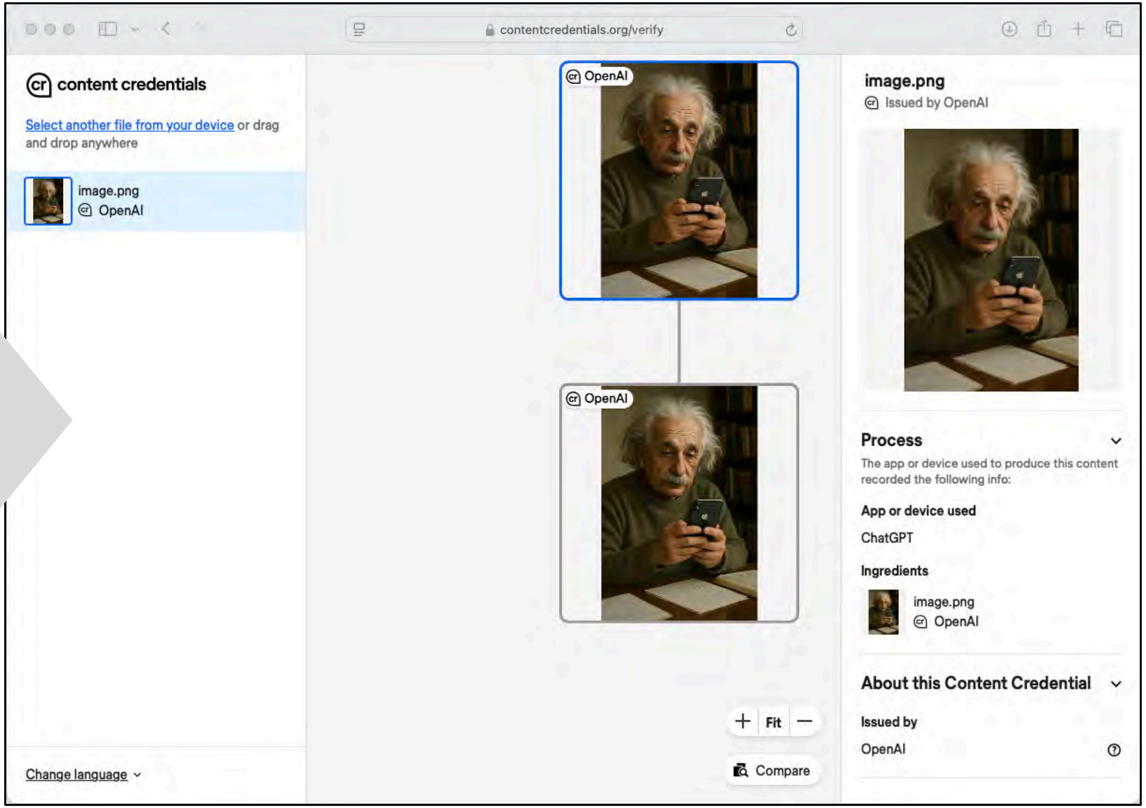
Sony a1 II
Sony a7s III
Sony a7 IV
Sony a9 III

Google Pixel 10

ChatGPT
Bing

https://contentcredentials.org/verify





content credentials

Select another file from your device or drag and drop anywhere

image.png
OpenAI

Change language

image.png
Issued by OpenAI

Process

The app or device used to produce this content recorded the following info:

App or device used
ChatGPT

Ingredients
image.png
OpenAI

About this Content Credential

Issued by
OpenAI

Statistical Steganalysis

Can we identify if content contains steganography?

Various attacks:

- **Histogram** (for LSB methods)
  - In natural images, histograms usually show smooth, continuous curves
  - Steganography (especially LSB embedding) can create unnatural patterns
- **Chi-square test**
  - LSB bits are normally not perfectly random
  - When secret data is embedded into LSBs (especially if it's random-looking encrypted data), it tends to randomize the LSB distribution
  - The chi-square test measures how far the observed distribution is from the expected one
- **Machine learning**: learn patterns of clean images to distinguish steganography

Payload capacity trade-offs

- The more data you hide, the greater the risk of introducing detectable artifacts

New Steganography Breakthrough Enables “Perfectly Secure” Digital Communications

SciTechDaily

University of Oxford • March 7, 2023

A group of researchers has achieved a breakthrough in secure communications by developing an algorithm that conceals sensitive information so effectively that it is impossible to detect that anything has been hidden.

The team, led by the University of Oxford in close collaboration with Carnegie Mellon University, envisages that this method may soon be used widely in digital human communications, including social media and private messaging. In particular, the ability to send perfectly secure information may empower vulnerable groups, such as dissidents, investigative journalists, and humanitarian aid workers.

…

Despite having been studied for more than 25 years, existing steganography approaches generally have imperfect security, meaning that individuals who use these methods risk being detected. This is because previous steganography algorithms would subtly change the distribution of the innocuous content.

To overcome this, the research team used recent breakthroughs in information theory, specifically minimum entropy coupling, which allows one to join two distributions of data together such that their mutual information is maximized, but the individual distributions are preserved.

https://scitechdaily.com/new-steganography-breakthrough-enables-perfectly-secure-digital-communications/

The End