A novel embedding-based framework improving the User and Entity Behavior Analysis

Thomas Anglade [1] and Christophe Denis [2] and Thierry Berthier [3]

[1] Data Scientist, iTrust, tanglade@itrust.fr
[2] Sorbonne University, LIP6, Paris, France, christophe.denis@lip6.fr
[3] CREC Saint-Cyr & Université de Limoges, thier.berthier@orange.fr

Abstract. Over the last few years, the number and the variety of cyber-attacks have been constantly growing. The landscape of cyber-attacks has become extremely large (DoS, DDoS, phishing, C&C, botnets, malwares, ransomwares, etc.). Today, UEBA (User and Entity Behavior Analysis[1]) is the best solution that companies need to use to adapt to these changes. Using UEBA, companies do not track security events or monitor devices; instead they track all the users and entities in the system. They use machine learning algorithms and statistical analyses to know when there is a deviation from established patterns.
This paper offers a novel embedding-based framework that facilitate UEBA by projecting sparse and unstructured log data into a new mathematical space in which numerous behavior trends and changes can be analyzed in a simpler and more visual way than using typical deep learning algorithms. We show that in this space, advanced cyber-attacks can be detected through a variation analysis of the fitted 2D-kernel density. The last part of the paper deals with the validation and the explanation of prediction obtained by black box Machine Learning methods. Indeed, the operational benefit of using Machine Learning methods is recognized but is hampered by the lack of understanding of their mechanisms, at the origin of operational, legal and ethical operational problems. This is largely dependent on the ability of engineers, decision-makers and users to understand the meaning and the properties of the results produced by these tools.

Keywords: cybersecurity, UEBA, Machine Learning, Explainable AI.

1 Introduction

Over the last few years, the number and the variety of cyber-attacks have been constantly growing [1]. The landscape of cyber-attacks has become extremely large (DoS, DDoS, phishing, C&C, botnets, malwares, ransomwares, etc.). In the same time, information systems have been quickly evolving IT perimeters used to be well-defined and they changed to become very porous and difficult to manage and oversee [2]. This is due to structural changes in the workplace, like remote work, BYOD, extensive use of leasing contracts for IT hardware, or even shared workplaces. As a result, it is now impossible to block all the cyber-attacks using old tools (web gateways, firewall, intrusion prevention tools, etc.). *Attackers will manage to penetrate the networks sooner or later.*

[1] This concept has been introduced by Gartner in 2015 : https://en.wikipedia.org/wiki/User_behavior_analytics

Today, UEBA (User and Entity Behavior Analysis) is the number 1 solution that companies need to use to adapt to these changes. Using UEBA, companies do not track security events or monitor devices; instead they track all the users and entities in the system. They use machine learning algorithms and statistical analyses to know when there is a deviation from established patterns. When the behavior of any entity in the system changes significantly (change is defined here using advanced mathematical and statistical tools and algorithms), an alert is raised and further investigated by security analysts. Artificial intelligence algorithms have proven to be a very performant way to learn normal entity behaviors using a variety of data sources, mainly logs (FW, AD, proxy, event logs, etc.) and accurately detect deviations from normality to raise alerts. This has led to the proliferation of machine learning and deep learning algorithms, mainly ranging among these categories:

- Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) for temporal behavior monitoring issued from NLP [3][4]
- Autoencoders for anomalous user behaviors and intrusion detection [5]
- Generative Adversarial Networks (GAN) to create synthetic data similar to input data in order to improve algorithms performance [6]
- Geometric deep learning and embedding methods in order to learn and predict the behavior of complex data structures like graphs and hypergraphs [7]

These algorithms have proven to be very effective for well-defined categories of attacks, but some limitations remain:

- the deep learning algorithms need to be fine-tuned for each specific type of attack, which is extremely time and resource-consuming
- the (deep learning) algorithms do not tend to be tuned for a specific task and will not necessarily generalize well to new types of attacks and complex attacks like APT
- the algorithms are often “black boxes”. As a result, mathematicians and cybersecurity analysts struggle to communicate about the results and to integrate expert feedbacks in the models

2 A novel embedding-based framework

In order to overcome these issues, we propose a novel embedding-based framework that facilitate UEBA by projecting sparse and unstructured log data into a new mathematical space in which numerous behavior trends and changes can be analyzed in a simpler and more visual way than using typical deep learning algorithms. The outcome of our method is a framework in which all the entities have time-evolving geometric coordinates, reflecting the nature and the quantity of the interactions observed between these entities. This framework is used as a "dialog box" for complex attack detections: ethical hackers are able to simulate attacks and understand how they translate in this space thanks to the help of data scientists. We use "traditional" machine learning algorithms (density-based methods, clustering and neighbor analysis) to track how complex attack patterns translate in this mathematical space, constructed through embedding methods.

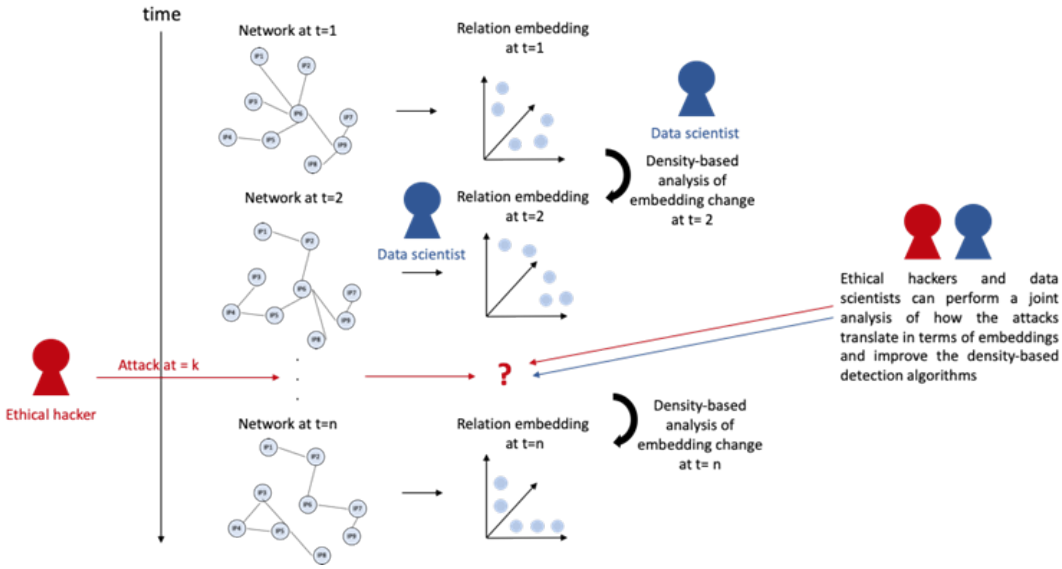


Fig. 1. Overview of our embedding-based detection framework

Embedding is one of the successful uses of deep learning in the last years. The method aims at representing discrete variables and data concepts as continuous vectors. This method works by mapping similar values close to each other in the embedding space, thus revealing the intrinsic properties of the categorical variables. It has been popularized by the great success of the Word2Vec algorithm, developed by Tomas Mikolov at Google in 2013 [8]. This algorithm extends the skip-gram model and allows to capture a large number of precise syntactic and semantic word relationships. It contributes to mitigate the "curse of dimensionality" by representing variables of interest in a denser space. The use of embedding techniques for cybersecurity is at its beginning. Research has been carried regarding the construction of knowledge graphs [9].

Our method processes IP-IP communication graphs data through an extension of the node2vec algorithm. We create sliding windows, and for each window the algorithm:

- Encodes each IP behavior in a M1 multidimensional vector
- Encodes each IP-IP relation in a M2 multidimensional vector (with M2 >= M1)
- Projects the IP and the relations in a 2-dimensional space allowing to perform density analysis in a visual way.

The proposed method is a combination of node2vec and t-SNE algorithms and will be extended in the future using autoencoder models. We refer to the work of [10] to define desired properties for the embeddings:

- temporal smoothness: IP changes their latent positions gradually over time
- network embedding: if a couple (IP1; IP2) interacts a lot in the network, the distance between Embedding_IP1 and Embedding_IP2 will be small in the embedded space (and vice versa)
- latent homophily: IPs who are close to each other in latent space interact with one another more frequently than two faraway members

We extend the latent homophily property to add a new constraint:

- latent relationship homophily: if the *nature* of the relation (IP1; IP2) is close to (IP3; IP4), their embeddings should be close. This should be temporally coherent.

The novelty of our method consists in the fact that our algorithm takes into account variables describing the *nature* of the relation between two IPs: port of communications, volume of data exchanged, duration of the connection. All these pieces of information are processed to compute the embedding. To our knowledge, this type of industry-wide research has never been done before.

3 Details and results

3.1 Description of data and attacks

Our model has been tested using 17 days of firewall IP communication data (19GB) belonging to a medium-sized company (2 days on the 15th and 16th April 2019 + 15 days from the 2nd May 2019 to the 16th of May 2019). The dataset contains 80 private IPs and 9000 public IPs in the network. We use raw data with the following fields (FortiGate firewall):

- Timestamp
- Source IP
- Destination IP
- Port of communication
- Number of bytes & packets sent
- Duration

Different types of attacks have been run by ethical hackers:

- ping scan *(source: 192.168.120.7 victim: 192.168.140.0/24 UTC time: April 15Th 15h45)*
- TCP port scan (*source: 192.168.120.7 victim: 192.168.140.109 UTC time: April 15th 15h25*)
- port scans with speed lowering (*source: 192.168.120.7 victim: 192.168.140.109 UTC time: from April 15th 13h56 to April 16th 08h07*)
- botnet C&C and data exfiltration through the DNS (*source: 192.168.120.7 DNS: 192.168.140.1 UTC time: from May 7th 12h53 to May 7th 13h24*)
- botnet C&C and direct data exfiltration (*source: 192.168.120.7 DNS: 192.168.140.1 UTC time: from May 7th 14h30 to May 7th 15h02*)

3.2 Details of the method:

The purpose of the method is to find a 2-dimensionnal embedding per IP address per hour. This will enable us to analyze the situation of the network and its temporal evolution. We aim at detecting events and link them to the attacks, by analyzing:

- clusters of IP addresses
- trajectories of IP addresses

Out method combines node2vec and T-SNE algorithms. We apply node2vec to determine 128 (or 64) dimensional embeddings. We then project these embeddings in a 2-dimensionnel space for visual analysis. The node2vec algorithm uses random walks to

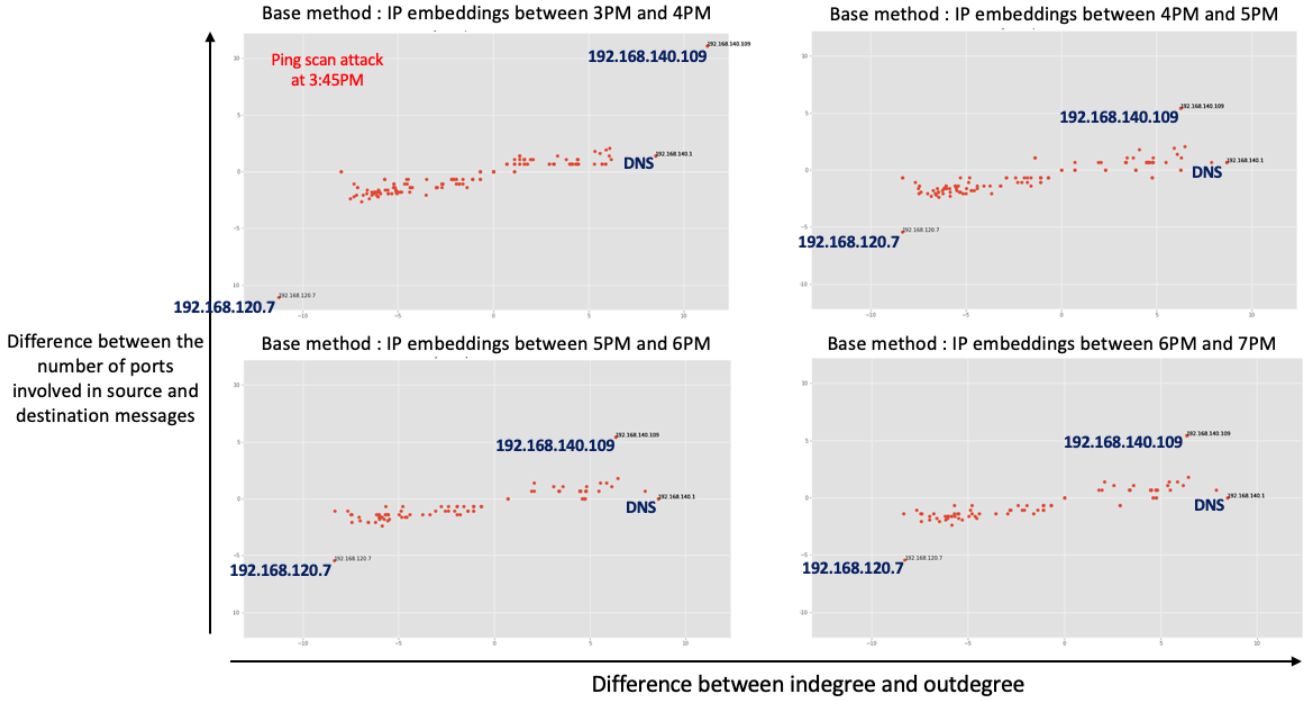
determine each node's embedding. The choice of the parameters has to be consistent with the types of cyber-attacks we aim to detect. For instance, when detecting "*botnet C&C and data exfiltration through the DNS*", the walks:

- are weighted by the number of bytes exchanged between nodes
- have a maximum length of 4 (corresponding to the types of exfiltration patterns and rebounds that we are looking for).

In order to benchmark our method, we define a baseline representation technique: 1 hour of an IP address activity is encoded in a 2-dimensional space using the following coordinates:

- X-axis: Number of source messages (indegree) – Number of destination messages (outdegree)
- Y-axis: 2 possibilities:
    - Number of ports linked to source messages – Number of ports linked to destination messages (with logarithmic transformation)
    - Amount of data out of the network through port 53 (DNS port)

This representation allows us to create IP clusters and determine receiving / emitting IP addresses. We refer to it as our "*base method*" in the rest of this article.

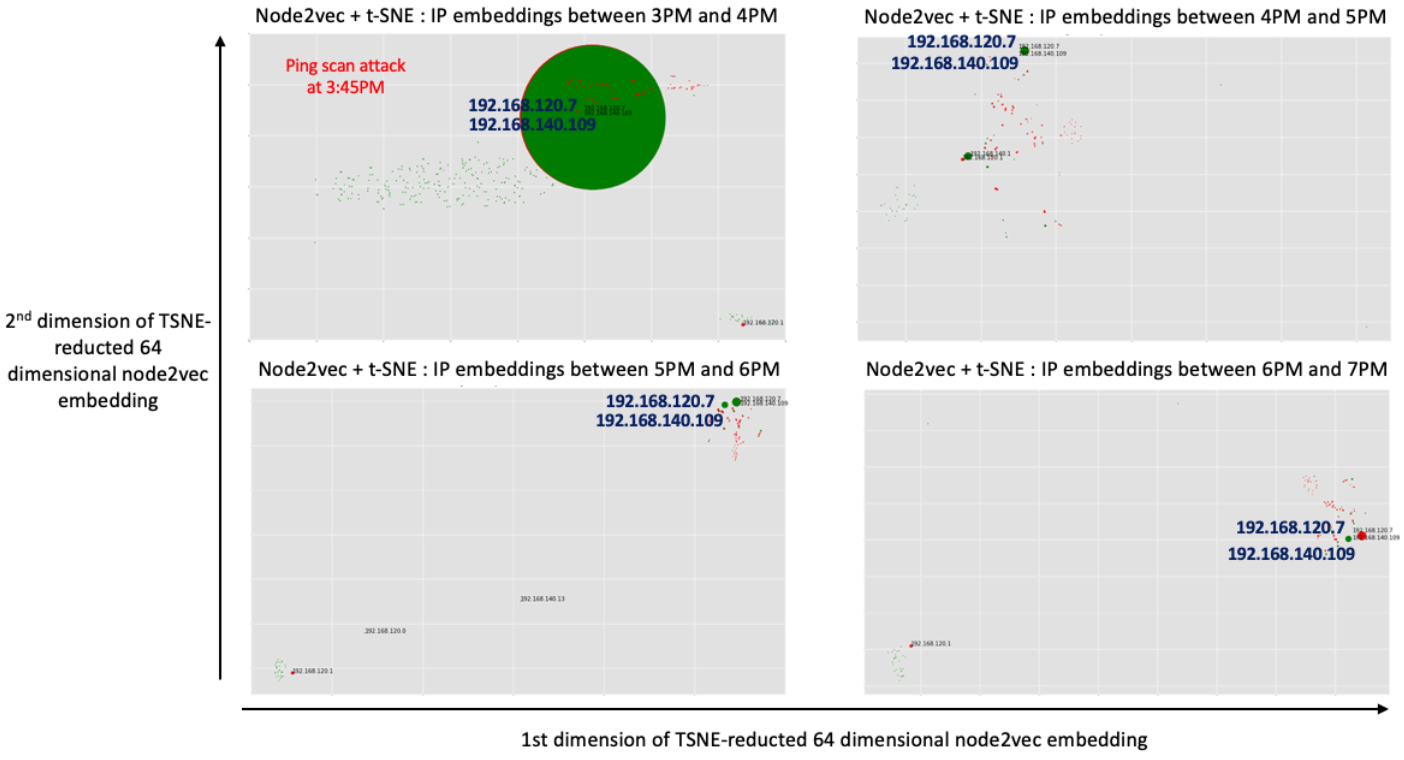


3.3 Visual results and interpretation:

3.3.1 "Network discovery" attacks :

Below the results for network discovery attacks occurring on April 15th and April 16th, for the base method:
On the top left picture IP 192.168.120.7 and IP 192.168.140.109 have very extreme coordinates. The attacking IP is at the bottom left of this picture (IP addresses sending a lot of messages and using a lot of source ports). The victim IP is at the top right (IP addresses receiving a lot of messages and using a lot of destination ports). The overall coordinate structure is consistent across the 4 pictures, allowing to analyze IP trajectories over time. Nonetheless, we would like to see 192.168.120.7 and 192.168.140.109 very close as they are highly connected during the attack.

Below the results for our embedding method: (the number of ports involved in communications for each IP is proportional to the size of the dot. Emitting IPs in red and receiving IPs in green)



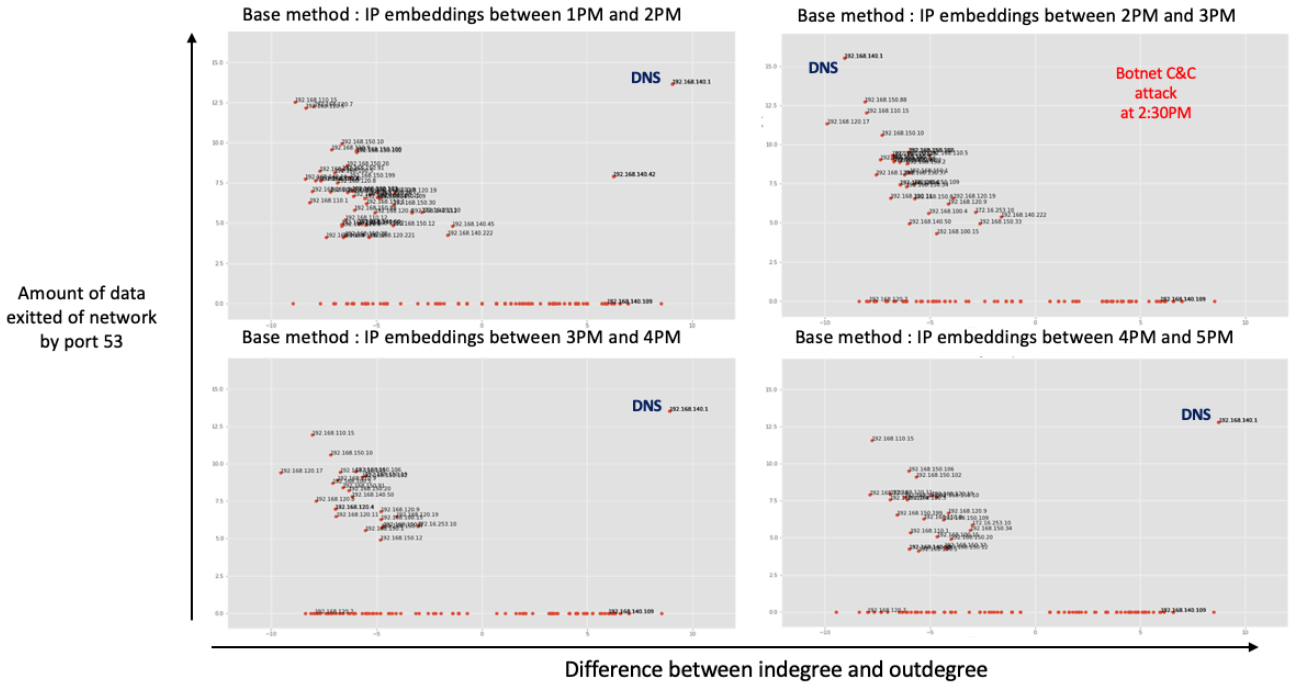
We can see the benefit of using consistent embeddings:

- Attacking and victim IP addresses are very close to each other in the embedding space
- IP addresses are clustered in "communication groups" (IP 192.168.120.1 is the most important dot in a cluster of 20 IP addresses. This pattern is time-consistent)
- We can follow the temporal evolution of the network
- We have a clear visualization of the attack.

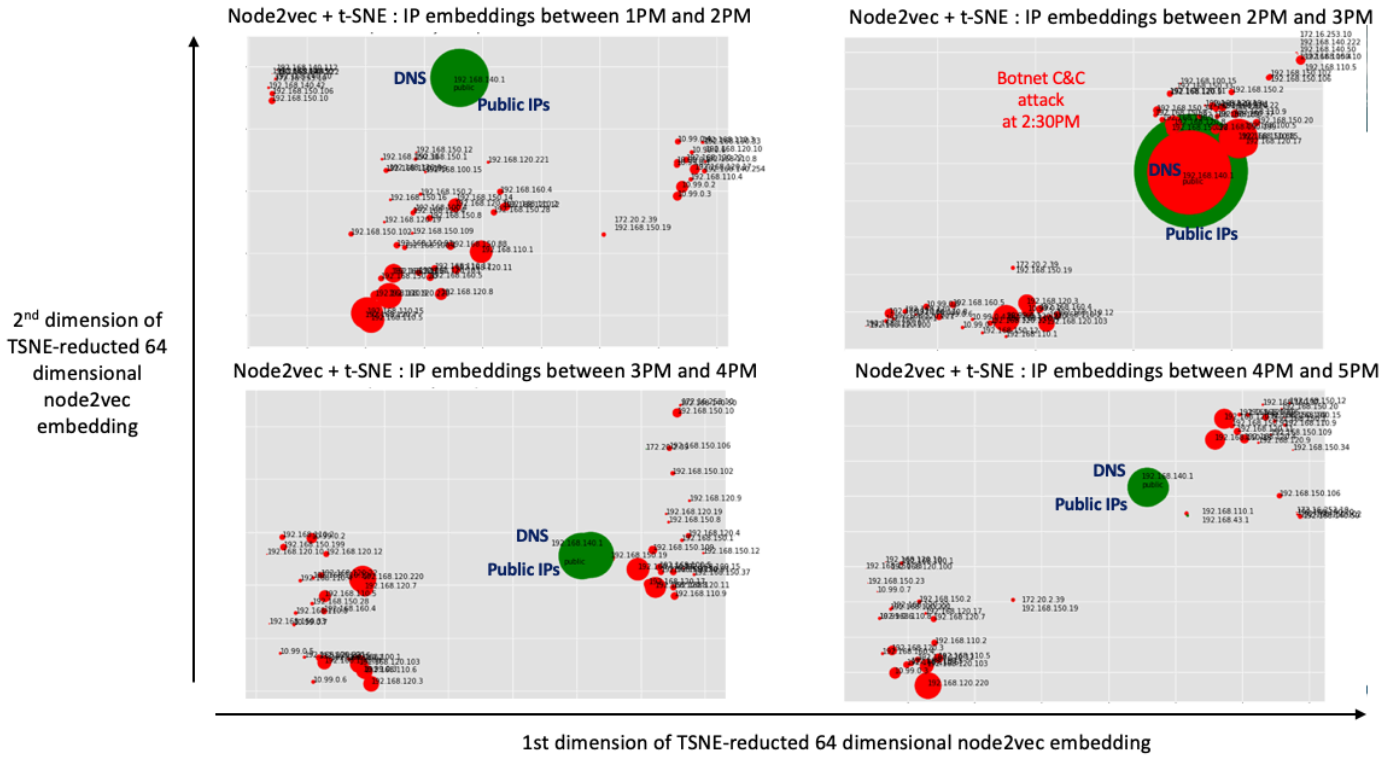
3.3.2 "Command & Control" attacks:

We now apply our method to more sophisticated attacks: botnet command & control, in order to exfiltrate data through the DNS. This is done by using the port 53, traditionally dedicated to DNS communications.

Below the results for DNS C&C attack occurring at 13h24 on May 7th, for the base method:



We analyze trajectory of the DNS (IP 192.168.140.1) over time: it tends usually to be in the top right corner (receiving IP on port 53, exchanging a lot of data), but suddenly mutates to the top left corner during the attack. This trajectory clearly exhibits a strange pattern.
Below the results with our embedding-based method (walk length = 4):



Thanks to our embeddings, we are able to visualize:

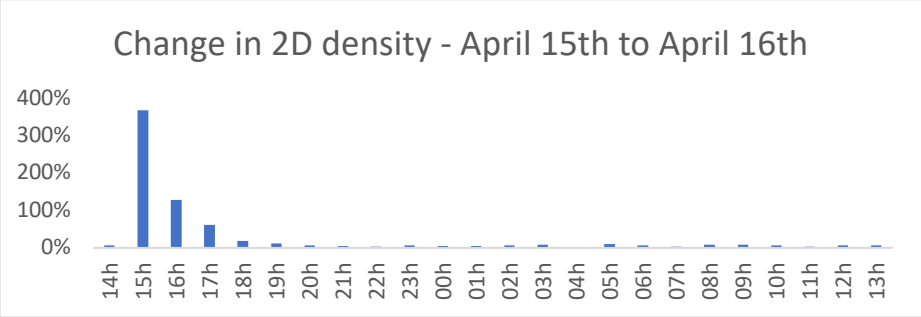
- The attack occurring at 2h30PM: the DNS and the public IP addresses become very close, with a very important weight, and their “attraction power” suddenly gets really high as a lot of emitting IP addresses mostly belonging to 192.168.120.xxx and 192.168.150.xxx subnets gets very close to those 2 IPs. The IP addresses attracted by the DNS are the “bots” who are being controlled and used by the attackant to exfiltrate data through the DNS
- The position of the private IP addresses and the public IP addresses (outside the network)
- The groups of IP addresses communicating on port 53.
- The temporal evolution of the network and the trajectories of the IP addresses over time.

3.4 Numerical detection methods:

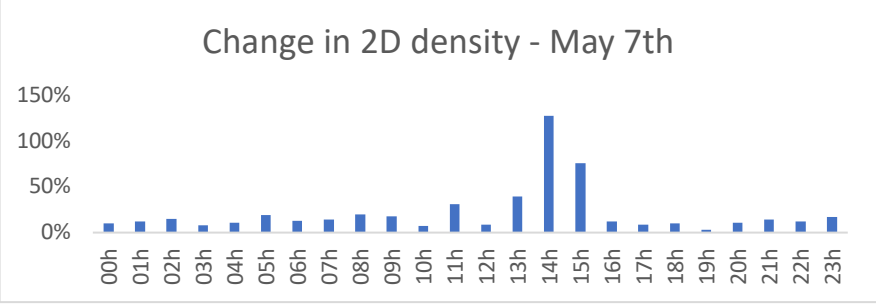
We detect anomalies by using clustering and density-based method. For every 1-hour screenshot of the network, we fit a 2-dimensional gaussian kernel and we estimate the change of the density function (2D probability density function). Our underlying assumption is that the variation of the 2D pdf should be smooth in time, unless something unexpected happens in the network. This allow us to detect the two types of attacks:

- Network discovery through port scan (attack at 15h45 on April 15th)
- Botnet C&C and data exfiltration through the DNS on port 53 (attack at 14h30 on May 7th)

Below the results for network discovery (on April 15th / April 16th):



Below the results for botnet C&C (on May 7th):



We manage to link both attacks to very high variations of the fitted density (128% for botnet C&C and 367% for network discovery). This enable to set up a threshold-based anomaly detection system for these attacks.

4 Conclusion and future work:

In this article, we presented a novel embedding-based framework designed to facilitate behavioral analysis to detect anomalies and link them to advanced cyber-attacks. This framework is composed of machine learning algorithm, it projects the representation of network's assets in a visual 2-dimensional space. The embeddings might be used by cyber analysts to improve anomaly detection. We also show that density analysis of the embeddings can be used as a detection system. The framework has been tested using two types of attacks, and give the following results:

- Network discovery through port scans: change of 367% in the fitted 2D-density
- Botnet C&C and data exfiltration through port 53: change of 128% in the fitted 2D-density

The next steps in this work are:

- Improving the temporal coherence of the embeddings (latent homophily and latent relationship homophily hypothesis)
- Defining a set of metrics used for anomaly detection and a set of associated rules (we showed in this article that a simple threshold-rule works for the attacks that we simulated, but we might want to include the temporal complexity of attacks)
- Including more features in our analysis. This could be done by using methods like the role2vec algorithm [11], which takes into account the attributes of graphs when computing the embeddings.

This work has been designed to enrich the toolbox of analysts working in security operational centers. The integration of ML methods into industrial processes gives hope for new growth drivers, in particular in the context of cybersecurity.
The operational benefit of using Machine Learning methods is recognized but is hampered by the lack of understanding of their mechanisms, at the origin of operational, legal and ethical operational problems. This highly affects the operational acceptability of AI tools. The ability of engineers, decision-makers and users to understand the meaning and the properties of the results produced by these tools will be a key success factor for the development of AI-enhanced security operational centers in the next years

References

1. Jang-Jaccard, J, Nepal, S : A survey of emerging threats in cybersecurity, Journal of Computer and System Sciences, Volume 80, Issue 5, 2014,
2. SAIFE, New-Paradigm-for-Securing-Todays-Porous-Perimeter, New-Paradigm-for-Securing-Todays-Porous-Perimeter,
3. Chawla, A, Lee, B, Fallon, S, Jacob, P : Host Based Intrusion Detection System with Combined CNN/RNN Model, In ECML PKDD 2018 Workshops, 2018
4. Yuan, Q, Wei, S,: Aligning Network Traffic for Serial Consistency and Anomalies with A Customized LSTM Model, In 2018 IEEE International Conference on Progress in Informatics and Computing (PIC), Suzhou, China, 2018
5. Mirsky, Y, Doitshman, T, Elovici, Y, Asaf Shabtai, A : Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection, In in Network and Distributed Systems Security Symposium (NDSS) 2018.
6. Sweet, C .R. : Synthesizing cyber intrusion alerts using generative adversarial networks, PhD-Thesis, Rochester Institute of Technology, 2019.
7. Adams, N, Heard, N, : Dynamic Networks and Cyber-Security, World Scientific, 2016.
8. Pingle, A : RelExt : Relation Extraction using deep learning approaches for cybersecurity knowledge graph improvement, arXiv2019
9. Mikolov, T, et al, Distributed representations of words and phrases and their compositionality, In NIPS 2016, 2016
10. Zhu, L et al, : Scalable temporal latent space inference for link prediction in dynamic networks, In IEEE Transactions on Knowledge and Data Engineering, 2016
11. Nesreen K. Ahmed and Ryan Rossi and John Boaz Lee and Theodore L. Willke and Rong Zhou and Xiangnan Kong and Hoda Eldardiry : Learning Role-based Graph Embeddings, arXiv2018