IDrISS: Intrusion Detection for IT Systems Security
Toward a semantic modelling of side-channel signals

Ngolè Mboula
Institut LIST
CEA, Université Paris-Saclay
Gif-sur-Yvette, France
fred-maurice.ngole-mboula@cea.fr

Erwan Nogues
Univ. Rennes, INSA Rennes, IETR - UMR CNRS 6164
DGA MI
Rennes, France
erwan.nogues@insa-rennes.fr

***Abstract*—This paper proposes a novel approach called IDrISS that exploits electromagnetic (EM) side-channel signals to design non-protocol based Intrusion Detection System (IDS). EM emanations side-channels are captured on power lines of an infrastructure. They are used to identify the presence of any type of electronic devices onto a physical network. IDrISS can learn the structure of the EM unintentional emanations of the legit devices composing the infrastructure as a reference profile. In a second step, it records and analyses the current emanations to compare and detect any kind of unwanted emanations. IDrISS is used as a Intrusion Detection System (IDS) that can trig an alarm as soon as a intrusion is detected. The results show that intrusion can be detected in various scenarios whatever the activity of the legit computers of the network. Furthermore, the capture device used is based on inexpensive off-the-shelf components that makes the deployment onto real network easy.**

***Index Terms*—Side-channel signals, intrusion detection, dictionary learning, sparsity, recurrent neural networks**

I. Introduction

A. Context and related works

Detecting intruders on a network is part of the analysis of Information Systems Security (INFOSEC). The goals of the intruder can be multiple: interception and listening of network traffic and exchanged data, commands injection, etc. Existing solutions for detecting intruders on a network are mainly based on the network's traffic analysis in order to detect any form of anomaly. Indeed, known techniques recover all network traces in order to filter legitimate traffic and detect traces that the intruder would generate (see for example [1], [2]). With the advent of wireless networks, intruders can now seek to integrate directly into wireless traffic [3]. However, the proposed approaches assume the analysis of the targeted protocol. Therefore, an intruder complying with the protocol cannot be detected by the system as being an intruder, especially if he is listening passively. Another approach could be to use current consumption analysis. This type of approach would be based on the electricity power consumption of the intrude device. However the intruder might have a consumption much lower than that of the network considered which would make it undetectable by current analysis.

Unlike network's analysis based approaches, our method does not rely on any a priori knowledge on the network. It also does not rely on electricity power consumption but instead on Electro Magnetic (EM) side-channel signals which contain much more information. In [4], [5], the EM signal is used to detect abnormal behavior on a chip like, for example, the execution of a malicious program. The proposed method is a local analysis with a EM probe . Moreover, whereas the former detects the intruder's activity effect on a single device's EM side-channel signal, we aim at detecting the intruder EM emanations in an aggregate of possibly several devices emanations. Thus, our method can find applications in many fields such as networked (or isolated) computer systems, control and data acquisition systems, the Internet of Things, wired and wireless networks. We detail this approach in section II, present some numerical experiments in section III and conclude in section IV.

B. Notations

We adopt the following notation conventions :

- lower case and bold letters are used for vectors;
- upper case and bold letters are used for matrices;
- by default, vectors are represented as columns;
- Greek letters are used for hyper-parameters;
- non-bold letters are elements in a matrix or in a vector.

II. Proposed method

A. Side-channel signals sensing

All electronic devices produce EM emanations that not only interfere with radio devices but also compromise the data handled by the information system. A third party may perform a side-channel analysis and recover the original information, hence compromising the system privacy. While pioneering work of the domain focused on analog signals [6], recent studies extend the eavesdropping exploit using an EM side-channel attack to digital signals and embedded circuits [7]. Side channels are used to retrieve the information completely. However, even though the information cannot be retrieved

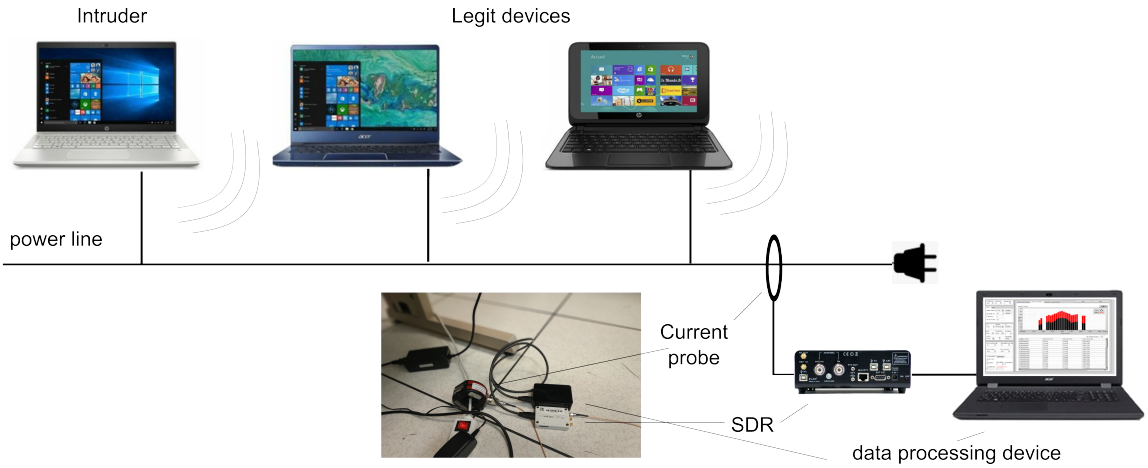

Fig. 1. Block diagram of the proposed method: the network is made of legit devices and an intruder. All generate EM emanations that couple onto the power lines. IDrISS captures the EM signal with a current probe and SDR system that sends the raw IQ data to a processing data device.

entirely, the EM emanation can embed another type of information such as the device type (computer, monitors, etc.) or activity cycle (sleep mode, idle, working, etc.)

We consider a generic scenario involving n legit devices and one unknown device (see Fig. 1). The EM emanations of both the legit and the unknown devices are sensed though an acquisition system which consists in a receiver and an analog digital converter (ADC). In our scenario, we consider a current probe to recover the EM by conduction. The receiver is a SDR device with its associated filters and ADC. This acquisition system senses the signals that are used for training and monitoring.

B. Sparse modelling of side-channel signals

The first step of the learning phase is to detect and extract from the signals received from the capture system, segments corresponding to periods of *activity* of equipment whose EM emanation were measured. On these segments, the signals are expected to exhibit particular morphologies. These signals are complex-valued. Although the phase is certainly informative, we focus the analysis on the amplitude signal. The previously mentioned segments are then the continuous and non-extendable periods over which the amplitude takes values greater than a given threshold. This threshold constitutes a sensitive parameter, the choice of which will be discussed later. A typical example of amplitude signal is given in Fig. 2.

Activity patterns are extracted, registered w.r.t. their time-wise barycenters as illustrated in Fig 3 and zero-padded so that they have the same length. These patterns are extracted for the all legit equipments and stacked into a matrix $\mathbf{P}$ of size $n \times m$, n being the number of activity patterns extracted and m their length. We factorize $\mathbf{P}$ into two sparse non-negative matrices: $\mathbf{P} = \mathbf{WD}$ where $\mathbf{W}$ and $\mathbf{D}$ are $n \times p$ and $p \times m$ matrices. We use the method presented in [8] via the python toolbox *NIMFA*[1]. This amounts to decomposing the activity patterns into simpler shared sub-features which are the are the p rows

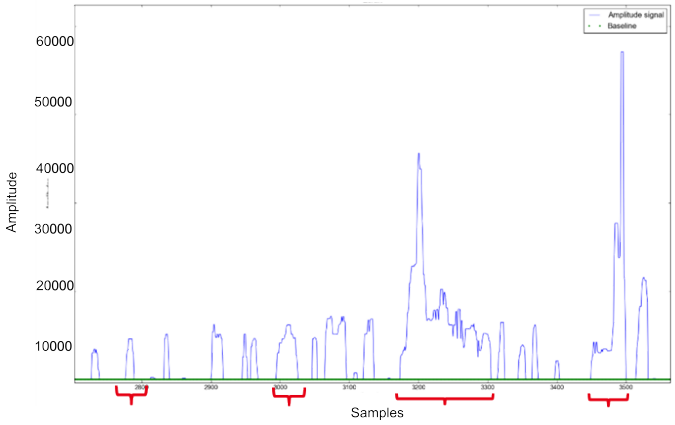

Fig. 2. Amplitude signal: the braces indicate some activity segments.

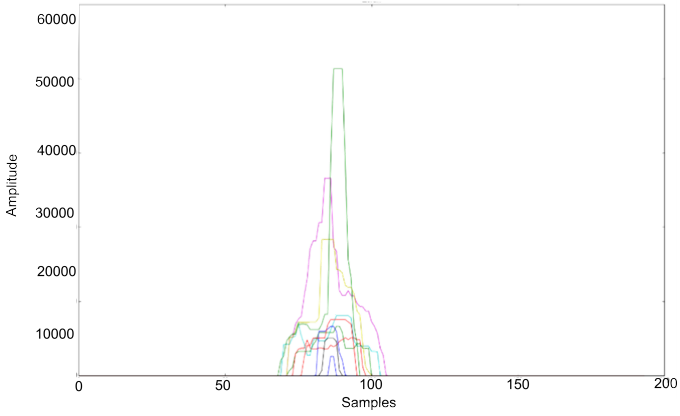

Fig. 3. Registered activity patterns.

[1]http://nimfa.biolab.si/index.html, accessed in Febuary 2020

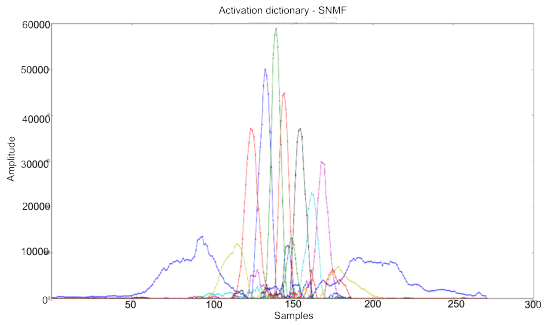

Fig. 4. Elementary activity features.

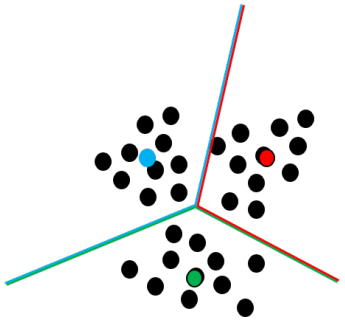

Fig. 5. Partitionning.

of the matrix $\mathbf{D}$. We can see examples of these sub-features in Fig. 4.

In the following, we refer to $\mathbf{D}$ as the dictionary and to $\mathbf{D}$'s rows as the atoms. We note these atoms $\mathbf{d}_1, \cdots, \mathbf{d}_p$. By design, these atoms can be used sparsely to reconstruct activity patterns. So for any given amplitude signal $\mathbf{x} = [x_1, \cdots, x_T]$ we can compute a sparse convolutive representation of $\mathbf{x}$ as follows:

$$\mathbf{x} = \sum_{i=1}^{p} \mathbf{h}_i * \mathbf{d}_i, \tag{1}$$

$*$ denoting the convolution product and $\mathbf{h}_i$ being a sparse non-negative weights vector of length T associated with the i^{th}. These weights determine the presence and the magnitude of the different atoms at different times in the magnitude signal.

We estimate them by solving an optimisation problem of the form:

$$\min_{\mathbf{h}_1, \cdots, \mathbf{h}_p} \frac{1}{2} \|\mathbf{x} - \sum_{i=1}^{p} \mathbf{h}_i * \mathbf{d}_i\|_2^2 + \sum_{i=1}^{T} w_i \|[\mathbf{h}_1[i], \cdots, \mathbf{h}_p[i]]\|_2, \quad \text{s.t. } \mathbf{h}_i \geq 0, \tag{2}$$

The weights $w_1, \cdots, w_T$ are hyper-parameters set in order to mitigate the l_2 norm induced bias, according to the strategy presented in [9]. Once this estimate has been made, we now have at each time i a weight vector $\mathbf{x}_i = [\mathbf{h}_1[i], \cdots, \mathbf{h}_p[i]]$ which characterizes the contribution of each atom to the magnitude of the amplitude signal at that time. The vector-valued signal $[\mathbf{x}_1, \cdots, \mathbf{x}_T]$ constitutes a new, more structured representation of the initial signal which will be used in the following for sequential modeling.

C. Sequential modeling of legit devices activity

At this stage, we have a new vector-valued representation of the amplitude signal. The first step is to quantize the new representation space in order to represent the amplitude signal using a finite vocabulary. In other words, we build a partition of this new space as illustrated in Fig. 5 We recall that we have at this stage a sparse vector-valued representation of the training amplitude signal (see Fig. 6). The previously mentioned partition should only be based on the vector values of significant l_2 norms, in other words, which are above background noise.

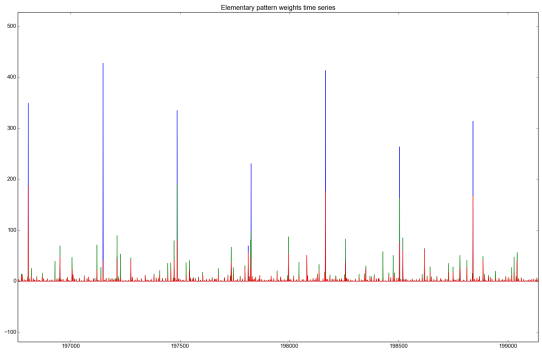

Fig. 6. Sparse vector-valued amplitude signal representation. Each color corresponds to one coordinate of the vector value at each time

These vector values are thus selected based on a threshold on their l_2 norms. This threshold should be high enough for the noise related vectors to be filtered out. It can be set according to l_2 norms values histogram. The selected vectors are first standardized i.e. the means and standard deviations of each of their coordinates are set to 0 and 1 respectively. The standardized vectors are then clustered using the k-means algorithm. The number of clusters is another important parameter that will be discussed later. The partitions are then defined as being the regions of space closest to each of the centroids of the formed clusters. In a second step, we return to non-standardized vectors and before thresholding. We make a new selection from a lower threshold than the first used in order to retain more complete, although potentially more noisy, information of the amplitude signal processed. Then we recenter and rescale the set of selected vectors using the means and standard deviation previously calculated. We extend each of these vector by adding to it the number of time steps separating it from the next selected vector and assign each vector a partition number.

Let $\mathbf{u}_i$ denote the vector corresponding to the time step t_i and l_i his partition number. Each partition can be interpreted as a particular state of the legit devices activity. In order to capture legit devices activities regularity, we train a Long-Short Term Memory (LSTM) to predict the state or partition number l_k at time t_k, based on the sequence of $\mathbf{u}_i$ observed

up to time t_{k-1}. This model is then used to detect anomalies as explained in the next subsection.

D. Intrusion analysis

The previous steps are performed using a reference training signal. Given a new amplitude signal, we compute its representation into the partitioned space previously built. It results into a sequence of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_m$, and the corresponding partitions numbers $l_{\mathbf{v}_1}, \cdots, l_{\mathbf{v}_m}$. Let p_i denote the probability estimated by our sequential model that the vector $\mathbf{v}_1$ belongs to the cluster number $l_{\mathbf{v}_i}$ conditionally to the preceding vectors of the sequence:

$$p_i = pr(l_{\mathbf{v}_i} | \mathbf{v}_{i-1}, \cdots, \mathbf{v}_1). \quad (3)$$

The sequence $\mathbf{p} = p_1, \cdots, p_m$ characterizes the compatibility of the new amplitude signal with the model learn:

- values close to 1 indicate a strong fit with the model;
- conversely, values close to 0 indicate outliers from the model point of view.

Hence, Thus, a consistent drop in the values on the p_is would be an indicator of abnormal activity, including the activation of unknown equipment.

III. Numerical experiments

A. Experimental Setup

The experimental setup is defined as follows: a current probe is installed the power strip where all the legit computers are plugged in as depicted in Fig. 1. A Radio Frequency (RF) power amplifier is inserted after the current clamp. The interception system is composed as follows the SDR device is an Ettus B205 mini receiving with a 20 MHz bandwidth to recover the emanations with a fine granularity [10]. The recovered signal is a digitized radio signal of the form 16-bit signed IQ samples. For the intrusion diagnosis, we consider a situation in which there are two legitimate computers (different from those used for dictionary learning) and potentially an *unknown* computer. Traces from the three computers are recorded separately and mixed offline following two scenarios. On the one hand, the legit computers complex traces are simply summed up. On the other hand, the traces are zeroed on different random slots before summation to emulate switching on and off. In both configurations, we use an uncorrupted segment of the synthetic trace to learn a sequential model, and we use it to analyze part of the trace to which we locally added the *unknown* computer's trace (see Fig. 7 and Fig. 8).

The training amplitude signal corresponds roughly to 1.6 seconds. The corrupted amplitude signal is one third of the training amplitude signal.

B. Results

a) Dictionary learning: The dictionary $\mathbf{D}$ of section II-B is computed based on complex traces recorded on the power cables of two computers and their screens using a current clamp as previously described. Activity segments are extracted using a threshold manually set to 40, based on the histogram of the values of the training amplitude signal. Thus the choice of

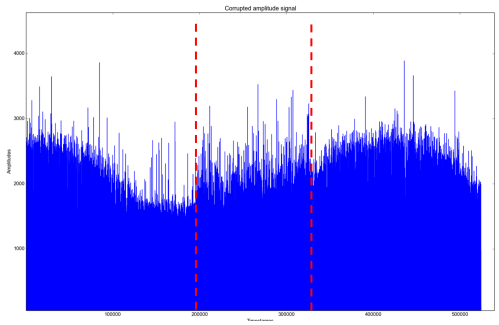

Fig. 7. Corrupted amplitude signal in the first scenario; the intruder is active in the time slot between the dashed lines.

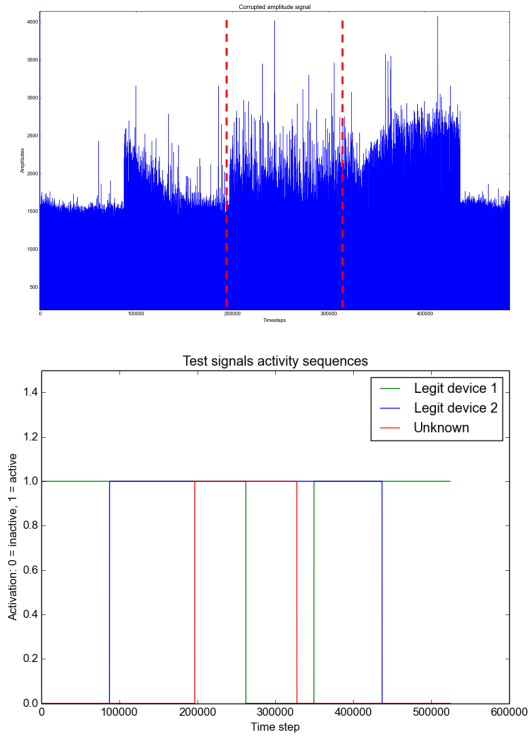

Fig. 8. Corrupted amplitude signal in the second scenario on the top; the intruder is active in the time slot between the dashed lines; on the bottom the periods of activity of each of the equipment are indicated.

this parameter can easily be automated. The number of atoms (parameter p) is chosen so that the matrix factorization error is negligible. We set it to 20. Actually, less atoms could have been computed using a different dictionary learning method (see for instance [11]), as we can see in Fig. 4 that several atoms are practically identical up to a shift.

b) Sequential modeling: In order to partition the new representation space previously described, we set the number of clusters to 100. The more clusters there are, the smaller the partitions and the more sensitive the detection is to small disturbances due to the activity of an intruder. However, a high number of clusters requires more data for learning the sequential model. There is therefore a compromise to be found,

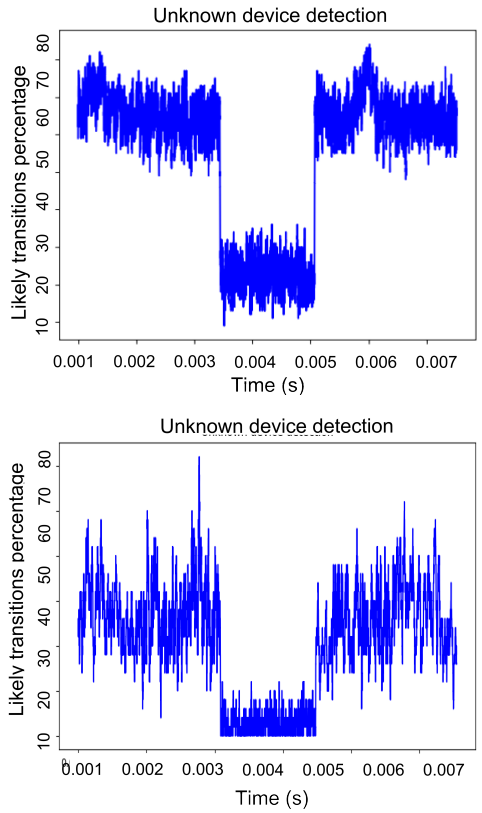

Fig. 9. Intrusion detection in the first scenario on the top and the second scenario on the bottom. The activation of the unknown equipment causes a detectable rupture thanks to the learned sequential model.

with a fixed amount of data. This hyperparameter has to be fine tuned accordingly, based on the prediction accuracy of the sequential model on the training data. The vector values are selected using a threshold of 100 for clustering and a threshold of 10 for labelling and sequential model learning. These thresholds are chosen manually based on the histogram of the time step wise l_2 norms of the vector-valued representation of the training amplitude signal obtained by solving problem 2. The later thresholding yields an average compression of roughly 1/10 of the original amplitude signals.

We train a two layers LSTM with a hidden and output layers size of 3. We get a prediction accuracy of more than 80% on a test uncorrupted signal, which confirms that the underlying regularities present in the amplitude signal have been well preserved. This accuracy is empirically stable, regardless of the non deterministic results of the k-means clustering.

c) Intrusion detection: Following the methodology described in Section , we calculate the signal $\mathbf{p}$ for the corrupted traces in the two scenarios. We then calculate in a sliding window of size 100 on the signal $\mathbf{p}$, the percentage of values greater than 0.5. One can see the results obtained in Fig. 9.

In both cases, activation of the unknown equipment results in a detectable increase in the number of implausible transitions in the analyzed sequence. However we observe that the break is less neat in the second scenario. This is simply due to greater statistical variability in the data in this case which makes the sequential model harder to train. In this case, therefore, more training data should be used. Besides, these results have to be consolidated with a thorough false detection rate analysis.

IV. Conclusion

We presented an intrusion detection method called IDrISS based on side-channel signal analysis. From these signals, recorded only for legit equipment, we learn a vocabulary and an operating syntax using a recurrent neural network. The learnt model then allows us to detect deviations from the expected operation, indicating the activity of an unknown equipment. We evaluated this methodology on realistic data. The capture device used is based on inexpensive off-the-shelf components. In a configuration with two legit and one unknown equipment and under two different scenarios, we obtained convincing detection results. The continuation of this work will focus on adapting the proposed methodology to the monitoring of an arbitrary number of legit equipment.

References

[1] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2027–2051, thirdquarter 2016.

[2] Zouheir Trabelsi and Khaled Shuaib, "Nis04-4: Man in the middle intrusion detection," *IEEE Globecom 2006*, pp. 1–6, 2006.

[3] Ahmedur Rahman, C. I. Ezeife, and A. K. Aggarwal, "Wifi miner: An online apriori-infrequent based wireless intrusion system," in *Knowledge Discovery from Sensor Data*, Mohamed Medhat Gaber, Ranga Raju Vatsavai, Olufemi A. Omitaomu, João Gama, Nitesh V. Chawla, and Auroop R. Ganguly, Eds., Berlin, Heidelberg, 2010, pp. 76–93, Springer Berlin Heidelberg.

[4] H. A. Khan, N. Sehatbakhsh, L. N. Nguyen, R. L. Callan, A. Yeredor, M. Prvulovic, and A. Zajic, "Idea: Intrusion detection through electromagnetic-signal analysis for critical embedded and cyber-physical systems," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2019.

[5] Nader Sehatbakhsh, Alireza Nazari, Monjur Alam, Frank Werner, Yuanda Zhu, Alenka Zajic, and Milos Prvulovic, "Remote: Robust external malware detection framework by using electromagnetic signals," *IEEE Transactions on Computers*, vol. 69, no. 3, pp. 312–326, 2019.

[6] W. Van Eck, "Electromagnetic radiation from video display units: An eavesdropping risk?," *Computers & Security*, vol. 4, no. 4, pp. 269–286, 1985.

[7] Florian Lemarchand, Cyril Marlin, Florent Montreuil, Erwan Nogues, and Maxime Pelcat, "Electro-magnetic side-channel attack through learned denoising and classification," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[8] Hyunsoo Kim and Haesun Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 05 2007.

[9] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec 2008.

[10] M. G. Kuhn, "Compromising Emanations of LCD TV Sets," *IEEE Transactions on Electromagnetic Compatibility*, vol. 55, no. 3, pp. 564–570, 2013.

[11] Morgan A. Schmitz, Matthieu. Heitz, Nicolas. Bonneel, Fred. Ngolè, David. Coeurjolly, Marco. Cuturi, Gabriel. Peyré, and Jean-Luc. Starck, "Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 643–678, 2018.