Research Paper #3 - Stegomalware A Systematic Survey of Malware Hiding and Detection in Images, Machine Learning Models and Research Challenges

Learning Objectives:

- Review knowledge of steganography
- Create a summary of notes for reference and understanding
- Gain an understanding of stegomalware, including its creation methods, operational mechanisms, and look at notable examples
- Analyse existing Machine Learning approaches to detecting Steganography
- Identify existing challenges and research gaps between stegomalware and steganography detection

Introduction

> *"For example, an adversary may inject the malicious payload into the cover medium to evade the antimalware solutions detection. The malware hiding in the cover medium is termed as 'stegomalware'"*

Definition: Stegomalware refers to malware that employs steganography to hide malicious code within multimedia files, such as images, to evade detection by security systems.

- Lokibot - malicious source code concealed in a PNG image file for malware installation; avoids detection from email security tools

- Multimedia formats such as image, audio, video, and text are known vectors for malware concealment. Network protocols like TCP, UDP, and ICMP data then be used to hide victim data and be sent over via network communication

More Stegomalware examples (Generated content by ChatGPT):

Turla (aka Snake or Uroburos)

- **How it used steganography:** This Russian APT group used JPEG images uploaded to social media (like Twitter and Instagram) to embed **commands** for infected machines.
- **Purpose:** To establish covert command and control (C2) channels.
- **Notable feature:** The malware would download an image and extract commands hidden inside using steganographic techniques.

Stegoloader (aka Win32/Gatak)

- **How it used steganography:** It embedded malicious code in image files (usually PNGs), which were downloaded by the initial malware loader.
- **Purpose:** To load additional payloads while bypassing traditional detection systems.
- **Discovered:** Around 2015, primarily targeting the healthcare sector.

Historical Malware Variants employing Steganography for exploitation

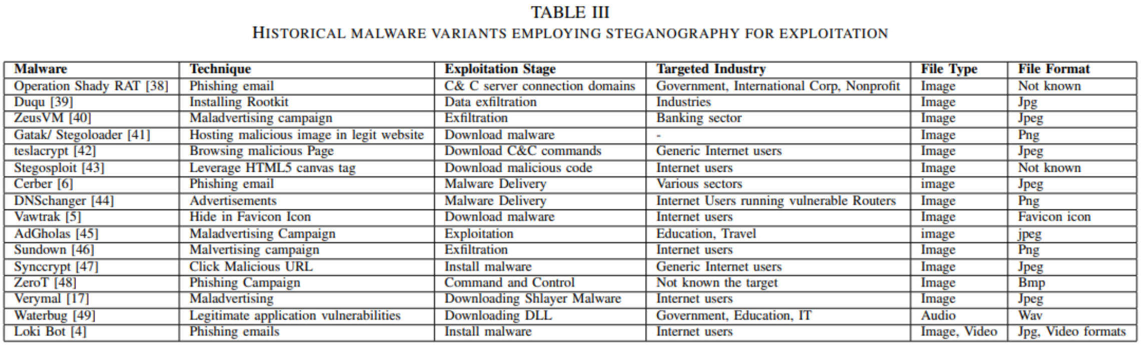


TABLE III
HISTORICAL MALWARE VARIANTS EMPLOYING STEGANOGRAPHY FOR EXPLOITATION

Malware	Technique	Exploitation Stage	Targeted Industry	File Type	File Format
Operation Shady RAT [38]	Phishing email	C& C server connection domains	Government, International Corp, Nonprofit	Image	Not known
Duqu [39]	Installing Rootkit	Data exfiltration	Industries	Image	Jpg
ZeusVM [40]	Maladvertising campaign	Exfiltration	Banking sector	Image	Jpeg
Gatak/ Stegoloader [41]	Hosting malicious image in legit website	Download malware	-	Image	Png
teslacrypt [42]	Browsing malicious Page	Download C&C commands	Generic Internet users	Image	Jpeg
Stegosploit [43]	Leverage HTML5 canvas tag	Download malicious code	Internet users	Image	Not known
Cerber [6]	Phishing email	Malware Delivery	Various sectors	image	Jpeg
DNSchanger [44]	Advertisements	Malware Delivery	Internet Users running vulnerable Routers	Image	Png
Vawtrak [5]	Hide in Favicon Icon	Download malware	Internet users	Image	Favicon icon
AdGholas [45]	Maladvertising Campaign	Exploitation	Education, Travel	image	jpeg
Sundown [46]	Malvertising campaign	Exfiltration	Internet users	Image	Png
Synccrypt [47]	Click Malicious URL	Install malware	Generic Internet users	Image	Jpeg
ZeroT [48]	Phishing Campaign	Command and Control	Not known the target	Image	Bmp
Verymal [17]	Maladvertising	Downloading Shlayer Malware	Internet users	Image	Jpeg
Waterbug [49]	Legitimate application vulnerabilities	Downloading DLL	Government, Education, IT	Audio	Wav
Loki Bot [4]	Phishing emails	Install malware	Internet users	Image, Video	Jpg, Video formats

Further Reading:

- https://votiro.com/blog/the-rise-of-ai-powered-steganography-attacks/#:~:text=By embedding malicious code within,of defense without raising alarms.

Challenges with Stegomalware

> *“The existing solutions are designed to mainly focus on the signature and behavior analytics of the executable files for the malware identification. Additionally, the number of known malware samples hidden in multimedia file format are very less for applying the ML/DL techniques and predict the malware files."*

- Current antivirus looks for known attack signatures or behaviours that it can flag
- There are very few examples of malware that is hidden inside images, videos, or audio files using steganography
- Because of the limited sample size, it is difficult to train ML or DL models effectively

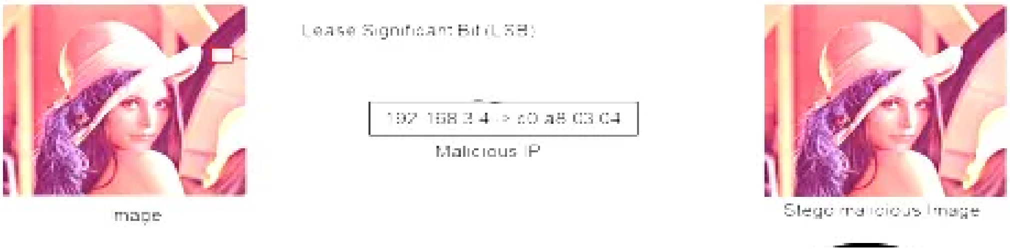
> *“Security professionals may find it even difficult to perform stegomalware mitigation activities such as blocking file formats at Firewall, Intrusion Detection/Prevention System or endpoint security level, as the image, audio and video files formats are extensively used in the enterprise for business operations and transactions”*

- Trusted file formats (PDFs, JPEGs, etc) are required and operated at the enterprise level
- Mitigating these file formats would block stegomalware but would also significantly disrupt business activities

> *“The file signature update in security tools also may not be a viable option, as the malware*
> *constantly evolves to evade the detection”*

- antivirus tools need to be frequently updated with new attack signatures to remain proactive with threats
- Creators of malware will regularly change or obfuscate their code to create new variants that will bypass detection by security tools

Example of Stegomalware Usage



- The cover image here is embedded with a malicious command and control server IP address payload using the Least significant bits algorithm (LSB)
- Existing malware on the system could connect to that malicious IP address, which points to an attackers command and control or C2 server
- After connecting, the server could exfiltrate data, receive and send commands or download more malware onto the system
- The user does not need to interact with the image for stegomalware of this calibre to be impactful

Steogmalware Creation/Multimedia File Formats

Popular Tools used for Steganography (Supports JPEG, BMP file formats)

- Steghide
- OpenStego
- Hide’N’Send

- SSutie Piscel
- Camouflage
- Xiao
- Openpuff

JPEG (.jpg, .jpeg)

- Most common format for image sharing.
- Uses lossy compression via Discrete Cosine Transform (DCT).
- Structure includes:
    - SOI (Start of Image) and EOI (End of Image)
    - Compression markers: DHT, DQT, SOS
    - APP1 segment often contains Exif metadata (device info, thumbnails)

How is malware hidden?

- DCT coefficients are subtly modified to embed data
- Exif metadata can also be used:

> *“In particular, the Exif data and thumbnail images storage give more space to embed the malware content.”*

- Thumbnail images in Exif are exploited to hide code or C2 configuration.

GIF (.gif)

- Supports animations through multiple frames.
- Uses a global and/or local color table to manage pixel data.
- Lightweight and widely used for banners and ads.

How is malware hidden?

- Comment blocks, application extensions, or after the end marker (trailer `0x3B`):

> *"An adversary may add the malicious payload after the trailer marker to hide the content and infect the victim*

> *machines when triggered in the scheduled action of items"*

- Some malware hides malicious URLs or scripts in these fields.
- GIFs in malvertising can be common — a user clicks an ad, triggers the embedded payload or redirection.

> *"...can be used by adversary to host fake ads with embedded malicious URLs so that the victims are redirected to the malicious web pages when the user click the ad page"*

WAV (.wav)

- Follows RIFF (Resource Interchange File Format).
- Structure:
  - `RIFF` chunk (file descriptor)
  - `fmt` chunk (format: mono/stereo, sample rate, etc.)
  - `data` chunk (raw audio samples)

How is malware hidden?

> *"The data may be embedded with malicious payload to hide the stegomalware and use audio format WAV as a carrier to deliver it to the victim device"*

Deep Learning Techniques for Steganalysis (Detection)

> *"The universal steganalysis techniques such as ML or DL based solution techniques can be applied to learn the behavior of the image using statistical features or unique feature taken from the sample data"*

- Deep learning (DL) approaches are increasingly applied in image steganalysis, particularly for detecting stegomalware, due to their ability to automatically extract complex, subtle patterns that reveal hidden data.

CNN-Based Steganalysis

> *“The image steganalysis review includes statistical feature and ML based solutions in spatial and JPEG domain, Rich models with Ensemble classifier and Deep learning based steganalysis solutions”*

- CNNs are the backbone of most DL-based steganalysis solutions
  - Automatically extracts spatial patterns from pixel values — helpful in detecting the slight perturbations introduced by steganography

SRNet - Spatial Rich Network

This model is particularly highlighted in the paper due to its performance:

> *“The authors showed that JPEG domain NS (J-Cov-NS) achieved high embed capacity and security when tested with DCTR and SRNet steganalysis”*

- Designed specifically for image steganalysis.
- Unlike traditional CNNs, SRNet uses preprocessing filters tailored to expose stego noise.
- Detects subtle spatial/frequency domain perturbations

GAN Based Steganalysis

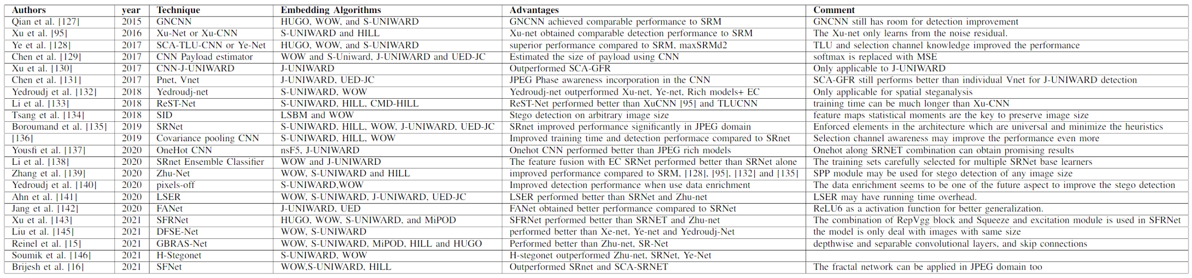
> *“GAN based image steganography typically includes the generator, discriminator and steganalysis modules to iteratively generate stego images such that minimize the distortion between stego and cover image”*

Generative Adversarial Networks (GANs) are dual-purpose in this domain:

- Steganography: Used to generate indistinguishable stegoimages.
- Steganalysis: The discriminator network (from GAN) is repurposed to classify stego vs. non-stego images.

Summary of Deep Learning Models

TABLE X
DEEP LEARNING STEGANALYSIS



Authors	year	Technique	Embedding Algorithms	Advantages	Comment
Qian et al. [127]	2015	GNCNN	HUGO, WOW, and S-UNIWARD	GNCNN achieved comparable performance to SRM	GNCNN still has room for detection improvement
Xu et al. [95]	2016	Xu-Net or Xu-CNN	S-UNIWARD and HILL	Xu-net obtained comparable detection performance to SRM	The Xu-net only learns from the noise residual.
Ye et al. [128]	2017	SCA-TLU-CNN or Ye-Net	HUGO, WOW, and S-UNIWARD	superior performance compared to SRM, maxSRMd2	TLU and selection channel knowledge improved the performance
Chen et al. [129]	2017	CNN Payload estimator	WOW and S-Uniward, J-UNIWARD and UED-JC	Estimated the size of payload using CNN	softmax is replaced with MSE
Xu et al. [130]	2017	CNN-J-UNIWARD	J-UNIWARD	Outperformed SCA-GFR	Only applicable to J-UNIWARD
Chen et al. [131]	2017	Pnet, Vnet	J-UNIWARD, UED-JC	JPEG Phase awareness incorporation in the CNN	SCA-GFR still performs better than individual Vnet for J-UNIWARD detection
Yedroudj et al. [132]	2018	Yedroudj-net	S-UNIWARD, WOW	Yedroudj-net outperformed Xu-net, Ye-net, Rich models+ EC	Only applicable for spatial steganalysis
Li et al. [133]	2018	ReST-Net	S-UNIWARD, HILL, CMD-HILL	ReST-Net performed better than XuCNN [95] and TLUCNN	training time can be much longer than Xu-CNN
Tsang et al. [134]	2018	SID	LSBM and WOW	Stego detection on arbitrary image size	feature maps statistical moments are the key to preserve image size
Boroumand et al. [135]	2019	SRNet	S-UNIWARD, HILL, WOW, J-UNIWARD, UED-JC	SRnet improved performance significantly in JPEG domain	Enforced elements in the architecture which are universal and minimize the heuristics
[136]	2019	Covariance pooling CNN	S-UNIWARD, HILL, WOW	Improved training time and detection performace compared to SRnet	Selection channel awareness may improve the performance even more
Yousfi et al. [137]	2020	OneHot CNN	nsF5, J-UNIWARD	Onehot CNN performed better than JPEG rich models	Onehot along SRNET combination can obtain promising results
Li et al. [138]	2020	SRnet Ensemble Classifier	WOW and J-UNIWARD	The feature fusion with EC SRNet performed better than SRNet alone	The training sets carefully selected for multiple SRNet base learners
Zhang et al. [139]	2020	Zhu-Net	WOW, S-UNIWARD and HILL	improved performance compared to SRM, [128], [95], [132] and [135]	SPP module may be used for stego detection of any image size
Yedroudj et al. [140]	2020	pixels-off	S-UNIWARD,WOW	Improved detection performance when use data enrichment	The data enrichment seems to be one of the future aspect to improve the stego detection
Ahn et al. [141]	2020	LSER	WOW, S-UNIWARD, J-UNIWARD, UED-JC	LSER performed better than SRNet and Zhu-net	LSER may have running time overhead.
Jang et al. [142]	2020	FANet	J-UNIWARD, UED	FANet obtained better performance compared to SRNet	ReLU6 as a activation function for better generalization.
Xu et al. [143]	2021	SFRNet	HUGO, WOW, S-UNIWARD, and MiPOD	SFRNet performed better than SRNET and Zhu-net	The combination of RepVgg block and Squeeze and excitation module is used in SFRNet
Liu et al. [145]	2021	DFSE-Net	WOW, S-UNIWARD	performed better than Xe-net, Ye-net and Yedroudj-Net	the model is only deal with images with same size
Reinel et al. [15]	2021	GBRAS-Net	WOW, S-UNIWARD, MiPOD, HILL and HUGO	Performed better than Zhu-net, SR-Net	depthwise and separable convolutional layers, and skip connections
Soumik et al. [146]	2021	H-Stegonet	S-UNIWARD, WOW	H-stegonet outperformed Zhu-net, SRNet, Ye-Net	
Brijesh et al. [16]	2021	SFNet	WOW,S-UNIWARD, HILL	Outperformed SRnet and SCA-SRNET	The fractal network can be applied in JPEG domain too

References

Chaganti, R., Ravi, V., Alazab, M., & Pham, T. D. (2021). Stegomalware: A Systematic Survey of MalwareHiding and Detection in Images, Machine LearningModels and Research Challenges. *Arxiv.org*. https://doi.org/10.48550/arXiv.2110.02504

ChatGPT Q&A

https://chatgpt.com/share/67f22f91-a790-8001-8f25-e5b7e7b34269