고도화된 스테가노그래피 공격기술 조사

이재욱[1], 최상훈[2], 박기웅[3†]
[1]세종대학교 SysCore Lab. 석사과정
[2]세종대학교 SysCore Lab. 연구교수
[3]세종대학교 정보보호학과 교수
dlwodnr59@naver.com, csh0052@gmail.com, woongbak@sejong.ac.kr

An Investigation into Advanced Steganography Attacks

Jae-Wook Lee[1], Sang-Hoon Choi[2], Ki-Woong Park[3†]
[1-2] SysCore Lab., Sejong University
[3]Dept. of Computer and Information Security, Sejong University

요 약

정보기술 분야의 발전으로 인해 정보의 중요성, 보안성, 기밀성, 개인 정보에 대한 보호가 더욱 강조되고 있다. 여러 정보보호 기술 중 스테가노그래피는 이미지, 텍스트, 오디오, 비디오 파일과 같은 적절한 멀티미디어 매체에 비밀 데이터를 숨기는 기술이다. 하지만 최근 악성코드는 탐지를 회피하기 위한 기술로 스테가노그래피를 사용하고 있다. 또한 인공지능의 발전으로 인해 인공지능을 활용한 스테가노그래피 공격 기법과 연구들이 점차 증가하고 있다. 본 논문에서는 스테가노그래피를 이용한 공격 기법과 해당 기술이 적용된 악성코드들의 사례에 대해 분석하고, 스테가노그래피를 활용한 악성코드의 대응 방법에 대한 향후 연구를 소개한다.

1. 서론

정보기술의 발전으로 정보의 중요성, 보안성, 기밀성, 그리고 개인정보 보호의 필요성이 강조되고 있다. 이러한, 정보들은 공격에 노출되어 심각한 문제가 되었으며 문제를 해결하기 위해 정보를 보호하는 위한 많은 기술이 등장했다. 이러한 기술에는 암호화와 스테가노그래피 기술이 포함되어 있다 [1]. 스테가노그래피는 이미지, 텍스트, 오디오, 비디오 파일과 같은 적절한 멀티미디어 매체에 비밀 데이터를 은닉하는 기술이다 [2].

하지만, 최근에는 스테가노그래피 기법이 암호화와 같이 공격자가 본인의 악성코드 혹은 페이로드를 은닉하는 기법으로 발전하고 있다. CSO에서 보도한 자료에 따르면 우크라이나와 관련된 조직을 표적으로 삼은 해커 그룹은 이미지 파일의 픽셀 내에 은닉된 악성 페이로드를 전달해왔다 [3]. 해당 해킹 그룹은 IDAT라는 악성코드 로더의 일부로 탐지를 피하기 위해 스테가노그래피를 사용하였다 [3]. 또한 TA558로 추적되는 해커그룹은 스테가노그래피를 난독화 기법으로 활용하여 Agent Tesla, FormBook, Remcos RAT, LokiBot, GuLoader, Snake Keylogger, XWorm 등 다양한 악성코드를 배포하는 것으로 확인되었다 [4]. 이와 같이 스테가노그래피를 데이터를 보호하는 것이 아닌 악성 행위를 하기 위해 사용되는 경우도 확인할 수 있다.

최근 인공지능의 발전으로 스테가노그래피 기술 또한 정교해지면서 고도화되고 있다. 이러한 기술의 발전을 바탕으로 인공지능을 활용한 스테가노그래피 공격 기법은 점점 더 증가하고 있고, 관련 연구 또한 다수 등장하고있다 [5]. 본 논문에서는 점점 발전하는 스테가노그래피를 활용한 공격 기법에 대한 연구를 조사하고 스테가노그래피를 적용한 악성코드의 동향을 분석한다.

본 논문의 구성은 다음과 같다. 제2장에서는 스테가노그래피를 활용한 공격 기법을 살펴보고, 제3장에서는 스테가노그래피 기술이 사용된 악성코드에 대해 기술한다. 제4장에서는 결론을 제시하고 향후 연구 방향에 대해 논의한다.

†교신저자: 박기웅 (세종대학교 정보보호학과 교수)

2. 스테가노그래피를 이용한 공격 연구

2.1 에코 은닉 기반 백도어

24년도 Mengyuan Zhang 외 5인은 Echo Hiding을 활용한 주파수 영역 기반 백도어 공격 기법을 제안하였다 [6]. Echo Hiding은 오디오 스테가노그래피 기법중 하나이고, 사람의 청각 시스템이 에코를 자연스로운 소리로 인식하는점을 이용하여 은닉 데이터가 탐지되기 어렵게 한다. 오디오 파일에 에코를 추가하고, 해당 주파수에 은닉 데이터를 삽입하여 백도어 트리거를 생성한다. 실험에서는 구글 Speech Commands Dataset을 사용하여 제안된 기법을 평가하였다. 실험 결과, 제안된 방법은 백도어 공격 성공률이 94% 이상 기록했으며, 이는 매우 낮은 비율의 오염된 데이터로도 높은 성공률을 달성할 수 있음을 보여주었다. 또한 사람의 귀로는 쉽게 인식이 되지 않으며 원본 오디오의 품질도 크게 저하되지 않음을 보였다.

2.2 FDNet

24년도 Liang Dong 외 6인 DNN에 대한 백도어 공격을 주파수에서 구현하는 FDNet 방식을 제안하였다. 기존의 백도어 공격은 주로 공간 영역에서의 데이터 변화 트리거 패턴에 의존해 쉽게 탐지되기 때문에 배포에 어려움이 있었으나, 본 연구는 이미지의 주파수에서 자연적으로 발생하는 노이즈를 통합하는 방식을 사용하여, 트리거가 이미지의 의미적 내용에 미미한 변경만을 가져오고 거의 감지되지 않도록 한다. 또한, 네거티브 샘플링 기법을 도입하여 신경망이 트리거 패턴으로서 더 풍부한 차이를 학습하도록 유도하며, 이를 통해 기계 기반 방어 메커니즘에 대한 강인성을 제공한다.

실험은 컨볼루션 신경망, 비주얼 트랜스포머, MLP-Mixer 모델과 MNIST, CIFAR-10, GTSRB, ImageNet 데이터셋에서 수행되었으며, 실험 결과는 높은 공격 성공률을 보여주었다. 특히, All-to-one 시나리오에서 거의 100%의 성공률을 달성하였고, All-to-all 시나리오에서는 90% 이상의 성공률을 기록하였다(단, ImageNet 제외). 이 연구는 DNN이 인간에게 거의 감지되지 않는 주파수의 불일치를 포착할 수 있음을 알아냈으며, 주파수에서의 공격 방법은 기존의 공간 기반 공격보다 높은 은닉성과 강인성을 제공한다 [7].

2.3 SDriBA

24년도 Weixuan Tang 외 4인은 INN을 이용한 스테가노그래피 기반의 백도어 공격 프레임워크인 SDriBA를 제안하였다. 해당 프레임워크는 백도어 공격에서 트리거의 주입과 인식을 데이터 은닉 및 은닉 데이터 추출 과정으로 모델링 하여, 공격의 은밀성과 효과성을 높이는 것을 목표로 한다. 해당 연구에서는 INN을 사용하여 은닉 데이터 삽입 및 추출 과정을 수행한다. 첫 번째로 깨끗한 이미지와 트리거 이미지를 입력을 받아서 오염된 이미지를 생성하고 이 과정에서 노이즈를 생성한다. 이후 오염된 이미지는 변환 레이어를 거쳐서 최종적으로 완성된 포이즌 이미지가 생성된다. 해당 과정은 트리거가 피해 모델에 의해 인식될 수 있도록 설계하였다. SDriBA의 효과를 검증하기 위해 다양한 실험을 진행하였다. 실험에서는 훈련 세트에서 무작위로 선택된 이미지에 트리거를 주입하고 이를 통해 해당 모델을 훈련시켰다[8].

2.4 SBA

24년도 Weida Xu 외 2인은 스테가노그래피 기반의 SAB(Stealing and Robust Backdoor) 공격 기법을 제안하며, 연합 학습 환경에서 백도어 공격을 수행하는 새로운 방식을 소개한다. 이미지 스테가노그래피를 사용해 트리거를 생성하고, 다중 손실 계산을 통해 백도어를 은닉한다. CIFAR-10, CIFAR-100, Fashion-MNIST 데이터셋을 사용한 실험에서, SAB는 기존 BadNets 및 DBA보다 높은 공격 성공률(ASR)과 악성코드 지속성을 보였으며, 공격이 중단된 이후에도 일정 기간 동안 높은 ASR을 유지했다 [9].

2.5 악성코드 트리거 프레임워크

22년도 Lamia Almehmadi 외 3인은 스테가노그래피를 활용해 이미지 파일에 악성코드를 은닉하고, 특정 조건에서만 악성코드를 활성화할 수 있는 위치 기반 트리거링 시스템을 제안하였다. PNG 형식의 이미지 파일에 JavaScript 기반 익스플로잇 코드와 지오로케이션 코드를 삽입한 후, Stegosploit toolkit을 사용하여 HTML과 PNG를 결합한 polyglot 파일을 생성한다. 이 파일을 취약한 웹 브라우저에서 열면 악성코드가 즉시 실행된다. 다양한 지리적 위치 추적 기술을 통해 실험을 진행한 결과, 특정 위치에서만 악성코드가 실행됨을 확인하였다. 단점으로는 위치 정보의 정확성이 부족할 경우 트리

거가 실패할 수 있다는 점이다 [10].

3. 스테고멀웨어 공격 사례

최근 고도화 된 악성코드는 탐지를 회피하기 위해 스테가노그래피 기술을 사용하고 있다. 이와 같은 스테가노그래피 기술이 적용된 악성코드를 스테고멀웨어(stegomalware)라고 한다. 스테고멀웨어는 감염된 시스템에서 흔적을 감추는 것뿐만 아니라, 적대적인 인프라와의 통신 역시 은닉한다. 다음은 스테고멀웨어의 사례에 대해 설명한다 [11].

3.1 Magecart

해당 공격자는 악성코드를 은닉하기 위해 스테가노그래피를 사용하였으며, JavaScript 코드를 공백, 탭, 줄바꿈 등의 보이지 않는 문자로 숨겨 CSS 파일에 삽입하였다. 감염된 jQuery 스크립트는 이 숨겨진 페이로드를 추출하고 실행하여 사용자의 정보나 결제 데이터를 탈취했다. 이 방법은 감지하기 어려운 텍스트 스테가노그래피 기법을 사용하였으며, 웹 기반의 데이터 탈취 공격이 점점 더 복잡해지고 있음을 보여준다 [12].

3.2 Webbfuscator

해당 악성코드는 Golang으로 작성된 악성코드이다. 감염은 악성 문서 “Geos-Rates.docx”가 첨부된 피싱 메일로 시작되며, 이 메일에는 템플릿 파일이 다운로드된다. 이 파일에는 Office 제품군에서 매크로가 활성화된 경우 자동으로 실행되는 난독화된 VBS 매크로가 포함되어 있다. 이후, 코드는 원격 리소스(“xmlschemeformat[.]com”)에서 JPG 이미지(“OxB36F8GEEC634.jpg”)를 다운로드하고, certutil.exe를 사용하여 실행 파일(“msdllupdate.exe”)로 디코딩한 후 실행한다. 해당 악성코드는 이미지 스테가노그래피 기법을 사용하였다 [13].

3.3 VinSelf

해당 악성코드는 먼저 Google Docs에서 "colors.bmp" 파일을 가져온다. 이후, 이 함수는 각 픽셀의 각 색상에서 LSB(Least Significant Bit, 최하위 비트)를 추출하여 이미지의 픽셀당 3비트의 데이터를 생성한다. 모든 LSB가 추출된 후, 비트 스트림의 각 바이트가 역순으로 변경된다. 추출한 데이터를 해독하여 C&C 서버의 정보를 획득할 수 있다 [14].

3.4 PNGLoader

해당 악성코드는 CLRLoader 악성코드를 메모리에 실행시킨 후 PNGLoader인 DLL을 실행시킨다. PNGLoader는 PNG 파일에 포함된 바이트를 추출한 후 추출한 데이터를 사용해 두 개의 실행 파일을 조합한다. Avast에 따르면 해당 악성코드는 최하위 비트(LSB)인 이미지 스테가노그래피를 이용한 방식이라 언급하였다. PNG 파일에 숨겨진 페이로드는 백신 프로그램이 검색할 수 없는 PowerShell 스크립트였고, 두 번째 페이로드는 Drop Box 파일 호스팅 서비스를 악용하여 C2 통신, 파일 유출 등을 유발하였다 [15].

3.5 스테고멀웨어 공격 사례 분석

<표 1> 악성코드별 사용한 스테가노그래피 기술

	Text	Image	Audio	Video
Magecart	O	X	X	X
Webbfuscator	X	O	X	X
VinSelf	X	O	X	X
PNGLoader	X	O	X	X

결과적으로 표 1과 같이 최근 고도화 된 악성코드는 이미지 스테가노그래피 기술을 적용한 사례가 많음을 확인할 수 있다. 스테고멀웨어는 악성 행위를 하는 페이로드를 보안 프로그램에 탐지되지 않도록 하기 위해 스테가노그래피 기술을 적용한다.

4. 결론 및 향후 연구

본 논문에서는 스테가노그래피를 활용한 공격 연구와 스테고멀웨어의 실제 사례 및 동작 과정에 대한 조사를 수행하였다. 조사 결과, 스테가노그래피가 백도어 공격에 자주 사용되는 것을 확인할 수 있었으며, 실제 사례에서는 다양한 스테가노그래피 기법들이 적용되고 있음을 알 수 있었다. 향후 연구에서는 이미지 스테가노그래피가 적용된 악성코드에 대한 재생성이라는 주제로 스테고멀웨어의 악의적인 행위에 대해 저지하는 연구를 진행할 것이다.

Acknowledgement

본 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원(IITP)의 정보통신방송혁신인재양성사업(Project No. 2021-0-01816, 50%), 정보통신방송기술 국제공동연구(Project No. RS-2022-00165794, 30%), 한국연구재단(NRF) 중견후속연구사업(Project

No. RS-2023-00208460, 20%)의 지원을 받아 수행된 연구임.

참고문헌

[1] Khalaf, A. M., & Lakhtaria, K. (2024, February). A review of steganography techniques. In *AIP Conference Proceedings* (Vol. 3051, No. 1). AIP Publishing.

[2] Cheddad, A., Condell, J., Curran, K., & Mc Kevitt, P. (2010). Digital image steganography: Survey and analysis of current methods. *Signal processing*, *90*(3), 727-752.

[3] CSO Staff. (2024, February 26). Hacker group hides malware in images to target Ukrainian organizations. CSO Online. https://www.csoonline.com/article/1309858/hacker-group-hides-malware-in-images-to-target-ukrainian-organizations.html

[4] Hacker News Staff. (2024, April 16). TA558 hackers weaponize images for wide-scale phishing attacks. *The Hacker News*. https://thehackernews.com/2024/04/ta558-hackers-weaponize-images-for-wide.html

[5] S. M. J. a. Abdalwahid, W. A. Hashim, M. G. Saeed, S. A. Altaie and S. W. Kareem, "Investigating the Effectiveness of Artificial Intelligence in Watermarking and Steganography for Digital Media Security," *2024 21st International Multi-Conference on Systems, Signals & Devices (SSD)*, Erbil, Iraq, 2024, pp. 552-561, doi: 10.1109/SSD61670.2024.10549272.

[6] M. Zhang, S. Ji, H. Cai, H. Dong, P. Zhang and Y. Li, "Audio Steganography Based Backdoor Attack for Speech Recognition Software," *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, Osaka, Japan, 2024, pp. 1208-1217, doi: 10.1109/COMPSAC61105.2024.00161.

[7] Dong, L., Fu, Z., Chen, L., Ding, H., Zheng, C., Cui, X., & Shen, Z. (2024). FDNet: Imperceptible backdoor attacks via frequency domain steganography and negative sampling. *Neurocomputing*, *583*, 127546.

[8] Tang, W., Li, J., Rao, Y., Zhou, Z., & Peng, F. A Trigger-Perceivable Backdoor Attack Framework Driven by Image Steganography. *Available at SSRN 4886265*.

[9] Xu, W., Xu, Y., & Zhang, S. (2024). SAB: A Stealing and Robust Backdoor Attack based on Steganographic Algorithm against Federated Learning. *arXiv preprint arXiv:2408.13773*.

[10] Almehmadi L, Basuhail A, Alghazzawi D, Rabie O. Framework for Malware Triggering Using Steganography. Applied Sciences. 2022; 12 (16):8176. https://doi.org/10.3390/app12168176

[11] Knöchel, M., & Karius, S. (2024, June). Text Steganography Methods and their Influence in Malware: A Comprehensive Overview and Evaluation. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security* (pp. 113-124).

[12] Jscrambler. 2022. Steganography in a Magecart Attack. https://jscrambler.com/ blog/steganography-in-magecart-attack

[13] Bill Toulas. 2022. Hackers Hide Malware in James Webb Telescope Images. BleepingComputer. https://www.bleepingcomputer.com/news/security/hackershide-malware-in-james-webb-telescope-images/

[14] Airbus. 2022. Vinself Now with Steganography - Airbus Defence and Space Cyber. Airbus. https://www.cyber.airbus.com/vinself-now-steganography/

[15] Bill Toulas. 2022. Worok Hackers Hide New Malware in PNGs Using Steganography. BleepingComputer. https://www.bleepingcomputer.com/news/security/worokhackers-hide-new-malware-in-pngs-using-steganography/