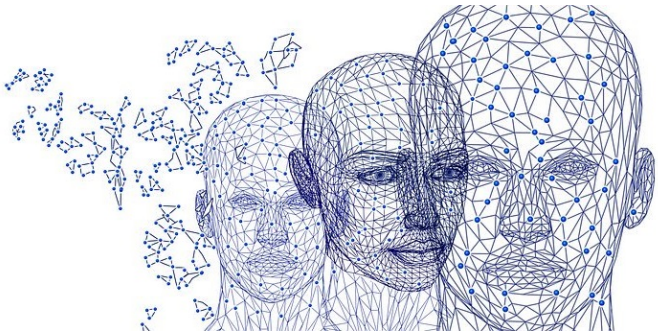
March 2021, Issue Code NL-2021-5

DDD (Digital Data Deception) Technology Watch Newsletter

Table of Contents



"*All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near.*"

— Sun Tzu, *The Art of War*

Editors: Enes Altuncu, Virginia Franqueira, Sanjay Bhattacherjee and Shujun Li
Affiliation: Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), University of Kent, UK
Contact Us: ddd-newsletter@kent.ac.uk

Editorial

In this fifth issue of the Digital Data Deception (DDD) Technology Watch Newsletter, we cover 3 main topics relevant for DDD.

The first topic is "Attacks in AI Models". We revisit *Backdoor Attacks*, a topic that was covered in the last issue of the newsletter. We also focus on *Data Poisoning Attacks* and *Model Poisoning Attacks* which are classes of adversarial attacks affecting AI-based systems.

The second topic is "Fact-Checking Technology", which is increasingly relevant to effectively and efficiently detecting deceptive content and preventing misinformation and disinformation. This threat has become a global issue which can hinder public health (`https://disinformationindex.org/2021/02/ad-funded-covid-19-conspiracy-sites-a-look-at-the-eu/`, and negatively impact inter-state relations and democracy or even encourage financing of extremist groups (`https://www.theguardian.com/world/2021/mar/10/us-far-right-extremists-millions-social-cryptocurrency`). This part focuses on *Fact-Checking Based on Crowd-Sourcing* and *Fact-Checking Based on Other Approaches*.

The third and final topic is "Information Hiding". The last newsletter focused specifically on "Information Hiding in Images", while this newsletter extends the topic to cover a variety of digital media used to hide information from anyone, except the sender and agreed receiver, and escape detection mechanisms. The section covers *Steganography Across Different Media*, *Coverless Steganography*, and *Steganalysis*. Two recent examples of steganography in practice highlight the relevance of its use for deception with malicious intent. The blog post `https://threatpost.com/magecart-attackers-stolen-data-jpg/164815/` reports criminals concealing stolen credit card data in jpg images available on a compromised website. A new campaign of the "ObliqueRAT Remote Access Trojan" (`https://www.zdnet.com/article/obliquerat-trojan-now-hides-in-images-on-compromised-websites/`) also uses steganography to hide their malware payload in bmp image files. This issue summarises 22 research papers published since 2019, and cites two extended versions of summarised papers. The search for papers followed a venue-driven systematic literature review (SLR) approach.

We hope you enjoy reading this issue. Feedback is always welcome, and should be directed to ddd-newsletter@kent.ac.uk.

List of Acronyms

- ABC: Australian Broadcasting Corporation
- ABS: Artificial Brain Stimulation
- AI: Artificial Intelligence
- API: Application Programming Interface
- BPW: Bit Per Word
- CIFAR: Canadian Institute for Advanced Research
- CR: Credibility Review
- DCT: Discrete Cosine Transform
- DNN: Deep Neural Network
- EER: Equal Error Rates
- GAN: Generative Adversarial Network
- GTSRB: German Traffic Sign Recognition Benchmark
- ICS: Industrial Control Systems
- ILSVRC: ImageNet Large Scale Visual Recognition Challenge
- IMDB: Internet Movie Database
- ISIC: International Skin Imaging Collaboration
- IT: Information Technology
- JPEG: Joint Photographic Experts Group
- OT: Operational Technology
- LSTM: Long Short-Term Memory
- MF: Matrix Factorisation
- MNIST: Modified National Institute of Standards and Technology
- MOISE: Model of Organization for Multi-agent Systems
- NC: Neural Cleanse
- NeuMF: Neural Matrix Factorisation
- OT: Operational Technology
- PLC: Programmable Logic Controllers
- PubFig: Public Figures
- QIM: Quantization Index Modulation
- ReLU: Rectified Linear Unit
- ResNet: Residual Neural Network
- RNN: Recurrent Neural Network
- RONI: Reject on Negative Impact
- SCAn: Statistical Contamination Analyzer
- TaCT: Targeted Contamination Attack
- UA: Unified Architecture
- VGG: Visual Geometry Group

Attacks in AI Models

Introduction

As a continuation of the previous issue, this section focuses on adversarial attacks performed in the training time of AI models. More precisely, the section covers the publications from top security and AI venues, regarding backdoor attacks, data poisoning attacks and model poisoning attacks as well as possible detection and mitigation techniques.

Backdoor Attacks

Wang et al. [22] presented a detection and mitigation system, namely *Neural Cleanse (NC)*, for Deep Neural Network (DNN) backdoor attacks. NC aimed to identify backdoors and reconstruct possible triggers. As shown in Figure 1, the system measures the minimum amount of perturbation necessary to change all inputs from each region (B and C) to the target region (A) in order to detect backdoors. In addition, the proposed framework considered multiple backdoor mitigation techniques, including input filters, neuron pruning and unlearning. The authors evaluated NC by using different tasks (and datasets) such as hand-written digit recognition (Modified National Institute of Standards and Technology database (MNIST)), traffic sign recognition (German Traffic Sign Recognition Benchmark (GTSRB)) and face recognition (YouTube Face, Public Figures (PubFig) and Visual Geometry Group (VGG) Face). For each task experimented, the results showed that the proposed system was able to decrease the attack success rates from 97-99% to 0-6%. Lastly, the authors discussed advanced backdoor attacks challenging the limitations of the proposed design. These are complex triggers, larger triggers, multiple infected labels with separate triggers, single infected label with multiple triggers and source-label-specific (partial) backdoors.

Editorial Comments

NC, proposed by Wang et al. [22], was one of the first systems for backdoor attack detection (together with Artificial Brain Stimulation (ABS) [14] mentioned in the previous issue). Although it produced promising results for conventional backdoor attacks where an image with a trigger is injected into the training data, more recent backdoor attacks follow more complex methods to perform the attacks, which can bypass detection systems such as NC and ABS. However, these systems are still used as a benchmark in relevant studies.

Saha et al. [21] proposed a backdoor attack where poisoned data is labelled correctly and contains a hidden trigger, making the poisoned data visually unidentifiable at training time. More precisely, the authors optimised the attack with poisoned images that are close matches to target images with regard to the pixel space and source images that have been patched by the trigger in the feature space. The steps of the proposed attack are shown in Figure 2 with an example. The authors also managed to generalise the proposed attack to unseen images and random trigger locations. To evaluate the proposed hidden trigger backdoor attack, they used ImageNet and Canadian Institute for Advanced Research (CIFAR)10 datasets for the experiments they performed. The experiments with various image classification settings demonstrated that the proposed attack was able to decrease the classification accuracy drastically. Lastly, it was reported that the proposed attack was still successful against a detection algorithm which uses spectral signatures.

Di et al. [8] proposed a *targeted contamination attack* (TaCT) which aimed to cause backdoor attack images to be less distinguishable from benign images to make protection harder. The idea behind the attack is adding correctly labelled samples stamped with the backdoor trigger (i.e., cover samples) into the training data in addition to poisoning training data with a set of attack images (i.e., images from the source classes merged with the trigger and assigned with the target label). Moreover, the authors introduced *statistical contamination analyzer* (SCAn) technique which is a backdoor detection technique based upon statistical properties of the representations produced by an infected model. For the evaluation of the proposed techniques, the authors used four datasets which were GTSRB, ImageNet Large Scale Visual Recognition Challenge (ILSVRC)2012, MegaFace and CIFAR-10. The experiments showed that TaCT achieved an attack success rate of 97%,

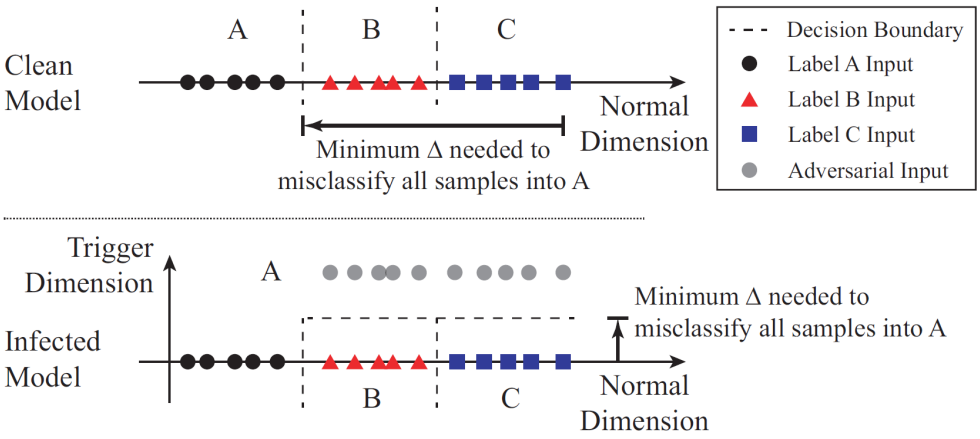


Figure 1: A simplified illustration of the key intuition in detecting backdoor, proposed by Wang et al. [22].

and none of the existing solutions used in the experiments were able to reduce its attack success rate. Furthermore, SCAn was shown to be robust against not only TaCT, but also various backdoor attacks, including blending-trigger, poison frogs and multiple target-trigger attacks.

Lovisotto et al. [15] introduced *template poisoning attacks* for biometric systems. Templates represent sensitive user data in biometric authentication systems, which need to be stored securely to protect the secrecy of the biometric trait. Therefore, user templates are kept in user's devices, which prevents them from being changed directly by an attacker without compromising the device. In this manner, the authors proposed a backdoor attack for face recognition systems to exploit the template update procedure in which the system replaces or adds recently seen samples to the user template. The attack was evaluated by considering three different DNNs: FaceNet, the VGG16 Face descriptor and Residual Neural Network (ResNet)-50, and using VGGFace2 dataset. Results showed that the proposed attack was successful in over 70% of cases with less than ten injection attempts, in white-box scenarios. Lastly, the authors designed a poisoning detection technique, and tested it with a set of intra-user variability factors, i.e., pose, facial hair and (sun)glasses, extracted with Google Vision Application Programming Interface (API). In consequence, the proposed detection method achieved Equal Error Rates (EER) of 7-14% for the detection and identified over 99% of attacks after only two sample injections.

Editorial Comments

Template poisoning attacks, mentioned by Lovisotto et al. [15], are performed against template update procedures used in face recognition systems, including Apple's FaceID and TouchID. The proposed attack focused on a real-world case which makes the study quite important. As the authors mentioned, the transferability of the proposed attack needs to be improved for black-box settings to increase the applicability of the attack.

Data Poisoning Attacks

Huang et al. [13] proposed a data poisoning attack for deep learning based recommender systems, where an attacker injects fake users with carefully crafted ratings to a recommender system so that target items are promoted to a large number of users. The authors formulated the proposed attack as an optimisation problem, such that the injected ratings would maximise the number of normal users to whom the target items are recommended. The authors evaluated the attack by using MovieLens-100K, MovieLens-1M and Last.fm datasets, and Neural Matrix Factorisation (NeuMF) as the target recommender system. Based on the experiments, they compared the performance of the proposed attack with existing attacks, including random attack, bandwagon attack and matrix-factorisation (MF) attack. Results showed that the proposed attack effectively promoted target items and outperformed the existing attacks in both white-box and gray-box settings. Lastly, the authors explored the detection of

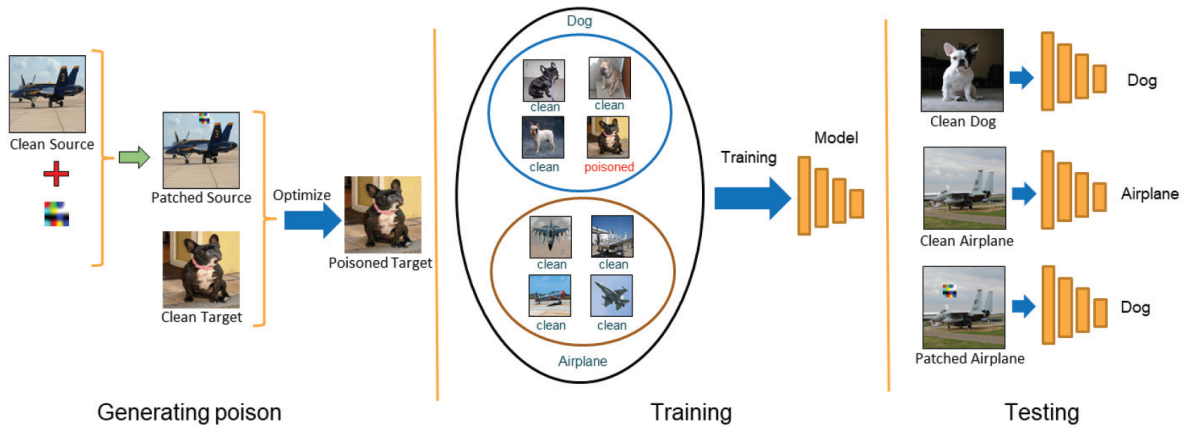


Figure 2: The steps of the generation of hidden trigger backdoor attack, proposed by Saha et al. [21].

fake users via statistical analysis of their rating patterns. While they reported that such a method can detect fake users based on the experimental results, they also showed that the method falsely identified 30% of the fake users constructed by the proposed attack as normal users, meaning that the attack was still effective.

Editorial Comments

The data poisoning attack proposed by Huang et al. [13] is different from other studies in terms of the area it focused on. While most of the studies on backdoor and poisoning attacks concentrated on image-based problems, the proposed attack targeted recommender systems and used text-based rating datasets.

Model Poisoning Attacks

Fang et al. [10] studied *local model poisoning attacks* to Byzantine-robust federated learning. In federated learning, each client devices, i.e., *worker devices*, maintains a *local model* for its local training dataset, and they desire to jointly learn a *global model* maintained by a master device, e.g., a cloud server. Assuming that the attacker has control of some worker devices, the authors aimed to compromise the integrity of the learning process in the training phase by manipulating the local model parameters. The proposed attack was evaluated on four recent Byzantine-robust federated learning methods including Krum, Bulyan, trimmed mean, and median, and they utilised MNIST, Fashion-MNIST, CHMNIST and Breast Cancer Wisconsin (Diagnostic) datasets. The results demonstrated that the proposed attack substantially increased the error rates of the global models. Lastly, the authors explored potential means of defending the proposed attack by generalising two existing defences against data poisoning attacks, Reject on Negative Impact (RONI) and TRIM.

Pang et al. [19] introduced a unified attack framework which jointly optimises adversarial inputs and poisoned models. The authors also explored mutual reinforcement effects between the two attack vectors by conducting experiments on CIFAR10, Mini-ImageNet, International Skin Imaging Collaboration (ISIC) and GTRB datasets together with ResNet18 and ResNet101 DNN models. Results from the experiments showed that the interactions between the two attack vectors demonstrated intriguing mutual reinforcement effects: when launching the unified attack, leveraging one attack vector significantly amplified the effectiveness of the other. According to the authors, such effects can be used to enhance the existing attacks. Lastly, the authors discussed the potential countermeasures for the proposed attack model.

Fact-Checking Technology

Introduction

In today's information age, false or misleading information can propagate rapidly through social media, which can cause serious harm to society. To fight against mis/disinformation, fact-checking at-scale and with high accuracy has become crucial. This section focuses on fact-checking based on human-machine teaming, i.e., crowd-sourcing and other approaches such as social media monitoring and multi-agent organisation.

Fact-Checking Based on Crowd-Sourcing

Hassan et al. [11] analysed 543 posts and 10,221 comments on the Reddit page of an online fact-checking community called *politicalfactchecking*. The community relies on crowd-sourcing to find and verify check-worthy facts relating to U.S. politics. The aim of the study is to explore the roles of crowd, moderators and automation in fact-checking processes. In this manner, the authors also developed an automated argument classification model to analyse the content and separate factual evidence from opinionated or evidence lacking judgement. The model developed using Gradient Boosting was evaluated by performing a 5-fold cross-validation. This resulted with a reported accuracy of 85%. The authors argued that stance detection (identifying someone's reaction towards a claim, e.g., positive, negative or neutral) gives opportunity for automation, although it is a harder problem. They also pointed out that a small number of trained fact-checkers can help validate a large number of facts if they can properly lead crowds in the right direction.

Bhuiyan et al. [3] investigated news credibility assessments made by crowds versus those made by experts to understand when and how ratings between them differ. In this manner, the authors collected a dataset of 4,050 credibility assessments taken from two crowd groups –journalism students and Upwork workers– and two expert groups –journalists and scientists– on a set of 50 news articles. The authors examined the ratings, and they found that factors such as assessor demographics and political leaning, the genre of the article and partisanship of the publication caused differences in rating performance. In addition, as shown in Figure 3, science experts considered accuracy, evidence, and grounding presented in the article while journalism experts put emphasis on publication reputation as the criteria they used to rate news articles. Based on the findings, the authors suggested that different crowd rating tasks might align with different experts rather than broadly rating credibility. Moreover, they reported that a combination of person-oriented strategies (e.g., filtering by demographics), followed by process-centric strategies (e.g., training raters) can facilitate high-quality and at-scale credibility assessment.

Editorial Comments

Bhuiyan et al. [3] considered news articles related to a single topic – climate change. This choice of the topic might represent a bias, since it is known that climate change experts tend to have a consensus on key issues, i.e., they disagree with each other less than on many other topics. This potential issue was also acknowledged by Bhuiyan et al. in [3]. Therefore, further validation of the results reported in [3] on more topics will be necessary.

Pinto et al. [20] proposed a fact-checking process that uses crowd-sourcing to perform the filtering, analysis and classification of news. While the proposed system relies on majority voting by the crowd for veracity assessment, it also contains an optional step for requesting opinion from a specialist. The specialist opinion guides the crowd when proposing a label regarding the veracity of the news. The authors aimed to develop a self-sustainable and democratic process. However, the proposed system has not been implemented yet.

Editorial Comments

The system proposed by Pinto et al. [20] relied on majority voting for decision making. While it allows taking input from a specialist regarding a news item, this does not affect the decision directly. Therefore, it is likely that the verdict on the veracity can be easily manipulated by the crowd - especially for political news where the political leaning of the majority of the crowd members could skew the verdict.

Fact-Checking Based on Other

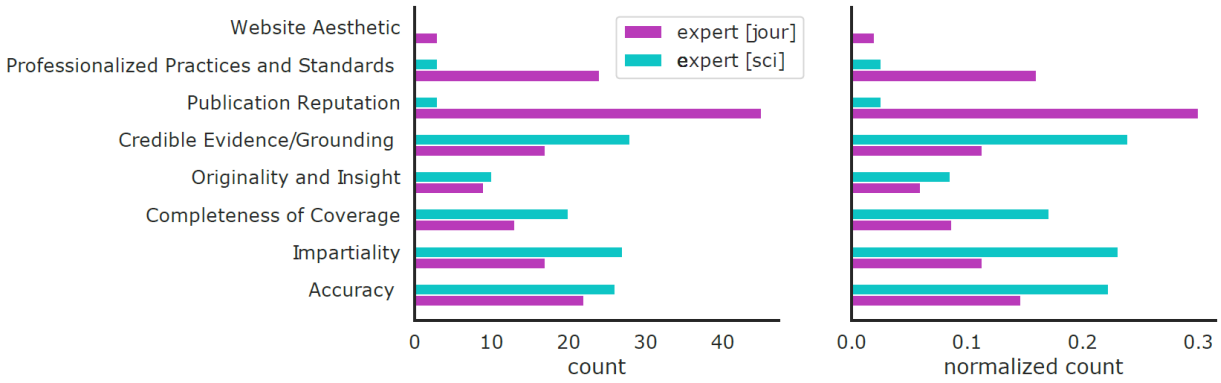


Figure 3: Frequency of the categories in expert explanations for journalists versus scientists from the study of Bhuiyan et al. [3]. While the left graph shows the raw counts, graph on the right shows the counts normalised by the number of explanations made by journalists versus scientists in total.

Approaches

Cerone et al. [5] developed a social media monitoring tool called *Watch 'n' Check* in collaboration with fact-checking experts from the Australian Broadcasting Corporation (ABC) (`https://www.abc.net.au/news/factcheck/`). It aims to assist fact-checkers to detect and target misinformation online. Watch 'n' Check filters tweets by matching the specified keyword or phrase to perform a quantitative real-time analysis on specific topics. The authors preferred to use Twitter as the target platform since it allows for extraction of a representative sample of published information. To evaluate the tool, the authors collected a dataset consisting of 182.1M tweets, extracted between December 1, 2019 and May 1, 2020, and they performed a quantitative analysis for two topics – bushfires in Australia and COVID-19. The proposed tool is still being implemented.

Editorial Comments

The social media monitoring tool developed by Cerone et al. [5] currently considers only Twitter as the target platform. In addition, the authors used Twitter Stream API which provides only 1% of the publicly available tweets. The authors shared the codes of their prototype publicly (`https://github.com/rmit-ir/watch-n-check`).

Wild et al. [23] defined a multi-faceted fact-checking process that considers several credibility indicators and formalised it using the Model of Organization for Multi-agent Systems (MOISE). It stipulates multi-agent interaction and coordination between the individual fact-checking efforts of humans and autonomous agents. In the proposed system, disinformation tackling involves 46 distinct questions which are assigned to six processes, each focusing on a different dimension of evaluating the credibility of an online information resource. The authors used EUFACTCHECK (`https://eufactcheck.eu/`) project as a starting point and abstracted it from its domain-specificity and reliance on experts. Regarding the implementation of the proposed system, the authors have designed and implemented a browser plugin that allows agents to self-assess the credibility of online information. However, the implementation is still ongoing.

Denaux and Gomez-Perez [6] proposed an architecture based on a core concept of *Credibility Reviews* (CRs) that can be used to build networks of distributed bots that collaborate for misinformation detection. The aim of the study was to develop automated systems that can produce machine readable results to enable better collaboration between the stakeholders. The authors implemented the proposed architecture on top of lightweight extensions to *Schema.org*. They also implemented a series of CR bots capable of collaborating to review text-based content such as articles, tweets, sentences and websites. (The source code is available at `https://github.com/rdenaux/acred`.) The authors evaluated the proposed architecture by using Clef'18 CheckThat! Factuality Task, FakeNewsNet and coinform250 datasets. The results were consistent across different datasets, therefore validated the proposed design. However, in order to be used in

real-world settings, the implemented system needs to be improved in terms of precision. The authors extended the proposed system by adding features of crowd-sourcing based detection of inaccurate credibility reviews and error analysis of erroneous reviews [7].

Information Hiding

Introduction

This section covers *Information Hiding* – i.e., Steganography and Steganalysis – expanding on the publications reviewed in the last newsletter issue which focused on "Information Hiding in Images". In terms of terminology, the purpose of steganography is to conceal the existence of hidden communication between a sender and a receiver, creating *covert channels*, where the hidden data/information (called "payload") is embedded into digital media (called "carrier" or "cover") resulting in a "stego carrier" (or "stego-image" or simply "stego"). The stego, containing hidden data, may be protected with a "stego key", pre-agreed between parties.

This section is structured in three parts. The first sub-section summarises papers where data is hidden in a range of covers (i.e., carriers), namely video, audio, network datagrams, and industrial control systems. The second sub-section concentrates on the novel development of "coverless steganography". This overcomes existing steganalysis methods by embedding the hidden data into the stego so that it does not modify its properties and probability distribution, in comparison with the more traditional cover-based methods. The final sub-section reviews developments in steganalysis.

Steganography Across Different Media

Fan et al. [9] pointed out that videos uploaded to social networks are transcoded using lossy mechanisms. As a result, most video steganographic methods are rendered ineffective. They proposed a robust method for video steganography that can function together with the transcoding mechanisms implemented by social networking sites. For the cover, they used the luminance component of the raw video. To hide a message, they used the block statistical feature based Quantization Index Modulation (QIM) algorithm. They designed an iteration in the local transcoder that would allow determining the minimum quantisation step for each video. This is an optimisation technique that provides a beneficial trade-off between the robustness and the visual quality. To further improve robustness and security, they proposed a strategy for selecting (robust) frames from the video. A side channel was built for extracting the correct steganographic messages so that no information had to be shared beforehand between the sender and the receiver. The method proposed in the paper provides robust protection against transcoding in social networks. In the experiments conducted by the authors, the average bit error rate recorded is less than 1%. In terms of security of the covert communication through social networking (like YouTube, Vimeo etc.), they claimed their method is quite robust. Figure 4 provides an overview of the steganographic framework of the method proposed in this paper.

Editorial Comments

The adaptive embedding method proposed in this paper, uses a traditional statistical tool - partitioning the data into mutually exclusive and exhaustive blocks and then computing the statistic for each such block. After that, each block is treated as a homogeneous unit. This usually makes the data coarser than the original data, and every block can henceforth be treated as a unit.

Hildebrandt et al. [12] pointed out novel attacks possible on Industrial Control Systems (ICS). The automation of cyber-physical systems is done primarily using ICSs. Such systems blend the capabilities of information technology (IT) with operational technology (OT), often using analogous principles in analogous contexts. They can provide controlling capabilities in applications with a wide range of complexities - traffic lights, elevators, car manufacturing, nuclear power plants, etc. In many cases, Ethernet-based connectivity is being used for ICSs. Many attack vectors and their corresponding defence mechanisms in IT naturally carry over to such ICSs. This is true for information hiding techniques as well. There have been such attacks on OT systems over the past decade, presumably conducted by nation state actors. In the case of attacks on IT networks, information hiding mechanisms are commonly used by attackers to hide the fact that they have managed to compromise a network. When the attacker is able to maintain control for an extended period of time, we say the compromise is persistent. ICSs are increasingly facing similar attacks. The observed attacks vary in magnitude from being simple automated at-

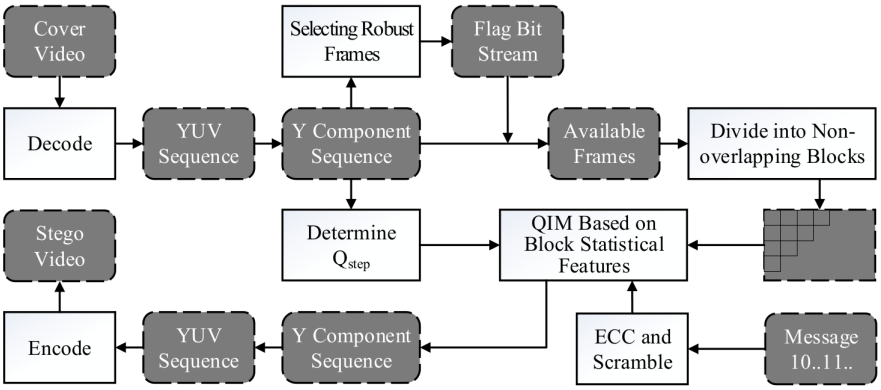


Figure 4: The steganographic framework of the method proposed in Fan et al. [9].

tacks to very targeted attacks by nation state actors to damage components or infrastructures. Information hiding embeds hard-to-detect backdoors to achieve persistent communication channels. Once such an attack has been executed, the attacker gains control of the system in a hidden fashion so that damages may be caused pro-actively as and when necessary. In this paper, the authors focused particularly on potential attacks that may be executed on Programmable Logic Controllers (PLCs) using communication through the OPC Unified Architecture (OPC UA) network protocol. The attack is on a supply chain involving an OPC UA server and a Siemens S7-1500 PLC used as an OPC UA client. The hidden storage channel communicates encrypted control sequences by embedding them in source timestamps that can be set to arbitrary values. The attack solely relies on the programming of the PLC and does not require access to the firmware. For mitigation of the attack, the authors investigated potential approaches for detecting hidden storage channels for a designated node called *warden*. Their approach is based on machine learning. They conducted an experiment with around 46,000 OPC UA read responses that were without a steganographic message and around 7,500 OPC UA read responses with an embedded steganographic message. They used a One-Class-Classifier that yields a detection performance of 89.5% with zero false positives within their experiment.

Editorial Comments

The work of Hildebrandt et al. [12] provided a niche perspective on how attacks on one kind of technology can be easily carried over to another with similar usability.

Lu et al. [16] proposed a new technique for steganography that is based on network data flow, such that the information is protected from interception by unauthorised parties. Their method is inspired by steganographic techniques that use video files as cover. They hide the information on the network into the relationship between timestamps of consecutive packets of a complete session. The technique talks about transferring hidden data and performing secondary identity authentication. Their method can be used on all network data traffic that has a timestamp option bit. However, it does not influence the access of a user to the real data in the network. The authors claimed that there are no known techniques to detect such steganographic techniques. They have experimentally verified and evaluated the steganographic concealment assurances offered by their methods. They claimed that absence of obvious statistical characteristics of the traffic to be one of the main reasons for the assurance.

Editorial Comments

The work of Lu et al. [16] makes use of unused fields in network data traffic for hiding information. While the techniques are nice, the level of assurances of information being hidden in unused fields in packets are not very convincing. The commonly unused fields are well known and if the sender is not careful about the probability distribution of messages, such techniques can be easily detectable. So, the success of such techniques

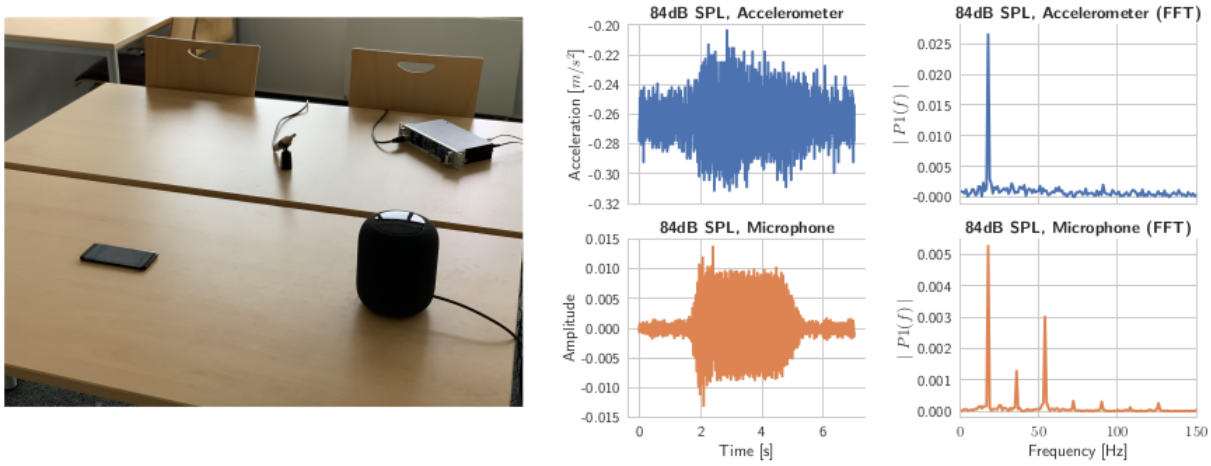


Figure 5: In Matyunin et al. [18], the authors have provided this image to show examples of the experimental setup and the recorded vibrations. They have also provided accelerometer measurements (blue) and sound (orange) that were recorded while playing an 18Hz sine wave at 84dB SPL in 50cm between devices.

may be heavily dependent on the statistical properties of the information being sent rather than the steganographic technique itself.

Matyunin et al. [18] examined the use of sub-bass and infrasonic range acoustic signals in establishing a covert channel of communication using vibrations. The communication channel is established between computers with a typical consumer speaker (that has the common capability of producing low-frequency signals) and mobile devices. Since these signals are not perceivable by humans, such communication will usually go undetected. The authors also showed that the sounds invariably vibrated the speaker itself and the surface on which the speaker has been placed. Such vibrations may go unnoticed by humans but can be recognised and read by a mobile device that is placed on the same surface as the speaker. The mobile device uses its accelerometer sensor for such communication. Using the above method, one may encode data into low-frequency sounds to be played by the speaker. When received by the mobile device's accelerometer, it can decode the vibrations and the data therein. Given that access to the accelerometer does not require any special permissions from the mobile device, such communication may go unnoticed. The authors claimed to have evaluated the setup (as shown in Figure 5) for different speakers and have applied it to several application scenarios.

Editorial Comments

The premise for information hiding in the work by Matyunin et al. [18] is based on vibration on surfaces. The tractability of such physical methods for covert communication are subject to human perception. The robustness of such methods is also subject to the physical properties of the medium - in this case the firmness of the surface.

Coverless Steganography

Cao et al. [4] proposed a coverless steganography method to hide information within *attribute labels of anime characters*. As illustrated in Figure 6, the anime character with embedded hidden information (Figure 6(b)) displays similar image clarity comparable to the original anime (Figure 6(a)), and fulfils the same set of attributes although holding 12 bits of hidden data. The method uses 3 components that are either based on off-the-shelf libraries or techniques proposed by other researchers: Illustration2Vec (used for label extraction from anime characters), Long Short-Term Memory (LSTM) network (used for converting hidden data into binary labels), and Generative Adversarial Network (GAN) (used for generation of anime characters). The stego-image uses an NxN matrix of anime characters, some containing hidden data and some not. This allows flexibility to exchange more or less hidden data between

a. original image **b**. generated image

Figure 6: Original anime character image (a), and generated anime character image (b) according to the coverless steganography method proposed by Cao et al. [4]. Both anime have the same attributes (i.e., long hair, black hair, blue eyes, blush and ribbon) but (b) contains 12-bit hidden information encoded in its attribute labels.

the sender and the receiver, according to their needs. In order to extract the hidden data, both parties obtain a map of the attributes used, and a set of indices for the matrix that contains the hidden data based on an agreed key. The LSTM network was trained with a set of 27,000 high quality anime characters obtained via `http://www.getchu.com/` or Web crawling. The method was evaluated in terms of image quality, hiding capacity, robustness against a number of attacks (e.g., rotation, scaling and different types of noise and filtering), and resistance to a steganalysis model. The authors claimed that results indicate no relevant deterioration of quality between the original and the anime characters with hidden data. In terms of hiding capacity, it outperformed the best four methods found in the literature, reaching up to 60 times better capacity when $N = 8$. Out of those four methods, it delivered better robustness in comparison to two of them although slightly worse than one of the other two. The authors emphasised that their focus was on increasing the hiding capacity while ensuring some level of robustness.

Editorial Comments

Using a variable length matrix of anime characters to hide information seems a great idea since it potentially scales well. The authors claimed a maximum data hiding capacity of 14 bits per anime image when $N = 1$, which is comparable to existing methods, and a hiding capacity of 896 bits per anime image when $N = 8$, which outperforms existing methods. The method, however, seems to be at its early stage. There were details not elaborated enough in the article (e.g., in relation to the shared key). The authors did not make their models available to other researchers. The evaluation provided was, to some extent, superficial (e.g., in regards to the hiding capacity) and lacked consistency (e.g., the hiding capacity was evaluated against four methods while the robustness was compared with three of the four methods used for evaluating the hiding capacity).

Luo et al. [17] presented a method that uses the realisation that an image to be hidden can be split into blocks visually similar to blocks in unmodified, real-time natural images. A set of natural images containing parts of the hidden image would be sent together with location information allowing the receiver to retrieve the relevant blocks and reconstruct the hidden image. The method involves the following main stages: (1) *feature extraction* applied to natural images and the image to be hidden (uses the DenseNet121 convolutional network pre-trained with natural images from ImageNet); (2) *hash sequence code generation* (uses a DCT-based (Discrete Cosine Transform) approach to generate a hash code for blocks – only the first 8-bits are considered); (3) *block matrix generation* (uses an inverted index structure to store blocks' hash code for each natural image); (4) *image hiding* (uses the matrix from stage (3) to determine the similarity between blocks from natural images and blocks from the image to be hidden using Euclidean distance to determine which blocks from natural images are relevant to represent the hidden image); and (5) *hidden image re-*

k	0	1	2	3	4	5	6
min max	6.9	12.9	15.8	16.2	18.4	20.1	19.5
Random	6.9	12.1	12.4	16.2	15.3	15.4	16.2
Last iteration	6.9	12.1	10.5	16.3	–	16.2	17.3

Figure 7: Error probability P_{err} for the MinMax strategy proposed by Bernard et al. [2], compared to two other strategies from the literature, for the first 5 iterations k. Results show that the MinMax strategy outperforms the other, given that a higher P_{err} indicates increased robustness.

construction (uses information from (4) and the set of relevant natural images to piece together the hidden image). The proposed method was evaluated in terms of capacity, performance, accuracy, and security. The authors claimed a higher hiding capacity of their method compared to four other coverless hiding methods. They argued a hiding capacity of 800 bits provided that the image to be hidden is divided into 100 blocks, and considering the 8-bits of hash code used to reconstruct the image. In terms of performance, the authors compared their method to one found in the literature, and their method achieved a faster return time. Accuracy was evaluated with two hidden images in terms of visual similarity between the hidden image at the sender and the receiver ends. Results showed a better accuracy of the proposed method ($\approx 53\%$), compared with the same method from the literature. In terms of security, the authors tested robustness against attacks such as JPEG compression, histogram equalisation, Gaussian filtering, and Gamma correction, and reached similar or slightly better results compared another method from the literature.

Editorial Comments

The article, as the previous one, proposed a method which seems to be in the early stages of development. What was lacking was a concrete example clearly demonstrating what is exchanged between the sender and the receiver to allow the reconstruction of the hidden image. This would have helped to understand better the applicability of the method in practice. The evaluation of the method was also somewhat limited, e.g., only two images were used to validate accuracy.

Steganalysis

Yang et al. [24] designed a solution aiming to address a gap identified in the state-of-the-art of generation-based linguistic steganalysis methods, i.e., they used features from the last layer only. To overcome this weakness, the authors proposed the use of *feature pyramids* which can leverage from all features at every layer. Three steps involved are: *feature extraction* (given a sequence of words, this step generates a sequence of a low-dimensional dense vector for each layer using bidirectional LSTM), *feature fusion* (this step concatenates and flattens all the vectors into a single one-dimension vector with all the features), and *feature classification* (this step uses a model with 2 dense layers with functions Rectified Linear Unit (ReLU) and Sigmoid to classify the given text into "cover" or "stego"). The authors built a dataset of cover text collected from Twitter, Internet Movie Database (IMDB), and `https://www.kaggle.com/snapcrack/all-the-news/data`, containing between 42k and 49k words each, to train their model. They then built a steganalysis dataset with 10,000 stego samples and 10,000 cover samples to evaluate performance of the model, using embedding rates ranging from 1 to 5 bit per word (bpw). Results comparing accuracy, precision and recall against 2 traditional steganalysis algorithms and 2 neural network based steganalysis models were presented; it achieved performance comparable (on average a bit better) to the best competing approach.

Bernard et al. [2] recognised that "steganography is adversarial by design" (i.e., steganography vs. steganalysis), and that steganography schemes aim to minimise distortion on the stego-image compared to the original cover to overcome steganalysis schemes (i.e., "distortion minimisation principle"). Therefore, they generalised the problem, drawing from *Game Theory*, as a MinMax strategy, and derived the distortion function by applying an iterative scheme in k steps. There is an initial embedding of hidden information and then, increasingly, a stronger stego-

image and adversaries are generated. The authors evaluated the approach using the following components: J-Uniward as the embedding algorithm, BossBase as the database containing grayscale JPEG images, XuNet implemented on top of TensorFlow as the steganalysis/classification tool (at step i, a new steganalyser f^i is trained to classify contents from the cover χ and the stego y^i), and ADV-EMB as the attack procedure (used to generate at step i stego contents y^i in respect to the MinMax strategy). A set of 5,000 pairs of cover and stego images were used for training and testing purposes. The Fisher Linear Discriminant classifier was used to compute the minimal total classification error probability (P_{err}). The target of the experiments consisted in evaluating the security of the proposed embedding under different steganalysis schemes, and evolution of the error with the number of iteration steps. The MinMax strategy was compared with the "random strategy" and with the "last iteration strategy". Figure 7 shows P_{err} (%) across different steps k with the MinMax strategy outperforming the others; a higher P_{err} indicates increased robustness.

Editorial Comments

This paper was extended by another article by the same authors published in 2021 [1]. We made the decision to focus on the first, since it lays the foundation of the proposed method; the second article provides further details such as algorithms. The fact that the proposed scheme was presented in detail gives confidence on the results reported. An experienced programmer should have no problem implementing their scheme and trying to reproduce their results.

References

[1] Solène Bernard, Patrick Bas, John Klein, and Tomas Pevny. 2021. Explicit Optimization of min max Steganographic Game. *IEEE Transactions on Information Forensics and Security* 16 (2021), 812–823. https://doi.org/10.1109/TIFS.2020.3021913

[2] Solène Bernard, Tomas Pevny, Patrick Bas, and John Klein. 2019. Exploiting Adversarial Embeddings for Better Steganography. In *2019 ACM Workshop on Information Hiding and Multimedia Security (IHMMSec' 19)*. ACM, 216–221. https://doi.org/10.1145/3335203.3335737

[3] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4. ACM, Article 93, 26 pages. https://doi.org/10.1145/3415164

[4] Yi Cao, Zhili Zhou, Q. M. Jonathan Wu, Chengsheng Yuan, and Xingming Sun. 2020. Coverless Information Hiding Based on the Generation of Anime Characters. *EURASIP Journal on Image and Video Processing* 2020, 36 (2020), 1–15. https://doi.org/10.1186/s13640-020-00524-4

[5] Assunta Cerone, Elham Naghizade, Falk Scholer, Devi Mallal, Russell Skelton, and Damiano Spina. 2020. Watch ' n' Check: Towards a Social Media Monitoring Tool to Assist Fact-Checking Experts. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 607–613. https://doi.org/10.1109/DSAA49011.2020.00085

[6] Ronald Denaux and Jose Manuel Gomez-Perez. 2020. Linked Credibility Reviews for Explainable Misinformation Detection. In *The Semantic Web – ISWC 2020*. Springer, 147–163. https://doi.org/10.1007/978-3-030-62419-4_9

[7] Ronald Denaux, Flavio Merenda, and Jose Manuel. 2020. Towards Crowdsourcing Tasks for Accurate Misinformation Detection. In *Advances in Semantics and Linked Data: Joint Workshop Proceedings from ISWC 2020*. CEUR Workshop Proceedings (CEUR-WS.org), 9. http://ceur-ws.org/Vol-2722/semiform2020-paper-2.pdf

[8] Tang Di, Wang XiaoFeng, Tang Haixu, and Zhang Kehuan. 2021. Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 18. https://www.usenix.org/conference/usenixsecurity21/presentation/tang-di

[9] Pingan Fan, Hong Zhang, Yifan Cai, Pei Xie, and Xianfeng Zhao. 2020. A Robust Video Steganographic Method against Social Networking Transcoding Based on Steganographic Side Channel. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*. ACM, 127–137. https://doi.org/10.1145/3369412.3395066

[10] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. 2020. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 1605–1622. https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[11] Naeemul Hassan, Mohammad Yousuf, Md Mahfuzul Haque, Javier A. Suarez Rivas, and Md Khadimul Islam. 2019. Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-Checking. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 1001–1006. https://doi.org/10.1145/3308560.3316734

[12] Mario Hildebrandt, Kevin Lamshöft, Jana Dittmann, Tom Neubert, and Claus Vielhauer. 2020. Information Hiding in Industrial Control Systems: An OPC UA Based Supply Chain Attack and Its Detection. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*. ACM, 115–120. https://doi.org/10.1145/3369412.3395068

[13] Hai Huang, Jiaming Mu, Neil Zhenqiang Gong, Qi Li, Bin Liu, and Mingwei Xu. 2021. Data Poisoning Attacks to Deep Learning Based Recommender Systems. In *Proceedings 2021 Network and Distributed System Security Symposium*. 17. https://doi.org/10.14722/ndss.2021.24525

[14] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. 2019. ABS: Scanning Neural Networks for Back-Doors by Artificial Brain Stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1265–1282. https://doi.org/10.1145/3319535.3363216

[15] Giulio Lovisotto, Simon Eberz, and Ivan Martinovic. 2020. Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 184–197. https://doi.org/10.1109/EuroSP48549.2020.00020

[16] Jiazhong Lu, Weisha Zhang, Ziye Deng, Shibin Zhang, Yan Chang, and Xiaolei Liu. 2021. Research on Information Steganography Based on Network Data Stream. *Neural Computing and Applications* 33, 3 (2021), 851–866. https://doi.org/10.1007/s00521-020-05260-4

[17] Yuanjing Luo, Jiaohua Qin, Xuyu Xiang, Yun Tan, Qiang Liu, and Lingyun Xiang. 2020. Coverless Real-Time Image Information Hiding Based on Image Block Matching and Dense Convolutional Network. *Journal of Real-Time Image Processing* 17 (2020), 125–135. https://doi.org/10.1007/s11554-019-00917-3

[18] Nikolay Matyunin, Yujue Wang, and Stefan Katzenbeisser. 2019. Vibrational Covert Channels Using Low-Frequency Acoustic Signals. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*. ACM, 31–36. https://doi.org/10.1145/3335203.3335712

[19] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. 2020. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 85–99. https://doi.org/10.1145/3372297.3417253

[20] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. 2019. Towards Fact-Checking through Crowdsourcing. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 494–499. https://doi.org/10.1109/CSCWD.2019.8791903

[21] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. 2020. Hidden Trigger Backdoor Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (2020), 11957–11965. https://doi.org/10.1609/aaai.v34i07.6871

[22] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 707–723. https://doi.org/10.1109/SP.2019.00031

[23] Antonia Wild, Andrei Ciortea, and Simon Mayer. 2020. Designing Social Machines for Tackling Online Disinformation. In *Companion Proceedings of the Web Conference 2020*. ACM, 650–654. https://doi.org/10.1145/3366424.3385770

[24] Hao Yang, YongJian Bao, Zhongliang Yang, Sheng Liu, Yongfeng Huang, and Saimei Jiao. 2020. Linguistic Steganalysis via Densely Connected LSTM with Feature Pyramid. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*. ACM, 5–10. https://doi.org/10.1145/3369412.3395067