

## Masterarbeit

# Predicting SSH keys in Open SSH Memory dumps

A report by

**Rascoussier, Florian Guillaume Pierre**

*Prüfer*

Prof. Dr. Michael Granitzer

Christofer Fellicious

Prof. Dr. Pierre-Edouard Portier

---

September 22, 2023

# Abstract

As the digital landscape evolves, cybersecurity has become an indispensable focus of IT systems. Its ever-escalating challenges have amplified the importance of digital forensics, particularly in the analysis of heap dumps from main memory. In this context, the Secure Shell protocol (SSH) designed for encrypted communications, serves as both a safeguard and a potential veil for malicious activities. This research project focuses on predicting SSH keys in OpenSSH memory dumps, aiming to enhance protective measures against illicit access and enable the development of advanced security frameworks or tools like honeypots.

This Masterarbeit is situated within the broader SmartVMI project, a collaborative research initiative with the objective to advance artificial intelligence-based mechanisms for attack detection and digital forensics. Specifically, this work seeks to build upon existing research on key prediction in OpenSSH heap dumps. Utilizing machine learning algorithms, the study aims to refine feature extraction techniques and explore innovative methods for effective key detection prediction. The objective is to accurately predict the presence and location of SSH keys within memory dumps. This work builds upon, and aims to enhance, the foundations laid by SSHkex [1] and SmartKex [2], enriching both the methodology and the results of the original research while exploring the untapped potential of newly proposed approaches.

This report encapsulates the progress of a year-long Master's thesis research project executed between October 2022 and October 2023. Conducted within the framework of the PhDTrack program between the University of Passau and INSA Lyon, the research has been supervised by Christofer Fellicious and Prof. Dr. Michael Granitzer from the University of Passau, as well as Prof. Dr. Pierre-Edouard Portier from INSA Lyon. It offers an in-depth discussion on the current state-of-the-art in key prediction for OpenSSH memory dumps, research questions, experimental setups, programs development as well as discussing potential future directions.

## Acknowledgements

A special acknowledgment goes to Christofer Fellicious, my engaged supervisor at the University of Passau, for his guidance, support and feedback during the Masterarbeit.

I want to express my sincere gratitude to my colleague and friend, Clément Lahoche, whose human and technical skills have been a great source of inspiration and motivation throughout this project; especially considering that we have been working on closely related subjects. It has been a great pleasure to share our ideas and insights, and to collaborate on the development of several programs necessary for the experimentations.

Another acknowledgments go to my esteemed supervisors Prof. Dr. Granitzer and Prof. Dr. Portier for their support and feedback during the Masterarbeit.

I would also like to express my sincere gratitude to all the persons that have helped me, even punctually, during the Masterarbeit with their valuable help, insights, discussions and contributions as well as all the persons involved in the PhDTrack program that made this Masterarbeit possible, including but not limited to:

- Lionel Brunie, Director of CS Department at INSA Lyon, that makes this PhDTrack program possible from the French side.
- Harald Kosch, Head of the Chair of Distributed Information Systems at the University of Passau, that makes this PhDTrack program possible from the German side.
- Natalia Lucari, PhDTrack coordinator at INSA Lyon, for her support and help during the PhDTrack program.
- Ophelie Coueffe, PhDTrack coordinator at the University of Passau, for her support and help during the PhDTrack program.
- Elöd Egyed-Zsigmond, PhDTrack coordinator at the University of Passau, for the subject selection and administrative support.
- All the other PhDTrack students for the great atmosphere, mutual help and the interesting discussions during almost two years.

Finally, my last acknowledgments go to my family and friends for their support and encouragements.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Questions . . . . .	2
1.2	Commitment to Open Science and Reproducibility . . . . .	3
1.2.1	GitHub Repositories . . . . .	3
1.3	Structure of the Thesis . . . . .	4
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	SSH and OpenSSH Implementation . . . . .	5
2.1.1	Basics of the Secure Shell Protocol (SSH) . . . . .	5
2.1.1.1	SSH design and origin . . . . .	5
2.1.1.2	SSH keys . . . . .	6
2.1.1.3	SSH key encryption . . . . .	6
2.1.2	OpenSSH Implementation . . . . .	7
2.1.2.1	OpenSSH components . . . . .	7
2.1.2.2	OpenSSH hashing . . . . .	7
2.1.3	The state of SSH security . . . . .	9
2.1.3.1	SSH security issues . . . . .	9
2.1.3.2	SSH vulnerabilities . . . . .	10
2.1.3.3	SSH and cyber-attacks . . . . .	10
2.1.4	The Imperative of SSH Honeypots in Cybersecurity Monitoring . . . . .	11
2.1.5	Research context and motivation for this Masterarbeit . . . . .	12
2.2	Graph-based memory modelization . . . . .	12
2.2.1	Defining memory concepts and modelization . . . . .	13
2.2.1.1	Endianness . . . . .	13
2.2.1.2	The role of entropy in forensic analysis . . . . .	14
2.2.2	Graphs and Knowledge Graphs . . . . .	15
2.2.2.1	Defining Graph Theory concepts . . . . .	15

2.2.2.2	Graphs types . . . . .	16
2.2.2.3	Graph vs Knowledge Graph . . . . .	16
2.2.2.4	Inductive Reasoning and Learning . . . . .	17
2.3	Data preprocessing for Machine Learning . . . . .	17
2.3.1	Feature engineering . . . . .	18
2.3.1.1	Types of Features . . . . .	18
2.3.1.2	The Curse of Dimensionality . . . . .	18
2.3.1.3	Feature Engineering techniques . . . . .	19
2.3.1.4	Evaluating Features with Correlation Tests . . . . .	20
2.3.2	Embeddings . . . . .	20
2.3.2.1	Embedding Creation vs Feature Engineering . . . . .	21
2.3.2.2	Embeddings for graphs . . . . .	21
2.3.2.3	Word Embeddings . . . . .	23
2.3.3	Other preprocessing techniques for Machine Learning . . . . .	24
2.3.3.1	Data Cleaning . . . . .	24
2.3.3.2	Dataset splitting and sampling . . . . .	25
2.3.3.3	Dealing with Imbalanced Datasets . . . . .	25
2.4	Machine Learning and Deep Learning . . . . .	26
2.4.1	Machine Learning . . . . .	26
2.4.1.1	What is Machine Learning . . . . .	27
2.4.1.2	Model Evaluation . . . . .	27
2.4.2	Machine Learning Models for Binary Classification . . . . .	28
2.4.3	Deep Learning . . . . .	29
2.4.3.1	Recurrent Neural Networks (RNN) . . . . .	30
2.4.3.2	Long Short-Term Memory (LSTM) . . . . .	30
2.4.3.3	Gated Recurrent Units (GRU) . . . . .	30
2.4.3.4	Convolutional Neural Networks (CNN) . . . . .	31
2.4.3.5	Graph Convolutional Networks (GCN) . . . . .	31
2.4.4	Graph Neural Networks . . . . .	31

2.4.4.1 Recursive Graph Neural Networks (RecGNNs) . . . . .	32
2.4.4.2 Convolutional Graph Neural Networks (ConvGNNs) . . . . .	33
<b>3 Related Work</b>	<b>35</b>
3.1 SSHKex . . . . .	35
3.2 SmartKex . . . . .	36
3.2.1 Baseline brute-force method . . . . .	36
3.2.2 SmartKex machine-learning method . . . . .	37
3.3 Objectives of the present work . . . . .	37
<b>4 Methods</b>	<b>39</b>
4.1 OpenSSH memory dumps dataset . . . . .	39
4.1.1 Assumptions . . . . .	43
4.1.2 Estimating the dataset balancing for key prediction . . . . .	43
4.1.3 Dataset validation . . . . .	44
4.1.3.1 Automatic annotation validation . . . . .	45
4.1.4 Exploring patterns in RAW heap dump files . . . . .	48
4.1.4.1 Detecting potential pointers . . . . .	48
4.1.4.2 Detecting potential keys . . . . .	51
4.1.4.3 Detecting potential data structures . . . . .	55
4.2 Hardware and software architecture . . . . .	57
4.2.1 Hardware development and testing environment . . . . .	57
4.2.2 Software, languages and tools . . . . .	58
4.2.2.1 Packaging and deployment . . . . .	59
4.3 Converting memory dumps to graphs . . . . .	60
4.3.1 Initial work from Python to Rust . . . . .	60
4.3.1.1 First steps . . . . .	60
4.3.2 Memory Graph Representation . . . . .	61
4.3.3 Graph Construction . . . . .	62
4.3.4 Exporting the Graph . . . . .	62

4.3.5 Semantic Embedding . . . . .	62
4.4 Feature engineering . . . . .	62
4.5 ML for key prediction . . . . .	62
<b>5 Results</b>	<b>63</b>
<b>6 Discussion</b>	<b>64</b>
<b>7 Conclusion</b>	<b>65</b>
7.1 Summary of Results . . . . .	65
7.2 Outlook on Future Work . . . . .	65
.1 Code . . . . .	66
.2 Math . . . . .	66
.3 Dataset . . . . .	66
<b>Acronyms</b>	<b>67</b>
<b>References</b>	<b>68</b>
<b>Additional bibliography</b>	<b>73</b>

# 1 Introduction

The digital age has brought with it an unprecedented increase in the volume and complexity of data that is being generated, stored, and processed. This data is often sensitive in nature, and its security is of paramount importance, making cybersecurity a critical focus area. This evolving landscape is fraught with challenges that continue to amplify the importance of digital forensics in IT systems. One area that stands out for its widespread use and importance is the Secure Shell protocol (SSH) and its most popular implementation, OpenSSH. SSH is a cryptographic network protocol widely used for secure remote access to systems. It is also used for secure file transfer, and as a secure tunnel for other applications. SSH is a key component of IT systems whose encryption capabilities are critical to the security of IT systems. However, it also presents a unique set of challenges, most notably the concealment of malicious activities.

A common case is when a unauthorized actor gains access to SSH keys so as to get access to a system. This can happen through a malicious human actor, but more commonly through automated processes such as malwares and botnets. This situations present a formidable and growing threat to cybersecurity, affecting a broad range of stakeholders from governments and financial institutions to individual users. In just 2019, the number of Command and Control (C&C) servers for botnets increased by 71.5%, leading to an estimated \$19 billion in advertising theft [66]. Many malwares and botnets «have in common that they have used as attack vector the Secure Shell (SSH) remote access service» [66].

At the heart of the issue lies the fact that SSH veils its communications through encryption, making it difficult to detect malicious activities. To be able to detect those potential malicious actors, it is possible to replace SSH by a honeypot that enable to monitor pseudo-SSH activities. There is a range of readily available honeypots, such as Kippo or Cowrie, which are designed to emulate a vulnerable SSH system and attract attackers [75]. The problem lies that thoses honeypots are not able to mimic perfectly a real system, which makes them easy to detect by experienced attackers. As stated by „Analysis of SSH Honeypot Effectiveness“: «The ability to collect meaningful malware from attackers depends on how the attackers receive the honeypot. Most attackers fingerprint targets before they launch their attack, so it would be very beneficial for security researchers to understand how to hide honeypots from fingerprinting and trick the attackers into depositing malware. [...] What is certain is that if a cautious attacker believes they are in a honeypot, they will leave without depositing malware onto the system, which reduces the effectiveness of the honeypot» [76].

There are other approaches that allows to decrypt SSH connections without relying on a honeypot, like the *man-in-the-middle* or *binary manipulation* with their own set of challenges [1]. Instead of relying on softwares that mimics or modify a real system, it is possible to use a real unmodified system directly. The idea is to be able to decrypt SSH connection channels, which is possible if the SSH keys are known. Since SSH encryption keys are typically stored in the main memory of a system, it is possible for the administrators to extract them through the exploitation of memory dumps of a targeted system. In this context, the ability to detect SSH keys in memory dumps, and specifically OpenSSH keys, is critical to the development of effective SSH honeypot-like systems. The research introduced by the SmartVMI project with SSHKex, SmartKex, the present thesis and the future related work could be used to develop such a new type of system-monitoring tools. This new kind of tools would be very difficult to detect by attackers, increasing their effectiveness, and wouldn't require the alteration of the system. The present report is focused on the SSH key detection in memory dumps, which is a key component allowing to decode SSH communications such that it become possible to intercept malicious communications and to detect malicious



activities.

## 1.1 Research Questions

At the very beginning of this thesis, the objective was to answer the following research questions:

- What is the state of the art in the field of security key detection in heap dump memory?
- What are the challenges of security key detection in heap dump memory?
- How can the existing methods for detecting SSH keys in OpenSSH heap dumps be improved?

The SmartVMI project has already made significant progress in the detection of SSH keys in OpenSSH heap dumps. An open dataset of memory dumps has been created, and a simple yet effective method for detecting SSH keys has been developed. The dataset has been used to train and test simple machine learning algorithms, and the results have been promising. The research has been published in the form of two papers, SSHkex [1] and SmartKex [2], which is the basis of this thesis.

However, there is still room for improvement, particularly in the area of machine learning algorithms. This thesis seeks to build upon the existing research by refining feature extraction techniques and exploring innovative methods for effective key detection prediction. The objective is to accurately predict the presence and location of SSH keys within memory dumps. Rooted in this context, this Masterarbeit aims to address several key research questions:

- **Memory graph:** How can we develop a memory graph representation to improve the prediction of SSH keys in memory dumps?
- **Memory graph embedding:** How can we develop a memory graph embedding representation to improve the prediction of SSH keys in memory dumps?
- **Feature importance:** What features are most indicative of SSH keys in memory dumps?
- **Feature extraction:** How can these features be extracted from memory dumps and used to train machine learning algorithms?
- **ML for key prediction:** How can machine learning algorithms be optimized for the prediction of SSH keys in memory dumps?
- **Graph Convolutional Networks for key prediction:** How can GCN be used to improve the prediction of SSH keys in memory dumps, and how does it compare to traditional machine learning algorithms?

By tackling these research questions, this thesis seeks not only to advance the academic understanding of SSH key prediction and digital forensics but also to provide practical insights that could lead to the development of more secure and effective systems.

## 1.2 Commitment to Open Science and Reproducibility

In alignment with the principles of Open Science, this thesis aims to be not just a scholarly report but also a comprehensive guide for anyone who wishes to understand, replicate, or extend the work presented. Open Science is a movement that advocates for the transparent and accessible sharing of scientific research, data, and dissemination processes [37]. It is built on six fundamental principles [36]:

1. **Open Methodology:** Detailed methodologies are provided to ensure that the experiments can be replicated.
2. **Open Source:** All code used in this research is available for scrutiny and reuse. As such, all code including the L<sup>A</sup>T<sub>E</sub>X code for the present report <sup>1</sup> is properly documented and can be accessed on GitHub.
3. **Open Data:** Raw data and the data processing techniques are made publicly available.
4. **Open Access:** The research is published in a manner that is free for all to read and download.
5. **Open Peer Review:** The peer review process is transparent. In the case of this Masterarbeit, the research is reviewed by the supervisors of the project.
6. **Open Educational Resources:** Any educational content produced is shared openly.

To ensure the highest level of reproducibility and accessibility, this thesis includes what might seem like exhaustive details, such as hardware or software specifications, precise shell commands and some code implementations used during the research. These are included to provide a complete picture and to minimize the friction for those who wish to replicate the experiments, whatever their level of expertise may be. By adhering to the principles of Open Science, this thesis aims to contribute to a more transparent, collaborative, and efficient scientific community.

### 1.2.1 GitHub Repositories

In the context of the present Masterarbeit, a number of GitHub repositories have been created to facilitate the sharing of code and data. These repositories are listed below:

- **masterarbeit\_report:** Repository containing the LaTeX code for the report as well as several small scripts related to dataset exploration. [https://github.com/0nyr/masterarbeit\\_report](https://github.com/0nyr/masterarbeit_report)
- **mem2graph:** Memory graph creation utility built in Rust, featuring different graph creation and embedding strategies. Collaboration with Clément Lahoche. <https://github.com/0nyr/mem2graph>
- **research-base:** Custom Python framework for developing programs that include all the basics of a research project, such as logging, environment and argument loading, result keeping, and more. Collaboration with Clément Lahoche. <https://github.com/0nyr/research-base>

---

<sup>1</sup>The present report repository can be found here: [https://github.com/0nyr/masterarbeit\\_report](https://github.com/0nyr/masterarbeit_report)

- **data\_processing\_masterarbeit**: Python program for data processing and machine learning for SSH key prediction. [https://github.com/0nyr/data\\_processing\\_masterarbeit](https://github.com/0nyr/data_processing_masterarbeit)
- **phdtrack\_project\_3**: Historical repository containing the first version of the memory graph creation utility and the first version of the dataset creation script. Collaboration with Clément Lahoche. [https://github.com/0nyr/phdtrack\\_project\\_3](https://github.com/0nyr/phdtrack_project_3)

## 1.3 Structure of the Thesis

# 2 Background

This section is dedicated to the background information needed to understand the work developed in the thesis. It provides the necessary context for the research, including the problem being solved, why it's important, and the related state-of-the-art and background information. It also includes fundamental concepts and theories, terminology definitions, and a high-level overview of the problem domain. It serves as a primer to the rest of the report, providing the necessary context for the research and is intended for readers who may not be experts in the specific area of the research but have some knowledge of the broader field.

## 2.1 SSH and OpenSSH Implementation

### 2.1.1 Basics of the Secure Shell Protocol (SSH)

The Secure Shell Protocol, commonly known as SSH, is designed to facilitate secure remote login and other secure network services over insecure networks. SSH has been design since its inception with security in mind, as a successor of the Telnet protocol, which is not secure, and other «unsecured remote shell protocols such as rlogin, rsh and rexec» [1].

#### 2.1.1.1 SSH design and origin

As stated by the authors of the *SSH Annual Report 2018*, «The founder of SSH, Tatu Ylönen, designed the first version of the SSH protocol after a password-sniffing attack at his university network. Tatu released his implementation as freeware in July 1995, and the tool quickly gained in popularity. Towards the end of 1995, the SSH user base had grown to 20,000 users in fifty countries. By 2000, there were an estimated 2,000,000 users of the protocol. Today, more than 95% of the servers used to power the Internet have SSH installed in them. The SSH protocol is truly one of the cornerstones of a safe Internet.» [67].

SSH is defined in *The Secure Shell (SSH) Protocol Architecture* [59]. It is divided into three major components:

- **Transport Layer Protocol:** This provides server authentication, confidentiality, and integrity. It can also optionally provide compression. Typically, the transport layer runs over a TCP/IP connection but can also be used on top of any other reliable data stream.
- **User Authentication Protocol:** Running over the transport layer, this protocol authenticates the client-side user to the server. Multiple methods of authentication such as password and public key are supported.
- **Connection Protocol:** This multiplexes the encrypted tunnel established by the preceding layers into several logical channels. Channels can be used for various purposes, such as setting up secure interactive shell sessions or tunneling arbitrary TCP/IP ports.

«The client sends a service request once a secure transport layer connection has been established. A second service request is sent after user authentication is complete. This allows new protocols to be defined and coexist with the protocols listed above» [59].

### 2.1.1.2 SSH keys

For the purposes of this Masterarbeit, a comprehensive understanding of SSH's key exchange and encryption mechanism is important. As outlined in SSHKex [1], the SSH protocol utilizes a key exchange procedure that culminates in a derived master key  $K$  and a hash value  $h$ . These components are pivotal for encrypting client-server communications and identifying sessions.

During the key exchange process, Diffie-Hellman is employed to negotiate an ephemeral shared key between the client and the server [59] [58]. This ephemeral key is then signed by the server using either RSA, DSA, or another suitable signature algorithm (see 2.1.2.2). The signed key confirms to the client that the negotiated key is indeed from the intended server and not an imposter or a middleman, thereby preventing man-in-the-middle (MITM) attacks.

In addition, the host key of the server is used to sign the Diffie-Hellman parameters. This key is not to be confused with the client key listed in the server's `authorized_keys` file, which is used later for client authentication.

The key exchange process results in multiple session keys computed for various purposes:

- **Initialization Vectors:** Key A and Key B are designated for initialization vectors from the client to the server and vice versa.
- **Encryption Keys:** Key C and Key D act as encryption keys for client-to-server and server-to-client communications, respectively.
- **Integrity Keys:** Key E and Key F are utilized to preserve the integrity of data transmitted between the client and server.

This approach provides forward secrecy: if the private key is stolen, it does not compromise the encryption of old sessions. This is because the Diffie-Hellman parameters are ephemeral and discarded once they are no longer needed. Therefore, the only long-lasting keypair is used for authentication purposes.

### 2.1.1.3 SSH key encryption

These keys are computed using hash functions that take the master key  $K$  and a hash value  $h$ , a unique letter (A, B, C, D, E, or F), and the session ID as inputs. This is summarized in „The OpenSSH Protocol under the Hood“: « The equations used for deriving the above vectors and keys are taken from RFC 4253 [60]. In the following, the `||` symbol stands for concatenation,  $K$  is encoded as mpint, `"A"` as byte and `sessionid` as raw data. Any letter, such as the `"A"` (in quotation marks) means the single character A, or ASCII 65.

- Initial IV client to server:  $HASH(K||H||"A"||session_id)$ .
- Initial IV server to client:  $HASH(K||H||"B"||session_id)$ .
- Encryption key client to server:  $HASH(K||H||"C"||session_id)$ .
- Encryption key server to client:  $HASH(K||H||"D"||session_id)$ .
- Integrity key client to server:  $HASH(K||H||"E"||session_id)$ .

- Integrity key server to client:  $HASH(K||H||F||session_id)$ .

» [61]. Details about the hash function are given in the next section.

The most interesting keys are the encryption keys, as they are used to encrypt the communication between the client and the server. The other keys are used for integrity checks and initialization vectors. Decrypting encrypted SSH communication necessitates either to retrieve these session keys and variables so as to recompute the keys, or to retrieve those keys directly, which is the focus of this Masterarbeit.

## 2.1.2 OpenSSH Implementation

OpenSSH (OpenBSD Secure Shell) is an open-source implementation written in C of the SSH protocol suite, and it is the most widely used SSH implementation [61]. It is the default SSH implementation on most Linux distributions, and it is also available for Windows. OpenSSH is used for a wide range of purposes, including remote command-line login and remote command execution. It is also used for port forwarding, tunneling, and transferring files via SCP and SFTP either manually or via automated processes, such as backup systems, configuration management tools, and automated software deployment tools.

### 2.1.2.1 OpenSSH components

OpenSSH is composed of several tools and daemons, including client and server components [65]:

- **ssh:** The basic client program that allows to log into and execute commands on a remote machine.
- **sftp:** An interactive file transfer program that uses SSH to secure the connection.
- **sshd:** This is the SSH daemon that runs on the server. This is used for connecting to a remote machine when using the SSH client from another system.
- **ssh-agent:** The program that holds private keys in memory, so one doesn't have to enter ones passphrase every time.
- **ssh-add:** A program for adding RSA or DSA identities to the authentication agent.
- **ssh-keygen:** A utility for creating and managing SSH keys.
- **ssh-keyscan:** A utility for gathering public SSH host keys from a number of hosts.
- **ssh-keychk:** A utility for checking the validity of SSH keys.
- Several other tools to support the SSH protocol and the OpenSSH implementation.

### 2.1.2.2 OpenSSH hashing

OpenSSH employs a variety of hash functions and algorithms to secure data, most commonly using SHA1. However, SHA1 is increasingly seen as weak due to its vulnerability to collision attacks [61]. In light of this, the contemporary standard leans towards SHA512. The hash functions

are used alongside cipher algorithms like «Advance Encryption Standard (AES) Cipher Block Chaining (CBC), AES Counter (AES-CTR), and ChaCha20» [2]. The Message Authentication Code (MAC) typically uses either MD5 or SHA1 hash algorithms in combination with a secret key. Since cybersecurity and cryptography are constantly evolving, so do SSH and OpenSSH. Depending on the version [61], the available hash options include:

- **ssh-dss:** (*disabled at run-time since OpenSSH 7.0 released in 2015*) SSH-1 version using Digital Signature Algorithm (DSA) from the Digital Signature Standard (DSS). Originally popular but phased out due to vulnerabilities to collision attacks for DSA Key in a 1024-bit modulus. As stated by *Key Exchange (KEX) Method Updates and Recommendations for Secure Shell (SSH)*: «These attacks are still computationally very difficult to perform, but it is desirable that any key exchange using SHA-1 be phased out as soon as possible» [3] [10].
- **ssh-rsa:** (*disabled at run-time since OpenSSH 8.8 released in 2021*) It refers to the use of RSA (Rivest-Shamir-Adleman) encryption algorithm. In the context of SSH-1, this version had to be replaced due to the related to key size issue similar to DSS: «RSA 1024-bit keys have approximately 80 bits of security strength»... «which may not be sufficient for most users.» [3] [15].
- **ecdsa-sha2-nistp256:** (*since OpenSSH 5.7 released in 2011*) Uses the SHA-2 family for hashing and the NIST P-256 curve. It is considered secure and efficient, with an Estimated Security Strength (ESS) of 128 bits [3] [7].
- **ecdsa-sha2-nistp384:** (*since OpenSSH 5.7*) Utilizes the SHA-2 family and the larger NIST P-384 curve for additional security at the cost of performance. It has an ESS of 192 bits [3] [7].
- **ecdsa-sha2-nistp521:** (*since OpenSSH 5.7*) Employs SHA-2 and the even larger NIST P-521 curve for maximal security with an ESS of 256 bits [3]. It is less commonly used due to performance considerations [7].
- **ssh-ed25519:** (*since OpenSSH 6.5 released in 2014*) Known for high security and performance efficiency; employs the Ed25519 elliptic curve with an ESS of 128 bits [3] which is similar to *ecdsa – sha2 – nistp256*, and has been more prevalent following the 2013 suspicions of NSA backdoors in NIST curves [6] following the Snowden revelations [4] [5] [9].
- **rsa-sha2-256:** (*since OpenSSH 7.2 released in 2016*) An upgrade from *ssh-rsa*, using SHA-256 (with ESS of 128 bits) for hashing to improve security without major performance hits [13].
- **rsa-sha2-512:** (*since OpenSSH 7.2*) Similar to *rsa – sha2 – 256* but employs SHA-512 for even stronger security, albeit with some performance cost [13].
- **ecdsa-sk:** (*since OpenSSH 8.2 released in 2020*) Security Key-enabled, uses NIST curves and is geared towards modern hardware-based authentication [14].
- **ed25519-sk:** (*since OpenSSH 8.2*) Similar to *ssh-ed25519* but integrates hardware-based Security Keys for an additional layer of security [14].
- **NTRU Prime-x25519:** (*since OpenSSH 9.0*) A new, highly secure algorithm focused on post-quantum cryptography, providing future-proof security [62] [57].

These hashes have fixed lengths such that key lengths range between 12 and 64 bytes [2]. Since high-quality random number generation is crucial to ensure that those keys are secure

and difficult to predict, it can thus be assumed that those keys have a high entropy [64]. This is a crucial assumption as it is the basis for the use of both brute force and machine learning algorithms to predict the presence and location of SSH keys in memory dumps.

The keys generated by these hash functions are pseudo-random numbers stored in the system's RAM. Following the Kerckhoffs' principle: that «a cryptosystem should be secure, even if everything about the system, except the key, is public knowledge», the code for the OpenSSH implementation is open-source and available on GitHub [65]. This allows for the analysis of the code and the identification of the memory structures where the keys are stored.

### 2.1.3 The state of SSH security

Since its origins, SSH has been developed with cybersecurity in mind, and is generally considered a secure method for remote login and other secure network services over an insecure network. However, as with any technology, it can be exploited if not configured or managed correctly. The protocol is used by system administrators to manage remote systems, and it is also used by automated processes to transfer data and perform other tasks. This makes SSH a valuable target for attackers. In fact, SSH has been a popular target for cyber-attacks. Due to being so prevalent, it is often used by threat actors either as a vector for initial access, as a means to move laterally across a network or as a covered exit for exfiltration of sensitive data [74]. The encrypted nature of its communications makes it an attractive option for attackers, as it can be difficult to detect malicious activity.

#### 2.1.3.1 SSH security issues

Here are some cases where SSH can be involved in cyber-attacks, although it's important to note that SSH itself is not inherently insecure:

- **SSH Brute-Force Attacks:** One of the most common types of attacks involving SSH is a brute-force attack, where an attacker tries to gain access by repeatedly attempting to login with different username-password combinations. These attacks are not sophisticated but can be effective if strong authentication measures are not in place. For instance, the botnet *Chabulo* was used to launch a large-scale brute-force attack «through compromised SSH servers and IoT devices» in 2018 [67].
- **SSH Key Theft:** In some advanced attacks, threat actors have stolen SSH keys to move laterally across a network after initial entry. This allows them to authenticate as a legitimate user and can make detection much more challenging. It can « occur when users have their SSH password or unencrypted keys stolen through a variety of methods (sniffed via a key-logging console program, shoulder-surfed via bad security awareness, poor key management practices, etc.).» [68].
- **Man-in-the-Middle Attacks:** Although SSH is designed to be secure, it can be susceptible to man-in-the-middle attacks if proper verification of SSH keys is not done during the initial connection setup [61].
- **Misconfiguration:** As with any technology, misconfiguration can lead to security issues. For example, leaving default passwords, using weak encryption algorithms, or enabling root login can all make an SSH-enabled system vulnerable [66].



### 2.1.3.2 SSH vulnerabilities

In cybersecurity, it is generally considered that any system that is connected to the Internet will be attacked at some point. Similarly, it is a common saying that no system is 100% secure. This is true for SSH as well. Although it is a secure protocol, it can be exploited if not configured or managed correctly.

Some vulnerabilities have also been discovered in the protocol itself, although these are rare.

- **SSH-1 Vulnerabilities:** A series of vulnerabilities in the first implementation of SSH were discovered from 1998 to 2001, with its subsequent fixes leading to unauthorized content insertion and arbitrary code execution. SSH-1 had many design flaws and is now considered obsolete. [70], [69].
- **CBC Plaintext Recovery:** A theoretical vulnerability discovered in 2008 affecting all versions of SSH, allowing the recovery of up to 32 bits of plaintext from CBC-encrypted ciphertext [71].
- **Suspected Decryption by NSA:** Leaked information in 2014 suggested that the NSA might be able to decrypt some SSH traffic, although the protocol itself was not confirmed to be compromised [72].

### 2.1.3.3 SSH and cyber-attacks

SSH has been used in many high-profile cyber-attacks and malwares, including the following:

- **Operation Windigo:** This was a large-scale campaign that infected over 25,000 UNIX servers. SSH was one of the vectors used for maintaining control over compromised servers. A report by ESET mentions that the OpenSSH backdoor Linux/Ebury was first discovered in 2011 as a component of the aforementioned operation. «This operation has been ongoing since at least 2011 and has affected high profile servers and companies, including cPanel - the company behind the famous web hosting control panel - and Linux Foundation's kernel.org - the main repository of source code for the Linux kernel» [73].
- **Linux/Hydra:** Initially unleashed in 2008, this malware is a fast login cracker that targets a range of popular protocols including SSH. Hence, SSH is one of its primary vectors to gain initial access to Internet of Things (IoT) devices. Once a device is infected by Linux/Hydra, it joins an IRC channel and initiates a SYN Flood attack [75].
- **Psyb0t:** Discovered in early 2009, Psyb0t is an IRC-controlled malware specifically designed to target devices with MIPS architecture, such as routers and modems. Notably, it was responsible for orchestrating a DDoS attack against the DroneBL service, infecting up to 100,000 devices for this purpose. The malware is equipped to conduct UDP and ICMP flood attacks and employs a brute-force attack mechanism against Telnet and SSH ports. Remarkably, it uses a pre-configured list of 6,000 usernames and 13,000 passwords to perform these attacks [75].
- **Chuck Noris:** Similar to Psyb0t in its objectives and methods, Chuck Noris targets routers and DSL modems, focusing on SoHo (small office/home office) devices. However, unlike Psyb0t, which uses ICMP flood attacks, Chuck Noris deploys ACK flood attacks. The malware carries out brute-force attacks on Telnet and SSH open ports, drawing parallels to the tactics employed by Psyb0t but with the specific variation in flooding techniques [75].

It's worth noting that in many of these cases, SSH was not the initial attack vector but was used at some stage in the attack lifecycle. Properly configured and managed SSH is still considered a secure and robust protocol for remote access and data transfer. In all those situations, a tool monitoring the SSH traffic could have detected the malicious activities and prevented the attack.

#### 2.1.4 The Imperative of SSH Honeypots in Cybersecurity Monitoring

SH (Secure Shell) has become an indispensable protocol for secure communication but can also conceal malicious agents. This reality underscores the urgency for robust monitoring mechanisms capable of identifying suspicious activities in real-time. Among various countermeasures, SSH honeypots have emerged as a particularly effective tool for monitoring and gathering intelligence on potential threats.

An SSH honeypot is a decoy server or service that mimics legitimate SSH services. The primary aim is to attract cybercriminals and study their tactics, thereby offering an active form of surveillance and data collection. Unlike traditional intrusion detection systems, honeypots do not merely identify an attack; they engage the attacker in a controlled environment, enabling detailed observation and logging of the intruder's actions. This allows for the collection of valuable information, such as the attacker's IP address, the tools used, and the techniques employed. This data can then be used to enhance security measures and develop more robust countermeasures [75].

SSH honeypots serve as an invaluable asset in the cybersecurity arsenal, providing not just a reactive but a proactive measure against evolving cyber threats. They can collect actionable intelligence on new hacking methods, malware, and exploitation scripts. This information can be crucial for proactively securing actual production environments. The data collected can also be used to trace back to the origin of the attack, facilitating legal pursuits against the perpetrators. By diverting attackers to decoy servers, honeypots also protect real assets from being targeted, saving both computational resources and administrative effort needed for post-incident recovery.

Popular SSH honeypots include Kippo, Cowrie, and HoneySSH. Cowrie is a fork of Kippo, with additional features such as logging of attacker's keystrokes and file transfer.

- **Kippo:** Kippo is a medium-interaction honeypot that logs the attacker's shell interaction. It specializes in capturing brute force and Telnet-based attacks [75].
- **Cowrie:** Serving as Kippo's successor, Cowrie emulates various protocols including SSH, SFTP, and SCP. It logs events in JSON format, making it particularly useful for detecting brute force and Telnet-based attacks, as well as spoofing attacks [75].
- **IoTPOT:** This IoT-focused honeypot supports multiple CPU architectures and can detect a variety of attacks including brute force, DoS, and sniffing attacks on Telnet, SSH, and HTTP ports [75].
- **HoneySSH:** HoneySSH is a low-interaction honeypot that emulates an SSH server and logs the attacker's IP address, username, and password [77].
- **Sarracenia (SSHKex):** Introduced in 2018, Sarracenia is a high-interaction SSH honeypot that has been enhanced by SSHKex. Instead of «requiring the VM to be paused for every incoming or outgoing packet, which degrades the server performance» [1], SSHKex allows for the extraction of derived SSH session keys. This reduces the performance degradation significantly, as the VM is paused less frequently [1] [8].

These honeypots are useful tools for gathering intelligence on potential threats. However, they are not without their limitations.

### 2.1.5 Research context and motivation for this Masterarbeit

Security and malware detection are active areas of research, with SSH honeypots being a particularly promising tool for gathering intelligence on potential threats. However, they are not without their limitations. For instance, they are not able to perfectly mimic a real system, such that attackers might be able to detect them „Analysis of SSH Honeypot Effectiveness“.

As explained in „Analysis of SSH Honeypot Effectiveness“: «As attackers become more sophisticated with their ransomware and malware campaigns, there is a significant need for security researchers to assist the greater community by running vulnerable honeypot machines to collect malicious software». Hetzler, Chen, and Khan explain that «the ability to collect meaningful malware from attackers depends on how the attackers receive the honeypot. Most attackers fingerprint targets before they launch their attack, so it would be very beneficial for security researchers to understand how to hide honeypots from fingerprinting and trick the attackers into depositing malware.» They conclude that «What is certain is that if a cautious attacker believes they are in a honeypot, they will leave without depositing malware onto the system, which reduces the effectiveness of the honeypot for security research.» We can extrapolate this conclusion to SSH honeypots, which are also vulnerable to fingerprinting and detection by attackers.

Hence, the need for more advanced SSH honeypots-inspired tools that can leverage data forensic and machine learning techniques so as to be able to use directly a real server as a honeypot, without the need to emulate a system. The current master’s thesis is aligned with this ongoing research (see 3), further enhancing the state of SSH honeypots. It aims to develop algorithms, proof of concepts and tools that can extract SSH keys from memory dumps of a real server, and use them to decrypt SSH traffic. This could lead to the development of new tools for SSH monitoring, as discussed in the future work section 7.2.

## 2.2 Graph-based memory modelization

In the following section, we present important concepts that will be used for the memory modelization of the heap dump.

Because the dataset used is composed of RAW heap dump files from OpenSSH, it is a critical aspect to understand how memory works at a low level point of view. This section aims to provide an in-depth understanding of how memory is managed in C, the language used in the OpenSSH implementation of SSH, for a *linux-x86\_64* architecture. We will explore the fundamental concepts of memory management, including the heap and the stack, memory allocation, and the role of pointers. These concepts will serve later as the foundation for our graph-based approach to memory modelization.

This section will also introduce many graph theory and Knowledge Graph (KG) concepts. We will explore the fundamentals of graph theory, including the definition of a graph, its components, and its properties. We will also discuss the concept of a Knowledge Graph (KG), which is a type of graph that stores information in the form of nodes and edges, and its applications in the field of machine learning.

### 2.2.1 Defining memory concepts and modelization

Memory management in C is a complex task that requires a deep understanding of the language's features, the operating system's capabilities and the compiler used. In C, memory is primarily managed through two built-in functions: `malloc` (memory allocation) and `free` (memory deallocation). These functions operate on two primary types of memory: the heap and the stack.

- **Heap vs Stack:** The heap is used for dynamic memory allocation, where variables are allocated and freed at runtime. In contrast, the stack is used for static memory allocation, where variables are allocated and deallocated automatically. The stack is faster but has a limited size, while the heap is more flexible but requires manual management to prevent memory leaks.
- **Heap Dump:** A heap dump is a snapshot of the heap's state at a given time. It provides valuable information about the memory layout, active pointers, and data stored in the heap. Analyzing heap dumps can help in debugging memory-related issues and understanding the program's behavior.
- **Memory Addresses:** Each location in memory is identified by a unique memory address. These addresses are usually represented in hexadecimal notation. Note that the address `0x0` is reserved for the `NULL` pointer, which is used to indicate that a pointer does not point to any memory location.
- **Pointer:** A pointer is a variable that stores the memory address of another variable. It is used to indirectly access the value of the variable it points to. Pointers are used extensively in C, particularly for dynamic memory allocation.
- **Data Structure:** A data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data. In the context of C programming, data structures are declared using the keyword `struct` and are byte aligned. This means that the size of the data structure is always a multiple of the size of the largest member of the structure. Structures can be nested within each other, and pointers can be used to indirectly access the members of a structure. Data structures are often stored in the heap using `malloc`.
- **Malloc headers:** When `malloc` is called, it allocates a block of memory in the heap and returns a pointer to the first byte of the block. The heap manager keeps track of these allocations through metadata, often stored in headers preceding the allocated blocks. These headers contain information such as the size of the allocated block and whether it is free or occupied. Note that the pointer returned by `malloc` in C points to the first byte of the block of memory that has been allocated for your use, not to the `malloc` header. The `malloc` header, is managed internally by the memory allocator and is not exposed to the programmer, but is visible in the heap dump.

#### 2.2.1.1 Endianness

Endianness refers to the byte order used to represent multi-byte data types. In a *little-endian* system, the least significant byte is stored first, while in a *big-endian* system, the most significant byte is stored first. Knowing the endianness of the system is crucial for interpreting the content of memory [63].

For instance, the hexadecimal value `0x56343a198000` (taken from `"HEAP_START"` of 4.3) is represented as 550179058774 ( $\simeq 5.50e + 11$ ) in decimal basis in a little-endian system, while it is represented as 94782313037824 ( $\simeq 9.48e + 13$ ) in a big-endian system.

**Little-Endian Conversion** The conversion of a hexadecimal number in *little-endian* format to a decimal number is given by the following formula:

$$\text{Decimal} = \sum_{i=0}^{N-1} (\text{HexDigit}_i \times 16^i)$$

**Big-Endian conversion** And the conversion of a hexadecimal number in *big-endian* format to a decimal number is given by following formula:

$$\text{Decimal} = \sum_{i=0}^{N-1} (\text{HexDigit}_{N-1-i} \times 16^i)$$

Here,  $\text{HexDigit}_i$  is the value of the  $i$ -th digit in the little-endian hexadecimal number, and  $N$  is the number of digits in the hexadecimal number. Note that  $\text{HexDigit}_i$  should be converted to its decimal equivalent ('A' becomes 10, 'B' becomes 11, etc.) before performing the calculation.

These formulas will be used later to convert pointer addresses from hexadecimal to decimal format in *mem2graph*.

### 2.2.1.2 The role of entropy in forensic analysis

Entropy plays a pivotal role in forensic analysis, particularly in the context of memory dumps analysis. It serves as a measure of uncertainty and randomness, which can be crucial for tasks such as endianness detection and identifying encrypted keys in memory.

As defined by Shannon in the realm of information theory, it is a measure of the uncertainty or randomness associated with a set of possible outcomes [63] [18]. In digital applications, when calculated using the logarithm to base 2, entropy represents the amount of bits of information in a message.

**Entropy Formula** The entropy  $H$  of a message is calculated using the formula:

$$H = - \sum_{i=1}^n p_i \log_2 p_i$$

Where  $p_1, p_2, \dots, p_n$  are the probabilities of the set of all possible messages [63].

Endianness, as presented before, refers to the byte order used to represent multi-byte data types in computer memory [63]. It is necessary to know the endianness of the heap dump to correctly interpret the content of memory, and especially the addresses of potential pointers. In this context, entropy can be used to infer the endianness of a system by analyzing the distribution of byte values in a memory dump, as presented in „Inference of Endianness and Wordsize From Memory Dumps“ [63].

Entropy is also a key element for encryption key detection, since those should be random byte sequences with high entropy [2]. By examining 8-byte aligned data for high entropy, it is possible to detect potential keys in a heap dump. Techniques such as the calculation of discrete differences

and logical operations can further refine this detection as described in „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ [2].

## 2.2.2 Graphs and Knowledge Graphs

In this project, we (i.e. Clément Lahoche and the author) have built a program called *mem2graph*, a generic tool developed in Rust for efficiency. It is used to convert a RAW memory dump into a graph representation. This graph is technically a directed Edge-labeled heterogeneous property graph, whose concepts are described later. As such, and while it is debatable whether or not *mem2graph* can be seen as building Knowledge Graphs (KGs), it relies on many concepts from the domain, necessitating a proper introduction.

### 2.2.2.1 Defining Graph Theory concepts

Graph theory is a mathematical field concerned with the study of graphs. It has applications in various fields, including computer science, social sciences, or linguistics. Graphs are used to model pairwise relations between objects, and the study of graphs involves analyzing the properties of these relations. In a few words, a graph is just a collection of nodes and edges.

A graph can be formally defined as: «a pair  $G = (S, A)$  where:

- $S$  is a finite set of vertices.
- $A$  is a set of pairs of vertices  $(s_i, s_j) \in S^2$ .

Graphs can be either directed or undirected. In a directed graph, the pairs  $(s_i, s_j)$  are ordered, representing arcs from  $s_i$  to  $s_j$ . In an undirected graph, the pairs are unordered, representing edges between  $s_i$  and  $s_j$  » [11]. Let's introduce some vocabulary to describe graphs:

- **Node (or Vertex):** A single entity in the graph, often represented as a circle.
- **Edge (or Arc):** A connection between two nodes, often represented as a line or arrow.
- **Degree:** The number of edges connected to a node.
- **Path:** A sequence of edges that connect two nodes.
- **Cycle:** A path that starts and ends at the same node. every pair of nodes.
- **Order:** The number of vertices in the graph.
- **Adjacency:** Two nodes are adjacent if there is an edge between them.
- **Loop:** An edge that connects a node to itself.
- **Weight:** A value assigned to an edge.
- **Ancestors (parents) and Descendants (children):** A node  $s_i$  is an ancestor of  $s_j$  if there is a path from  $s_i$  to  $s_j$ . A node  $s_j$  is a descendant of  $s_i$  if there is a path from  $s_i$  to  $s_j$ .

Various other terminologies and concepts exist in graph theory, but these are the most important ones for our purposes. For a more in-depth understanding of graph theory, the reader is encouraged to consult the work by Solnon as a quick introduction, [11] or a more in-depth one by West et al. [12].

### 2.2.2.2 Graphs types

Graphs offer a flexible way to conceptualize, represent, and integrate diverse and incomplete data. Many graph models exist, each with its own advantages and disadvantages, as well as graph properties <sup>1</sup>. Those different types of graphs include:

- **Directed Edge-labelled Graphs (DEL):** The classic graph, set of nodes and edges that connect the nodes with in certain way. RDF is a popular DEL data model.
- **Heterogeneous Graphs:** Each node and edge is assigned one type, allowing for partitioning nodes according to their type, which is useful for machine learning.
- **Property Graphs:** Allows a set of property-value pairs and a label to be associated with nodes and edges. This model is used in Neo4j and offers great flexibility but is harder to handle and query.
- **Graph Dataset:** A set of named graphs, with a default graph with no ID. Useful when working with different sources.
- **Hypergraphs:** Allow edges that connect sets rather than pairs of nodes.

A given graph can have a range of properties that give some insights into its structure. Here are some important properties of graphs:

- **Connected Graph:** A graph is connected if there is a path between every pair of nodes.
- **Disconnected Graph:** A graph is disconnected if there is at least one pair of nodes that are not connected by a path.
- **Cyclic Graph:** A graph is cyclic if it contains at least one cycle.
- **Acyclic Graph:** A graph is acyclic if it does not contain any cycles.
- **Complete Graph:** A graph is complete if there is an edge between every pair of nodes.

### 2.2.2.3 Graph vs Knowledge Graph

A Knowledge Graph (KG) is a specialized form of graph intended to accumulate and convey real-world knowledge. As said before, we are not technically building KGs, but we rely on many concepts from the domain. Research on KG has further accelerated in recent years, and introduced significant improvement to Graph Theory, especially in the practical applications of graph construction and use for Machine Learning or Deep Learning [78]. They have a number of benefits when compared with a relational model or NoSQL alternatives, such as the ability for data to evolve in a more flexible manner, and the capacity to organize data in a way that is not hierarchical. They can represent incomplete information, and does not require a precise schema [78, p.2], which is invaluable in the context of memory analysis, where the structure of the heap is not known in advance.

---

<sup>1</sup>Some parts of the following section are directly from a prior work for Seminar 5369S: Knowledge Graphs, written during summer 2023. As of the date of writing and to the best of my knowledge, this work has not been published, and as such, cannot be properly referenced.

While all KGs are graphs, not all graphs are KGs. However the line between the two is often blurry, and the distinction is not always clear. The term "knowledge graph" first appeared in 1973, but really gained popularity through a 2012 blog post about Google's Knowledge Graph [81]. Several definition attempts have been made, but none of them are universally accepted. Below are listed some of the most common definitions of Knowledge Graphs:

- «A graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities.» [78]
- «A graph of data consisting of semantically described entities and relations of different types that are integrated from different sources. Entities have a unique identifier. KG entities and relations are semantically described using an ontology or, more clearly, an ontological representation.» [80]

In KGs, edges are often labeled and may represent complex relationships like "is a subclass of" or "is married to," allowing for more expressive power. The very nature of KG makes any definition attempt difficult. Indeed, KG is a broad concept that can be applied to many different domains, use cases and can have diverse implementations. The definition of KGs is thus very context-dependent, and it's debatable where the line between a graph and a KG is drawn.

For the purpose of this thesis, we won't focus on this distinction and just consider that we deal with memory *graphs*, which are graphs that are not necessarily KGs, but that can be used to represent complex relationships between entities extracted from memory dumps and be leveraged using KG-related techniques for advance tasks like feature engineering, automated embedding, inductive reasoning and learning.

#### **2.2.2.4 Inductive Reasoning and Learning**

Inductive reasoning in Knowledge Graphs (KGs) involves techniques like embedding and Graph Convolutional Networks (GCNs) to learn the potential underlying structure of the graph. This is particularly useful for tasks like link prediction, node classification, and clustering.

For the sake of clarity and grouping, we will present the concepts of graph embedding and GCNs in the next sections, but it is important to note that they are not mutually exclusive. In fact, GCNs can be used to generate embeddings, and embeddings can be used as input for GCNs. As such, they are often used together in the context of KGs, or in our case, memory graphs.

## **2.3 Data preprocessing for Machine Learning**

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on the development of algorithms that can learn from data and make predictions. It is a powerful tool that has been used to solve a wide range of problems, including image classification, speech recognition, and natural language processing. Before we can apply machine learning algorithms to a dataset, we must first prepare the data by performing various preprocessing steps. In this section, we will discuss the most common data preprocessing techniques, including data cleaning, feature engineering, and dataset splitting, as well as the importance of feature selection and dimensionality reduction. All those elements are crucial for the development of effective machine learning models for key extraction.



### 2.3.1 Feature engineering

In the realm of machine learning and data science, *features* refer to individual measurable attributes or characteristics of the phenomena under study. Those features can be of different types, and can be used to predict the value of a target variable.

#### 2.3.1.1 Types of Features

Features can be of different types, depending on the nature of the data. The type of feature determines how it is processed and utilized by the model. The most common types of features include:

- **Numerical Features:** These are quantitative attributes representing measurements like height, weight, or age.
- **Categorical Features:** These are qualitative attributes representing discrete classes or labels, such as gender (Male, Female) or educational level (High School, Bachelor's, Master's).
- **Ordinal Features:** Similar to categorical features but with an inherent order, like ratings on a scale of 1 to 5.
- **Text Features:** These contain textual data and often require special preprocessing steps like tokenization and vectorization.
- **Temporal Features:** These are time-based attributes, such as timestamps, requiring special handling to capture time-dependent patterns.
- **Geospatial Features:** These attributes represent geographical or spatial coordinates.

Features are pivotal for the performance of machine learning models. The quality and pertinence of features can significantly influence the model's capability to discern underlying patterns in the data. Inadequately chosen or irrelevant features can lead to a poorly performing model, while carefully selected, relevant features can result in a robust and accurate model.

#### 2.3.1.2 The Curse of Dimensionality

The *curse of dimensionality* refers to a set of challenges that arise when dealing with high-dimensional data. As the number of features, or dimensions, in a dataset increases, the volume of the feature space grows exponentially. This exponential growth leads to several issues:

- **Data Sparsity:** In high-dimensional spaces, data points tend to be sparse, making it difficult for algorithms to identify patterns. The sparsity also means that the notion of "distance" becomes less meaningful, which is problematic for algorithms that rely on distance metrics.
- **Computational Complexity:** The exponential increase in volume demands significantly more computational power and memory, making it challenging to process and analyze the data efficiently.
- **Overfitting:** High dimensionality increases the risk of overfitting, where a model learns the noise in the data rather than the actual pattern. Overfit models perform poorly on unseen data.

- **Statistical Significance:** As dimensions increase, the amount of data required to achieve statistical significance also increases exponentially, often making it impractical to collect sufficient data.
- **Visualization and Interpretability:** High-dimensional data are difficult to visualize and interpret, making it challenging to derive intuitive insights.

Due to these challenges, many dimensionality reduction techniques have been developed to transform high-dimensional data into a lower-dimensional form, aiming to preserve as much of the relevant information as possible. These techniques are discussed in the next section.

### 2.3.1.3 Feature Engineering techniques

The meticulous process of selecting the most relevant features, or constructing new features from existing ones, is known as *feature engineering*. This step can encompass normalization, transformation, and the creation of interaction terms among features. This complex process requires a deep understanding of the domain of study and the datasets, and is often a crucial step in the development of machine learning models [17].

In the context of Machine Learning, features essentially serve as the input variables  $X$  that a machine learning model employs to make predictions or inferences about the output variable  $Y$ . But not all features are equally informative. Some features may be redundant, irrelevant, or even detrimental to the model's performance. Irrelevant features are those that do not contribute to the predictive power of the model, while redundant features are those that are highly correlated with other features. Feature engineering techniques, can be used to build, transform or eliminate features, thereby reducing the dimensionality of the data and enhancing model performance [16].

- **Scaling and Normalization:** Scaling and normalization are techniques used to transform the features to a similar scale. This is particularly important for algorithms that rely on distance metrics, such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). Scaling and normalization can also help accelerate the training process by reducing the number of iterations required for the model to converge. Some common techniques include min-max scaling, z-score normalization, and log transformation.
- **Feature Extraction:** This technique involves transforming the original set of features into a new set of features, which is usually of lower dimensionality. The new features are often combinations of the original features and aim to capture as much of the information in the original data as possible. Many methods exist, like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are often employed to transform high-dimensional data into a lower-dimensional form are commonly used for feature extraction [16].
- **Feature Selection:** Unlike feature extraction, feature selection aims to pick a subset of the most important features from the original set, without changing them [16]. The goal is to remove irrelevant or redundant features that do not contribute significantly to the model's performance. Techniques like Recursive Feature Elimination (RFE) and using importance scores from tree-based algorithms like Random Forest are popular methods for feature selection.

Both techniques have their own advantages and disadvantages, and the choice between the two often depends on the specific requirements of the task at hand. Feature extraction is generally

more suitable when the original features do not have much interpretability to begin with, or when transforming features can lead to a more compact and effective representation. On the other hand, feature selection is often preferred when it is important to maintain the interpretability of the features, or when computational efficiency is a concern [17].

#### 2.3.1.4 Evaluating Features with Correlation Tests

To assess the quality of features, various statistical measures can be employed. Correlation tests are statistical tests that measure the strength and direction of the relationship between two variables. Pearson, Kendall, and Spearman correlation coefficients are commonly used to quantify the linear or monotonic relationship between each feature and the target variable [20]. A high absolute value indicates a strong relationship, aiding in feature selection.

- **Pearson Correlation:** Measures the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a strong negative linear correlation, 1 indicates a strong positive linear correlation, and 0 indicates no linear correlation.
- **Kendall's Tau:** A non-parametric test that measures the strength and direction of a monotonic relationship between two variables.
- **Spearman's Rank:** Also a non-parametric test, it assesses how well an arbitrary monotonic function can describe the relationship between two variables without making any assumptions about the frequency distribution.

These techniques are useful for evaluating the relationship between each feature and allows to generate correlation matrices, which can be used to identify redundant features. It's also possible to evaluate each feature independently through univariate feature selection techniques. In Python's scikit-learn library [19], methods like *F-test* and the *p-value* are often used for this purpose. "

- **F-test value:** Measures the linear dependency between the feature variable and the target. A higher F-test value indicates a more useful feature.
- **p-value:** Indicates the probability of an F-test value this large arising if the null hypothesis is true. A smaller p-value suggests rejecting the null hypothesis, making the feature significant.

In summary, features are the foundational elements of any machine learning model. The quality of these features, along with how they are processed and utilized, can markedly impact the model's performance. The significance of feature engineering cannot be overstated. Properly engineered features can drastically reduce modeling errors, leading to more accurate and reliable predictions. It serves as a bridge between raw data and predictive models, ensuring that the models are fed with the most relevant and informative features.

### 2.3.2 Embeddings

Embeddings are low-dimensional vector representations of high-dimensional objects. They are often used to capture complex relationships between objects and are particularly useful for machine learning tasks like clustering, classification, and link prediction. Embeddings are

widely used in the field of natural language processing (NLP) to represent words, sentences, and documents [56]. In recent years, they have also been applied to graphs to learn node representations [78]. In this section, we will discuss the concept of embeddings and explore some common techniques for creating them.

### 2.3.2.1 Embedding Creation vs Feature Engineering

Both embedding creation and feature engineering are techniques used to transform data for machine learning models. However, they are different in terms of their goals and how they are achieved.

- **Embedding Creation:** Embedding creation is the process of learning a low-dimensional vector representation of a high-dimensional object. This involves mapping discrete objects, such as words in Natural Language Processing (NLP) or nodes in a graph in our case, to vectors of continuous values in a lower-dimensional space [56]. The goal is to capture the semantic or structural relationships between these objects. Specialized algorithms like Word2Vec for word embeddings or Node2Vec for graph embeddings are often used.
- **Feature Engineering:** As discussed before, it is a more general practice that involves creating new features or modifying existing ones to improve the performance of a machine learning model. While feature engineering can include creating embeddings, it also encompasses a wide range of other techniques like normalization, transformation, outlier detection, and handling missing values.

In general, feature engineering is a more manual process, while embedding creation is more automated. Feature engineering requires domain knowledge and understanding of the problem, while embedding creation can be done using a variety of machine learning techniques.

Thus, embedding creation is a specialized form of feature engineering aimed at mapping discrete objects to continuous vectors of numbers, usually for capturing complex relationships. Feature engineering, on the other hand, is a broader practice that can involve a variety of techniques, including but not limited to embedding creation.

### 2.3.2.2 Embeddings for graphs

Graph embedding techniques aim to map nodes and edges in a graph to vectors in a low-dimensional space [78]. The primary goal is to preserve the graph's structural properties, such as node connectivity and community structure, in the embedded space. These vectors can then be used for various machine learning tasks like clustering, classification, and link prediction. A quite complete overview of the different techniques can be found in „Knowledge Graphs“. Here are some common and advanced techniques:

**Translational Models** The first type of graph embeddings techniques is based around using translational models that interpret edge labels as transformations from subject nodes to object nodes „Knowledge Graphs“.

- **TransE:** is one of the earliest and most straightforward translational models. It represents entities as points in a vector space and relations as translations between these points. The

primary idea is that for a valid triple  $(h, r, t)$ , the equation  $h + r = t$  should hold, where  $h$  stands for the head entity,  $r$  represents the relation, and  $t$  is the tail entity. This model is simple and computationally efficient but is limited in its ability to capture complex relationships [78].

- **TransH:** TransH extends TransE by introducing relation-specific hyperplanes. This allows the model to capture more complex relationships by projecting the entity embeddings onto these hyperplanes before performing translations [78].
- **TransR:** TransR goes a step further by not only introducing relation-specific hyperplanes but also relation-specific translations. This allows for a more flexible representation of relations, accommodating various types of complexities [78].
- Other improvements include *TransD* or *MuRP*, which shows that research in this domain is still very active.

**Tensor Decomposition Models** Tensor decomposition models represent entities and relations as vectors or matrices in a low-dimensional space. These models are based on the assumption that the relationship between entities can be represented by a bilinear function. They are computationally efficient and can capture complex relationships, but they are also limited in their ability to model asymmetric and reflexive relations.

- **RESCAL:** RESCAL employs a bilinear model where each relation is represented by a full-rank matrix. This allows for capturing asymmetric and reflexive relations but at the cost of increased computational complexity [78].
- **DistMult:** DistMult simplifies RESCAL by assuming that the relation matrices are diagonal. This reduces the number of parameters and computational complexity, making it more scalable [78] [79].
- **Complex:** Complex extends DistMult by introducing complex-valued embeddings. This allows the model to capture asymmetric relations effectively while maintaining computational efficiency [78] [79].

**Neural Models** Neural models employ neural networks to learn the features of entities and relations. This provides a more flexible and adaptive approach to graph embeddings.

- **ConvKB:** ConvKB employs a convolutional neural network to automatically learn the features of entities and relations. It is technically a translational model as introduced before, but uses a convolutional layer to capture the interactions between entities and relations. This provides a more flexible and adaptive approach to graph embeddings [26].
- **RotatE:** RotatE uses complex rotations in the embedding space to model relations. This neural model captures the semantics of relations in a more expressive manner [79].
- **SDNE (Structural Deep Network Embedding):** SDNE employs a deep autoencoder to learn complex and non-linear node embeddings while preserving first-order and second-order proximities. It is particularly effective for capturing intricate patterns and structures in the graph [27].
- **R-GCN:** Relational Graph Convolutional Networks (R-GCNs) combine the strengths of GCNs and traditional embedding methods to capture both topological and semantic information

[28]. A more recent study by Degraeve et al. argues that the main contribution of R-GCN lies in its message passing paradigm rather than the learned weights. This paper introduces a variant called Random R-GCN (RR-GCN) [40].

- **ConvE:** ConvE employs convolutional layers to capture local and global interactions between entities and relations, offering a more expressive representation [79].

**Language Models** Language models utilize pre-trained language models to enrich the embeddings with contextual information. This approach leverages the recent developments of language models to capture the semantics of entities and relations.

- **BERT for KGE:** Utilizing pre-trained BERT models, this approach leverages the power of language models to enrich the embeddings with contextual information [29].
- **BART KGE:** Bidirectional and Auto-Regressive Transformers (BART) is a denoising autoencoder that can be used for various NLP tasks. This approach utilizes BART to learn entity and relation embeddings. The paper introducing this also compares other LLM like GPT-2 [30].
- Due to the recent developments in NLP, this domain is still very active, and new approaches are being developed regularly.

### 2.3.2.3 Word Embeddings

Word embeddings are vector representations of words in a low-dimensional space. They are often used as input for machine learning models in natural language processing (NLP) tasks like text classification, sentiment analysis, and machine translation. Word embeddings are typically learned from large text corpora using unsupervised learning techniques like Word2Vec and GloVe. These embeddings can then be used to capture semantic relationships between words and phrases, which is particularly useful for NLP tasks [56].

- **Word2Vec:** Word2Vec is a popular algorithm for learning word embeddings from text data. It employs a shallow neural network to learn the embeddings and is often used as a pre-processing step for NLP tasks [79].
- **GloVe:** Global Vectors for Word Representation (GloVe) is another popular algorithm for learning word embeddings. It is based on the co-occurrence matrix of words and utilizes matrix factorization to learn the embeddings [79].

**Graph Embeddings** Graph embeddings are vector representations of nodes in a graph. Graph embeddings are typically learned using unsupervised learning techniques like Node2Vec and DeepWalk. These embeddings can then be used to capture structural relationships between nodes, which is particularly useful for graph analytics tasks [78].

- **One-Hot Encoding:** This is a simple technique that represents each node as a vector of 0s and 1s, where the length of the vector is equal to the number of nodes in the graph. The vector contains a 1 at the index corresponding to the node and 0s everywhere else. As is, this method is not really suitable for large graphs as it results in a high-dimensional and sparse matrix representation [79].

- **Node2Vec:** This algorithm learns continuous feature representations for nodes by optimizing a neighborhood-preserving objective. It employs biased random walks and uses the Skip-gram model to generate embeddings. Node2Vec is particularly effective for capturing local structures and can be fine-tuned for specific tasks [31].
- **DeepWalk:** Similar to Node2Vec, DeepWalk uses random walks to generate node sequences. It employs the Skip-gram model from natural language processing to learn embeddings. Unlike Node2Vec, it does not use biased walks, making it more suitable for capturing global structures [41].
- **Spectral Clustering:** This technique is based on the spectral theory of graphs. It utilizes the eigenvalues and eigenvectors of the Laplacian matrix of the graph to find an optimal embedding. Spectral Clustering is particularly useful for community detection and can capture the global structure of the graph [47] [49].
- **LINE:** Large-scale Information Network Embedding (LINE) aims to preserve both local and global network structures. It optimizes two objectives: first-order and second-order proximities between nodes. LINE is scalable and can handle large graphs efficiently [32].
- **Graph Factorization:** This method directly factorizes the adjacency matrix of the graph to learn node embeddings, making it computationally efficient but less capable of capturing complex structures. It is often used for large-scale graphs where computational resources are limited [79].

### 2.3.3 Other preprocessing techniques for Machine Learning

When working with real-world data and datasets, it is common to issues like missing values, outliers, and imbalanced classes that can adversely affect the performance of machine learning models. In this section, we will discuss some common preprocessing techniques used to improve the quality of data and ensure that the models can learn effectively.

#### 2.3.3.1 Data Cleaning

Data cleaning is the process of detecting and correcting errors in the data. It is a crucial step in data preprocessing and involves various techniques like outlier detection, handling missing values, and data normalization. Data cleaning is necessary to ensure that the data is accurate and consistent, which is essential for machine learning models to learn effectively.

- **Outlier Detection:** Outliers are data points that deviate significantly from the rest of the dataset. They can be caused by errors in data collection or genuine variations. Outliers can skew the model's learning and should be identified and handled appropriately, either by removal or transformation.
- **Handling Missing Values:** Missing data can lead to biased or incorrect model training. Techniques for handling missing values include imputation, where missing values are replaced with statistical measures like mean, median, or mode, and deletion, where rows with missing values are removed.
- **Data Normalization:** Features with different scales can affect the performance of machine learning algorithms. Normalization rescales the features to a standard range, usually  $[0, 1]$ , or transforms them to have a mean of 0 and a standard deviation of 1.

- **Encoding Categorical Variables:** Many machine learning algorithms require numerical input and output variables. Categorical variables are converted to numerical format through techniques like one-hot encoding or label encoding.
- **Text Cleaning:** In natural language processing tasks, text data may require cleaning to remove irrelevant characters, correct typos, or standardize text format.
- **Duplicate Removal:** Duplicate entries can bias the model and should be identified and removed from the dataset.
- **Feature Engineering:** While not strictly data cleaning, feature engineering involves transforming existing features or creating new ones to improve model performance.

### 2.3.3.2 Dataset splitting and sampling

One other typical step is the division of the dataset into training and testing sets. This separation is crucial for evaluating the generalization performance of a model. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. Failing to separate these sets can lead to overfitting, where the model performs well on the training data but poorly on new, unseen data.

Various techniques exist for dataset splitting and sampling, each with its own advantages and disadvantages:

- **Random Split:** The dataset is randomly divided into training and testing sets based on a given ratio, such as 70% for training and 30% for testing. This method is simple but may result in imbalanced classes in the splits.
- **Stratified Split:** Similar to random split, but ensures that the distribution of classes is the same in both training and testing sets. This is particularly useful for imbalanced datasets.
- **k-Fold Cross-Validation:** The dataset is divided into 'k' subsets or "folds." The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time with a different fold as the testing set. The average performance metric is used for evaluation.
- **Leave-One-Out Cross-Validation (LOOCV):** A special case of k-Fold Cross-Validation where k is equal to the number of data points. Each data point is used once as the test set while the remaining points form the training set.
- **Bootstrapping:** Random samples are drawn with replacement from the dataset to create multiple training sets. The model is trained and tested on these sets, and the average performance is calculated.

### 2.3.3.3 Dealing with Imbalanced Datasets

Imbalanced datasets are those where the classes are not represented equally. This is a common issue in machine learning, especially in classification problems. An imbalanced dataset can lead to a biased model that may not effectively predict the minority class. Therefore, it is crucial to address this issue during data preprocessing.



- **Why it is Necessary:** In imbalanced datasets, machine learning algorithms tend to be biased towards the majority class, ignoring the minority class. This results in poor classification performance for the minority class, which is often the class of interest in problems like fraud detection, medical diagnosis, etc.
- **Oversampling:** This involves adding more copies of the minority class. Oversampling can be random with replacement, or it can involve generating synthetic samples.
- **Undersampling:** This involves removing some of the samples of the majority class. This method is generally not preferred as it can lead to loss of data.
- **SMOTE (Synthetic Minority Over-sampling Technique):** SMOTE is an oversampling method that creates synthetic samples in the feature space. It selects two or more similar instances (using a distance measure) and perturbing an instance one at a time by a random amount within the difference to the neighboring instances.
- **ADASYN (Adaptive Synthetic Sampling):** Similar to SMOTE, but it uses a weighted distribution for different minority class examples according to their level of difficulty in learning.
- **Cost-sensitive Learning:** This involves modifying the algorithm to increase the weight of the minority class during training, thereby making the algorithm more sensitive to it.
- **Ensemble Methods:** Methods like Random Forest and Gradient Boosting can be used with techniques like bagging and boosting to handle imbalanced datasets effectively.
- **Resampling Methods:** These involve randomly partitioning the data into subsets, balancing each subset, and then aggregating the results.

To conclude this section, data preprocessing is a crucial step in machine learning. The number of concepts and techniques discussed here is by no means exhaustive, which further highlights the importance and complexity of data preprocessing.

## 2.4 Machine Learning and Deep Learning

In the preceding sections, we have discussed the importance of data preprocessing and feature engineering for machine learning. In this section, we will discuss the machine learning pipeline and explore some common machine learning algorithms. We will also discuss deep learning and neural networks especially in the context of graphs, since they have gained popularity in recent years due to their superior performance on many tasks.

### 2.4.1 Machine Learning

Machine learning is a cornerstone in the field of artificial intelligence and has been instrumental in driving many of today's technological and scientific breakthroughs. From natural language processing to computer vision, machine learning algorithms play a critical role in making sense of large and complex data sets.

### 2.4.1.1 What is Machine Learning

Machine learning is a subfield of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. This learning process is based on the recognition of complex patterns in data and the making of intelligent decisions based on them [21].

The machine learning pipeline typically involves several steps including some pre-steps like data collection, preprocessing, feature extraction that we have already discussed before, usually followed by model training, evaluation, and deployment. Algorithms are trained on a dataset, and the learned patterns are used to make predictions or decisions without human intervention. ML algorithms can be broadly classified into three categories [21]:

- **Supervised Learning:** Algorithms are trained on labeled data, and the aim is to make predictions or map inputs to outputs.
- **Unsupervised Learning:** Algorithms are trained on unlabeled data, focusing on the underlying structure or distribution in the data.
- **Reinforcement Learning:** Algorithms learn to perform an action from state to state to maximize some type of reward or objective function.

In the following, we will focus on supervised and unsupervised learning, as they are the most relevant to our work.

### 2.4.1.2 Model Evaluation

Evaluating the performance of a machine learning model is crucial for understanding its effectiveness and suitability for a given task. It is not a trivial task as it depends on various factors like the type of data, the problem at hand, and the model itself [55]. In this section, we will discuss some common evaluation metrics for classification and regression models.

**Precision** Precision is the ratio of correctly predicted positive observations to the total predicted positives. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

**Recall** Recall is the ratio of correctly predicted positive observations to all the observations in the actual class. The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

**F-1 Score** The F-1 Score is the weighted average of Precision and Recall, and it ranges from 0 to 1. The formula for the F-1 Score is:

$$\text{F-1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Accuracy** Accuracy is the ratio of correctly predicted observations to the total observations. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}}$$

These are the most common metrics used for evaluating classification models. Other metrics like the Area Under the Curve (AUC) and the Receiver Operating Characteristic (ROC) curve are also used for evaluating binary classification models. For regression models, metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are commonly used [19].

## 2.4.2 Machine Learning Models for Binary Classification

Binary classification is a supervised learning task where the goal is to predict a binary outcome, i.e., one of two possible classes (0 or 1). This is a truly classical task for a ML model, with a range of applicable algorithms. As such, we rely on Python's Scikit-learn library for implementation.

**Logistic Regression** A specialized form of regression tailored for predicting binary outcomes. It employs the logistic function to map predicted values between 0 and 1. While straightforward and interpretable, its performance may be limited on complex, non-linear data [22].

The logistic function is defined as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Where  $\beta_0$  is the intercept and  $\beta_1$  is the coefficient for the predictor variable  $x$ .

**Decision Trees** These are versatile models used for both classification and regression. They partition the feature space into regions, making decisions at each node. While easy to visualize, they are susceptible to overfitting [23].

Decision Trees use metrics like Gini impurity or entropy to make splits:

$$\text{Gini}(T) = 1 - \sum_{i=1}^C p_i^2$$

Where  $T$  is a node and  $p_i$  is the proportion of class  $i$  instances among the training instances in node  $T$ .

**Random Forest** An ensemble technique that aggregates predictions from multiple decision trees. Known for its robustness and ability to handle large, high-dimensional data [24].

The final prediction is an average or majority vote from all trees:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$$

Where  $\hat{y}$  is the final prediction and  $\hat{y}_i$  is the prediction from the  $i^{th}$  tree.

**Support Vector Machines (SVM)** Effective for both classification and regression, SVMs find the hyperplane that best separates the data into classes. They excel in high-dimensional spaces [54].

The objective is to maximize the margin between classes:

$$\max_{w,b} \frac{2}{\|w\|}$$

Where  $w$  is the weight vector and  $b$  is the bias term.

**k-Nearest Neighbors (k-NN)** An instance-based algorithm that classifies a new point based on the majority class among its 'k' nearest neighbors [51].

The distance between points is often calculated using Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Where  $d(x, y)$  is the distance between points  $x$  and  $y$ , and  $n$  is the number of dimensions.

### 2.4.3 Deep Learning

Deep learning offers a powerful set of tools for automatically generating embeddings from the graph representation of heap dumps. Leveraging neural networks, we can build custom models using PyTorch to perform binary classification tasks, such as predicting key nodes in a heap dump. In this section, we will explore various neural network architectures that are well-suited for this task.

Neural Networks serve as the backbone of deep learning models [53]. They consist of interconnected nodes or "neurons" organized into layers. Deep learning extends this architecture by employing multiple hidden layers, enabling the model to learn complex representations from data. This is particularly useful when dealing with graph-based heap dump representations, where the relationships between nodes can be intricate [79].

Choosing Deep Learning models is dependent on the task. As such, we need to present the different types of neural networks and their advantages and disadvantages. Below are some common neural network architectures:

#### 2.4.3.1 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) are particularly effective for sequence-based data. In the context of heap dumps, if the memory addresses or keys exhibit some form of sequential pattern, RNNs can be employed to capture these temporal dependencies for binary classification [53, p. 10.2].

- **Advantages:**

- Good at capturing short-term dependencies in sequence data.
- Relatively simpler architecture.

- **Disadvantages:**

- Struggles with long-term dependencies due to the vanishing gradient problem.
- May require more data for effective training.

#### 2.4.3.2 Long Short-Term Memory (LSTM)

LSTMs are an extension of RNNs designed to capture long-term dependencies, making them suitable for more complex sequences [52].

- **Advantages:**

- Effective in capturing long-term dependencies.
- Less susceptible to the vanishing gradient problem.

- **Disadvantages:**

- More complex and computationally intensive than RNNs.
- May require fine-tuning of hyperparameters.

#### 2.4.3.3 Gated Recurrent Units (GRU)

Gated Recurrent Units (GRU) offer a compromise between the simplicity of RNNs and the power of LSTMs. They are effective in capturing both short-term and long-term dependencies but with a less complex architecture [50].

- **Advantages:**

- Simpler architecture compared to LSTM.
- Efficient in capturing long-term dependencies.

- **Disadvantages:**

- May not perform as well as LSTMs for very complex sequences.
- Still more computationally intensive than basic RNNs.

#### 2.4.3.4 Convolutional Neural Networks (CNN)

Though traditionally used in image processing, CNNs can also be adapted for sequence data like heap dumps. They are excellent at identifying spatial hierarchies or patterns in the data [53] [25].

- **Advantages:**

- Highly effective in identifying local patterns.
- Less prone to overfitting due to pooling layers.

- **Disadvantages:**

- May not capture global dependencies as effectively as RNNs or LSTMs.
- Requires a fixed-size input.

#### 2.4.3.5 Graph Convolutional Networks (GCN)

Graph Convolutional Networks (GCNs) are a specialized form of neural networks designed to work directly with graphs [79]. They are particularly useful for our task as they can automatically generate embeddings from the graph representation of heap dumps. These embeddings can then be used for binary classification to predict key nodes.

- **Advantages:**

- Capable of capturing both local and global graph structures.
- No need for manual feature extraction from graphs.

- **Disadvantages:**

- May require fine-tuning and a well-defined graph structure.
- Computationally intensive for large graphs.

Several Python libraries can be used to implement GCNs for our specific task. Some of the notable ones are PyTorch Geometric, Spektral, and DGL (Deep Graph Library). These libraries offer pre-built GCN layers and various utilities to facilitate the embedding and binary classification of heap dump graphs.

#### 2.4.4 Graph Neural Networks

A Graph Neural Network (GNN) constructs a neural network that mirrors the structure of the underlying data graph. In this architecture, nodes are linked according to their relationships in the data graph. The model is trained in a supervised fashion to map input features of nodes to their corresponding output features [79]. These output features can either be manually annotated or sourced from a knowledge graph.

Unlike traditional knowledge graph embeddings, GNNs offer the advantage of end-to-end supervised learning tailored for specific tasks.

With a dataset of labeled examples, GNNs can classify individual nodes or even entire graphs. The challenge lies both in how to convert a given graph to a format that can be fed into a neural network (see subsection 2.3.2.2) and how to design a neural network that can effectively learn from the graph data. Note that the distinction between the two is not always clear since two problems are often tackled together. As such, the present classification is not exhaustive and arbitrary, even though it is based on the literature and meta-analysis of the field, like „Knowledge Graphs (Extended)“ [79] and „A comprehensive survey on graph neural networks“ [43].

GNNs have been employed in various applications, ranging from traffic prediction, recommender systems, and software verification [43]. Remarkably, GNNs can also serve as substitutes for conventional graph algorithms. For instance, they have been used to identify central nodes in knowledge graphs through supervised learning [79]. The following sections delve into two specific types of GNNs: Recursive GNNs and Convolutional GNNs.

#### 2.4.4.1 Recursive Graph Neural Networks (RecGNNs)

Recursive Graph Neural Networks (RecGNNs) serve as the foundational approach to graph neural networks [79]. The model operates by passing messages between neighboring nodes to recursively compute results. The framework learns the functions that generate the expected output based on a training set of labeled nodes. Scarselli et al. [46] proposed a seminal GNN model that uses feature vectors for nodes and edges, along with state vectors for nodes. Two parametric functions, the transition function and the output function, are used to update the state vectors and compute the final output for nodes, respectively. These functions are applied recursively until a fixpoint is reached. The model is highly flexible and can be adapted in various ways, such as defining neighboring nodes differently or using distinct parameters for each node [46].

**Learning Process in GNNs** The learning process in GNNs can be divided into three steps: input computation, node state update, and output computation. These steps are repeated until a fixpoint is reached, and the final output is computed for each node. Essentially, this involves finding the optimal parameters  $w$  such as for  $\varphi_w$  to best approximate the data in the learning dataset  $\mathcal{L}$ .

The learning task can be formulated as the minimization of a quadratic cost function  $e_w$  through iterative gradient descent [46]:

$$e_w = \sum_{i=1}^p \sum_{j=1}^{q_i} (t_{i,j} - \varphi_w(G_i, n_{i,j}))^2$$

Where:

- $\mathcal{G}$ : The set of graphs.
- $\mathcal{N}$ : The subset of nodes.
- $\mathcal{D} = \mathcal{G} \times \mathcal{N}$ : The set of pairs of a graph and nodes.
- $G_i = (N_i, E_i) \in \mathcal{G}$ : The graph.

- $N_i$ : The set of nodes in graph  $G_i$ .
- $E_i$ : The set of edges in graph  $G_i$ .
- $n_{i,j} \in N_i$ : The  $j^{th}$  node of the graph  $G_i$ .
- $q_i \leq |N_i|$ : The number of supervised nodes in the graph  $G_i$ .
- $p \leq |\mathcal{G}|$ : The number of graphs in the learning set.
- $\mathcal{L} = \{(G_i, n_{i,j}, t_{i,j})\}$ : The learning set.
- $t_{i,j} \in \mathbb{R}^m$ : The target output for node  $n_{i,j}$ .
- $m$ : The number of outputs.
- $\mathbb{R}^m$ : The  $m$ -dimensional Euclidean space.
- $\varphi_w : \mathcal{D} \rightarrow \mathbb{R}^m$ : The function that maps the input to the output.
- $e_w$ : Error function.

For graph-focused tasks, a special node is used for the target ( $t$ ), whereas for node-focused tasks, supervision can be performed on every node. More information about the learning process can be found in [46].

#### 2.4.4.2 Convolutional Graph Neural Networks (ConvGNNs)

As introduced before, Graph Convolutional Networks extend the concept of convolution from images to graphs. «The core idea in the image setting is to apply small kernels (aka filters) over localised regions of an image using a convolution operator to extract features from that local region.» [79]. They are designed to work with non-Euclidean data and are particularly useful for semi-supervised learning tasks on graphs. GCNs aim to learn a function that maps nodes to a low-dimensional space while considering their local neighborhood and features.

**Convolution filters** A Convolutional Neural Network (CNN) is a type of neural network that uses convolutional filters to extract features from images. These filters are applied to local regions of the image to capture spatial patterns [34]. The same concept can be extended to graphs, where the convolution filters are applied to local regions of the graph to capture structural patterns.

The convolution operation on an image is historically defined as follows [39] [48]:

$$y_{i,j} = \sum_{k=-m}^m \sum_{l=-n}^n x_{i+k,j+l} \times w_{kl}$$

Where:

- $y_{i,j}$ : The output value at the coordinate  $(i, j)$  of the convolution output.



- $x_{i,j}$ : The input value at the coordinate  $(i, j)$  of the input image.
- $w_{kl}$ : The weight value (coefficient) of the kernel at the coordinate  $(k, l)$ .
- $m$ : The kernel width.
- $n$ : The kernel height.

Similar to this, ConvGNNs implement the transition function using convolutions. One of the challenges in ConvGNNs is defining regions of a graph, as nodes in a graph may have varying numbers of neighbors. Solutions to this problem involve using spectral or spatial representations of graphs [49] [79]. An alternative approach employs attention mechanisms to learn the most important features [33]. Unlike RecGNNs, ConvGNNs apply a fixed number of convolutional layers and can use different kernels/weights at each distinct step.

Here are some common ConvGNN architectures. I recommend reading „Beyond low-pass filtering: Graph convolutional networks with automatic filtering“ [44] for a more in-depth discussion of these models:

- **Vanilla GCN**: The basic GCN model consists of an input layer, one or more hidden layers, and an output layer. Each layer is associated with a graph convolution operation that updates the node features based on their neighbors [79].
- **GraphSAGE (Graph Sample and Aggregation)**: This model generalizes the GCN framework by allowing various aggregation functions like mean, LSTM, and pooling to combine information from a node’s neighbors [42].
- **ChebNet**: ChebNet uses Chebyshev polynomials to generalize the convolution operation in the spectral domain. This allows the model to capture a broader range of graph structures [43].
- **GAT (Graph Attention Networks)**: GAT introduces attention mechanisms into GCNs, enabling the model to weigh neighbors differently when aggregating information [33].
- **MoNet**: This model employs a mixture model to generalize the convolution operation, making it capable of handling graphs with diverse structures [45].
- **Graph Isomorphism Network (GIN)**: GIN is designed to capture the isomorphism between different graphs, making it powerful for tasks like graph classification [44].
- Research on the topic is still ongoing, and a lot of new architectures are being proposed: Cluster-GCN, TAGCN, AutoGCN, and many more [44]. It’s worth noting that many of these models are not mutually exclusive and can be combined to create more powerful models.

The background chapter has introduced a lot of concepts and techniques that are relevant to our work. We have discussed the importance of data preprocessing and feature engineering, explored various machine learning algorithms, and looked at some common neural network architectures. Although not exhaustive, this overview should provide a solid foundation for the methods, design and implementation that will be discussed in the next chapters.

## 3 Related Work

Now that the necessary background has been established, this chapter will present the related work in the field of machine learning for memory forensics in the context of OpenSSH. The chapter is divided into two parts. The first part will present the related work in the field of memory forensics, and the second part will present the related work in the field of machine learning for memory forensics.

### 3.1 SSHKex

SSHKex is a research project that aims to address the challenges of analyzing encrypted SSH traffic by leveraging Virtual Machine Introspection (VMI) techniques. Developed by Sentanoe and Reiser, the project focuses on extracting SSH keys and decrypting SSH network traffic in a stealthy, non-intrusive manner while maintaining evidence integrity [1]. This paper is itself a continuation of the work presented in „Sarracenia: Enhancing the Performance and Stealthiness of SSH Honeypots Using Virtual Machine Introspection“ [8], which introduced Sarracenia, a high-interaction SSH honeypot. It is also related to a range of other research projects and papers [1, section 5.6 and 6].

The SSHKex approach combines standard network traffic capturing methods with dynamic SSH session key extraction. It assumes that the SSH implementation running on the server is known, which is crucial for the key extraction process. The project employs VMI tools like LibVMI and Volatility to gain a complete and untainted view of all guest VM’s state information. This allows to efficiently locate SSH session keys in the main memory of a Linux machine.

Here is a summary of the SSHKex methodology for key extraction:

1. **Data Structure Information:** The method leverages detailed knowledge about the data structures used to store the keys. Specific debugging symbols corresponding to the SSH implementation version on the target system provide essential offset values to facilitate the extraction of key material. The structures of interest include `struct ssh`, `struct session_state`, `struct newkeys`, and `struct sshenc`. These structures store a range of information such as IP addresses, ports, session states, and encryption keys.
2. **Tracing OpenSSH Functions:** Function tracing is employed to identify the precise locations of data structures and to extract keys at the right time. The focus is on two key functions: `kex_derive_keys` (which initiates key generation) and `do_authentication2` (which kicks off user authentication).
3. **Breakpoints Injection:** Software breakpoints are intentionally placed in the program execution to facilitate debugging. SSHKex utilizes Virtual Machine Introspection (VMI) to inject these breakpoints at the initial points of the two aforementioned key functions.
4. **Key Extraction:** Upon calling the `kex_derive_keys` function, SSHKex initially stores the address of the `ssh struct`. The actual keys are extracted from memory when the `do_authentication2` function is subsequently called, adhering to the known structures.
5. **Key Indexing:** OpenSSH stores client-to-server and server-to-client keys in distinct indices of the `newkeys` structure. SSHKex extracts keys based on these specific indices.

6. **Handling Multiple Connections:** To manage multiple SSH connections, OpenSSH spawns child processes. SSHKex extends its key extraction strategy to each child process by identifying them through their unique process IDs.

One of the key strengths of SSHKex is its focus on stealthiness, preservation, and evidence integrity. The approach aims to be as unobtrusive as possible, avoiding any modifications to the system under investigation. This is particularly important in forensic contexts, where the integrity of the evidence is crucial [1].

## 3.2 SmartKex

SmartKex is a direct followup project that focuses on the extraction of SSH keys from heap memory dumps. Its primary objective is to automate the process of SSH key extraction from heap memory dumps. The project introduces a machine learning-assisted methodology that significantly improves the efficiency and accuracy of key extraction compared to traditional brute-force methods. This method is also significantly more straightforward to implement compared to the previous SSHKex approach, which requires detailed knowledge of the SSH implementation and the ability to inject breakpoints into the program execution.

SmartKex discusses two distinct methods for SSH key extraction:

- **Brute-Force Baseline Method:** This is a traditional approach that scans through the heap memory to identify potential keys based on known patterns.
- **Machine Learning-Assisted Method:** This approach uses a Random Forest algorithm trained on a highly imbalanced dataset using SMOTE balancing. The machine learning model is designed to identify SSH keys with high precision and recall rates, but is not exact as compared to the brute-force method since it is based on a probabilistic model.

### 3.2.1 Baseline brute-force method

Here is a summary of SmartKex's brute-force method for SSH key extraction from heap dumps [2]:

1. **Heap Dump Generation:** Heap dump binary files of OpenSSH server process have been generated (ASK HOW) and serves as the input for the key extraction process. The exact process and architecture is not described in SmartKex paper, but we suppose it was done on a *linux-x86\_64* architecture.
2. **Data Reduction:** To minimize the heap size, the method removes memory pages that are irrelevant (empty) based on Hamming distance.
3. **Brute-force key search:** Starting from the first byte, a key length of 128 bytes is taken from the heap dump as the potential key. The algorithm iterates over the entire heap, continuously updating the potential key until the heap's end is reached.
4. **Decryption Attempt:** For every potential key, an attempt is made to decrypt network packets. If decryption fails, the process is repeated with a new potential key.

Although the brute-force approach is exact, it is computationally expensive. It performs poorly especially when keys are located at the end of the heap dump [2, section 6.2].

### 3.2.2 SmartKex machine-learning method

The real innovation of SmartKex is its machine learning-assisted methodology for SSH key extraction. At the cost of exactness, this approach is significantly faster than the brute-force method and has a high degree of accuracy in identifying encryption keys. It also allows for the heap size to be reduced to less than 2% of its original size, further optimizing the extraction process.

Here is a summary of SmartKex's machine learning-assisted method for SSH key extraction from heap dumps [2]:

1. **Heap Dump inputs:** Similarly to the brute-force method, heap dump binary files of OpenSSH also serve as inputs for the key extraction process.
2. **Preprocessing:** The raw heap dump is resized into an  $N \times 8$  matrix. High entropy parts of the heap dump, which are likely to be encryption keys, are identified using the logical AND operation on the vertical and horizontal differences of adjacent bytes. This creates an array that flags potential key locations.
3. **Training:** A Random Forest algorithm is trained on 128-byte slices of the preprocessed heap. The dataset is imbalanced, with the slices that contain keys being rare. A stacked classifier approach is used, comprising a high precision classifier and a high recall classifier.
4. **Key Identification:** The machine learning model is used to predict which 128-byte slices of the heap dump are likely to contain encryption keys. These slices are then subjected to a brute-force method to actually extract the keys.

SmartKex is significantly faster than the brute-force method alone and has a high degree of accuracy in identifying encryption keys. It also allows for the heap size to be reduced to less than 2% of its original size, further optimizing the extraction process.

SmartKex has broad applications in the field of cybersecurity, particularly in memory forensics. Its machine learning-assisted methodology can be adapted for other types of sensitive data extraction, making it a versatile tool for researchers and practitioners alike. The project is open-source, with the code available on GitHub <sup>1</sup>.

## 3.3 Objectives of the present work

This Masterarbeit can be seen as a direct followup to the paper „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“. The present work aims to improve the SmartKex methodology by exploring new machine learning architectures and algorithms. The goal is to improve the accuracy of the machine learning model and to reduce the computational complexity of the overall process.

To do so, this work has significantly broaden the research area by exploring entirely new ways to deal with the dataset by leveraging memory graph representation, feature engineering, new machine learning and deep learning model architectures, and new training strategies. A range of different tools and script, with a focus on code quality and reproducibility with careful packaging

---

<sup>1</sup><https://github.com/smartvmi/Smart-and-Naive-SSH-Key-Extraction>

using Nix ensure that the present research can be easily extended and reproduced by other researchers.

# 4 Methods

In the preceding chapters, all the necessary background knowledge to understand the methods have been introduced. In this chapter, we will present an overview of the challenges, the methods, and the tools we have developed for this thesis. We will first describe the dataset we have used. Then, we will describe the programs developed for this thesis. Finally, we will describe how we have packaged and deployed our programs with Nix.

## 4.1 OpenSSH memory dumps dataset

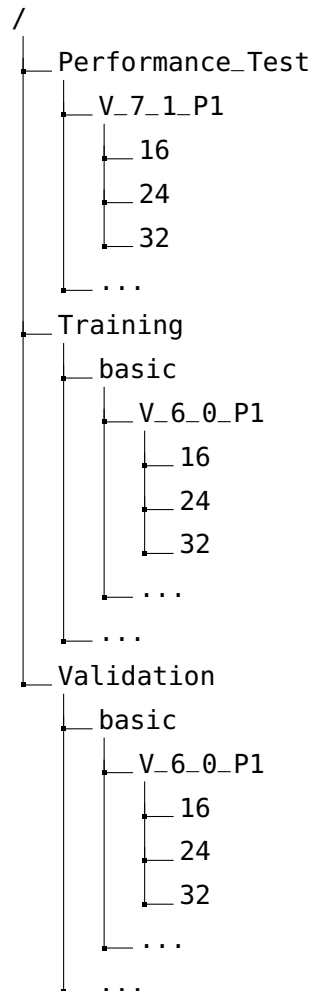
SmartKex has contributed to the research community by generating a comprehensive annotated dataset of OpenSSH heap memory dumps [2]. The dataset is publicly available on Zenodo <sup>1</sup>.

---

<sup>1</sup><https://zenodo.org/record/6537904>

The dataset is organized into two top-level directories: *Training* and *Validation* with an additional *Performance\_Test*. The first two main directories are further divided based on the SSH scenario, such as immediate exit, port-forward, secure copy, and shared connection. Each of these subdirectories is then categorized by the software version that generated the memory dump. Within these, the heaps are organized based on their key lengths, providing a multi-layered structure that aids in specific research queries.

Figure 4.1: Illustration of the Dataset Directory Structure



Two primary file formats are used to store the data: JSON and RAW. The JSON files contain meta-information like the encryption method, virtual memory address of the key, and the key's value in hexadecimal representation. The RAW files, on the other hand, contain the actual heap dump of the OpenSSH process.

Here is an example of content of a RAW memory dump file, displayed using *vim* and *xxd* commands:

```

1      00000000: 0000 0000 0000 0000 5102 0000 0000 0000 .....Q.....
2      00000010: 0607 0707 0707 0303 0200 0006 0401 0206 .....
3      00000020: 0200 0001 0100 0107 0604 0100 0000 0203 .....
4      00000030: 0103 0101 0000 0000 0000 0000 0000 0002 .....
5      00000040: 0001 0000 0000 0000 0000 0100 0000 0001 .....
6      00000050: 8022 1a3a 3456 0000 007f 1a3a 3456 0000 .":4V...:4V..
7      00000060: f040 1a3a 3456 0000 9032 1a3a 3456 0000 .@.:4V...2.:4V..
8      00000070: 608b 1a3a 3456 0000 9047 1a3a 3456 0000 '...:4V...G.:4V..

```

Listing 4.1: 16 bytes per line visualization of a Hex Dump from *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*

The original file contains the raw byte content of the heap dump of a specific version of OpenSSH. It is a binary file, which means that it is not human-readable. However, it can be converted to a human-readable format using the *xxd* command. The first column to the left represents the offset in hexadecimal. The last column represents the actual content of the bytes, in ASCII format. The columns in between represent the content of the bytes in hexadecimal format.

Since hexadecimal is a base-16 number system, each byte is represented by two hexadecimal digits. The ASCII representation of the bytes is displayed on the right, and is only used for reference, as it is not always possible to convert the bytes to ASCII. For instance, the bytes at offset 0x10 are not printable characters, and thus cannot be converted to ASCII. Each line represents 16 bytes, and the offset is incremented by 16 for each line.

For the purpose of this thesis, it will be more interesting to visualize the content of the heap dump as 8 bytes lines. This can be achieved by using the *xxd* command with the *-c* option, as shown in the following example:

The same example as before, a memory dump file, displayed using *vim* and *xxd -c 8* commands:

```

1      00000000: 0000 0000 0000 0000 .....
2      00000008: 5102 0000 0000 0000 Q.....
3      00000010: 0607 0707 0707 0303 .....
4      00000018: 0200 0006 0401 0206 .....
5      00000020: 0200 0001 0100 0107 .....
6      00000028: 0604 0100 0000 0203 .....
7      00000030: 0103 0101 0000 0000 .....
8      00000038: 0000 0000 0000 0002 .....
9      00000040: 0001 0000 0000 0000 .....
10     00000048: 0000 0100 0000 0001 .....
11     00000050: 8022 1a3a 3456 0000 .":4V..
12     00000058: 007f 1a3a 3456 0000 ...:4V..
13     00000060: f040 1a3a 3456 0000 .@.:4V..
14     00000068: 9032 1a3a 3456 0000 .2.:4V..
15     00000070: 608b 1a3a 3456 0000 '...:4V..
16     00000078: 9047 1a3a 3456 0000 .G.:4V..

```

Listing 4.2: 8 bytes per line visualization of a Hex Dump from *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*

This example shows the exact content of the preceding one.

To this RAW file is associated a JSON file, which contains its annotations.



Here is a example of content of a JSON annotation file that comes with the previous RAW file:

```
1      {
2        "SSH_PID": "5070",
3        "SSH_STRUCT_ADDR": "56343a1a4800",
4        "session_state_OFFSET": "0",
5        "SESSION_STATE_ADDR": "56343a1a8d30",
6        "newkeys_OFFSET": "344",
7        "NEWKEYS_1_ADDR": "56343a1aaa40",
8        "NEWKEYS_2_ADDR": "56343a1aab40",
9        "enc_KEY_OFFSET": "0",
10       "mac_KEY_OFFSET": "48",
11       "name_ENCRYPTION_KEY_OFFSET": "0",
12       "ENCRYPTION_KEY_1_NAME_ADDR": "56343a1a9db0",
13       "ENCRYPTION_KEY_1_NAME": "aes128-gcm@openssh.com",
14       "ENCRYPTION_KEY_2_NAME_ADDR": "56343a1a3fb0",
15       "ENCRYPTION_KEY_2_NAME": "aes128-gcm@openssh.com",
16       "key_ENCRYPTION_KEY_OFFSET": "32",
17       "key_len_ENCRYPTION_KEY_OFFSET": "20",
18       "iv_ENCRYPTION_KEY_OFFSET": "40",
19       "iv_len_ENCRYPTION_KEY_OFFSET": "24",
20       "KEY_A_ADDR": "56343a1a3170",
21       "KEY_A_LEN": "12",
22       "KEY_A_REAL_LEN": "12",
23       "KEY_A": "feb5fd4ef0759b034d69b858",
24       "KEY_B_ADDR": "56343a1a33e0",
25       "KEY_B_LEN": "12",
26       "KEY_B_REAL_LEN": "12",
27       "KEY_B": "f50b988297fa19709445c4ee",
28       "KEY_C_ADDR": "56343a1aa1b0",
29       "KEY_C_LEN": "16",
30       "KEY_C_REAL_LEN": "16",
31       "KEY_C": "f5b53280e944db0fe196668d877cd4c0",
32       "KEY_D_ADDR": "56343a1a4010",
33       "KEY_D_LEN": "16",
34       "KEY_D_REAL_LEN": "16",
35       "KEY_D": "ac4f18a963d9e72c857497b7dc9d088d",
36       "KEY_E_ADDR": "56343a1a7d90",
37       "KEY_E_LEN": "0",
38       "KEY_E_REAL_LEN": "0",
39       "KEY_E": "",
40       "KEY_F_ADDR": "56343a1a2f60",
41       "KEY_F_LEN": "0",
42       "KEY_F_REAL_LEN": "0",
43       "KEY_F": "",
44       "HEAP_START": "56343a198000"
45     }
```

Listing 4.3: Complete JSON example, from *Training/basic/V\_7\_8\_P1/16/5070-1643978841.json*

Those annotation files contain the meta-information about the heap dump, such as the encryption method, virtual memory address of the key, and the key's value in hexadecimal representation. Those annotations are invaluable for the development of machine learning models for key extraction.

The dataset is not just limited to SSH key extraction; it also serves as a resource for identifying essential data structures that hold sensitive information. This makes it a versatile tool for various research applications, including but not limited to machine learning models for key extraction and

malware detection.

#### 4.1.1 Assumptions

Before we dive in, let's make some assumptions about the dataset. We will use these assumptions to guide our exploration of the heap dump file.

- **Heap dump file size:** We will assume that the heap dump file size is a multiple of 8 bytes. This is because the heap dump file is a memory dump, and memory is allocated in chunks that are multiples of 8 bytes. This means that we can expect the heap dump file to be composed of a sequence of 8 bytes blocks. If this assumption is not met, we will assume that padding the last block with 0s will not change the results of our exploration.
- **Malloc header chaining:** We will assume that all the heap dump files have been generated using the same malloc implementation. It means that we can expect to find the same patterns in all the heap dump files. Especially, we expect all the heap dump files to start by zero or more blocks of 8 bytes of 0s, followed by the first malloc header. We can then follow the malloc header chaining to explore the heap dump file allocated memory chunks.
- **Dataset key annotation format:** We will assume that the JSON annotation files have been generated using the same program. This means that we can expect the same format for all the JSON annotation files. This is important, as we will use the JSON annotation files to get the key addresses for annotating memory graphs used for the embedding step. If the format is not the same, we will assume that the JSON annotation file is corrupted, and we will skip it.

In the scripts and programs that have been developed for the following thesis, we have implemented a number of checks and tests to ensure that these assumptions are met. If not, the programs will raise an error, log the problem and generally skip the data. This behavior is implemented to ensure that the programs are robust to unexpected data, and to ensure that the results are reliable. These assumptions, and related problems will be discussed and measured at several locations in the following sections.

#### 4.1.2 Estimating the dataset balancing for key prediction

First, let's quickly estimate what the dataset is composed about. This will later be used to estimate the balancing of data for our key prediction goal. Some quick linux commands can be used to get a general overview of the dataset.

A first command can quickly give us an idea of the number of files in the dataset:

```
1 find /path/to/dataset -type f | wc -l
```

Listing 4.4: Count all dataset files

Another command can be used to get the total size of the dataset:

```
1 du -sb /path/to/dataset
```

Listing 4.5: Get the total size of the dataset

The first command indicates that the dataset contains 208749 files, which represents, according to second one, a total of 18203592048 bytes, or around 18 Gigabytes.

We could just divide the number of files by the size of the dataset to get an average size of the files. However, this would not be accurate, as we are only interested in the size of the RAW files. Since JSON files are much smaller than RAW files, they would skew the average size of the files. Since we are only considering RAW files, we will use improved commands in order to determine the size of the RAW file only.

The following command can be used to get a better understanding of the dataset, concerning the number of RAW files and their size:

```
1 find /path/to/dataset -type f -name "*.RAW" | wc -l
```

Listing 4.6: Find the number of RAW files in the dataset

And the next one can be used to get the number of bytes of RAW files in the dataset:

```
1 find /path/to/dataset -type f -name "*.raw" -exec du -b {} + \  
2 | awk '{s+=$1} END {print s}'
```

Listing 4.7: Find the number of bytes of RAW files in the dataset

Where:

- `find phdtrack_data/ -type f -name "*.raw"` finds all the files in the dataset that have the extension `.raw`.
- `-exec du -b {} + | awk '{s+=$1} END {print s}'` executes the command `du -b` on each file found by the previous command, and sums the size of each file.

Theses commands indicate that the dataset contains 103595 RAW files, which represents a total of 18067001344 bytes, or around 18 Gigabytes. This shows that the vast majority of the data is contained in RAW files, with JSON files representing less than a percent of the dataset in term of byte size. As such, the average size for every RAW file is around 170 Kilobytes.

Now, considering that a given heap dump file is expecting to have only 6 keys (see subsubsection 2.1.1.2), with keys maximal possible size being of 64 bytes, we can estimate that we have at maximum 39780480 or around 40 Megabytes of positively labeled samples. This, considering the total useful size of around 18 Gigabytes, means that our dataset is very imbalanced, with an expected upper-bounded ratio of 0.0022% of positively labeled samples or around 2 : 1000.

Considering that, a frontal ML binary classification approach will not work. This is why the present report will discuss feature engineering and graph-based memory representation. The idea is to embed more information to our keys so as to be able to fight effectively the imbalanceness of the raw data.

### 4.1.3 Dataset validation

The dataset is nearly a collection of heap dump RAW files for different use cases and versions of OpenSSH. Each heap dump file goes along a JSON annotation file that has been generated by the creators of the dataset to provide additional information about the heap dump, and especially encryption keys.

However, it is worth noting that the dataset is not perfect, as some of the annotations are sometimes missing. This is probably due to the fact that the annotations were generated automatically, and some of the keys are missing. For instance, in *Training/basic/V\_7\_8\_P1/16/*, literally the first file of the dataset contains an incomplete annotation file, as some of the keys are missing. This is a limitation of the dataset that should be kept in mind when using it for research purposes.

Here is an example of content of a JSON annotation file with missing keys, and with missing annotations (like address or lenght) for the keys that are present:

```
1      {
2          "ENCRYPTION_KEY_NAME": "aes128-ctr",
3          "ENCRYPTION_KEY_LENGTH": "16",
4          "KEY_C": "689e549a80ce4be95d8b742e36a229bf",
5          "KEY_D": "76788e66a56d2b61eec294df37422fcb",
6          "HEAP_START": "5589d41e0000"
7      }
```

Listing 4.8: Missing keys in JSON annotation file *Training/basic/V\_6\_0\_P1/16/24375-1644243522.json*

#### 4.1.3.1 Automatic annotation validation

So as to determine really how much of the dataset can be really for machine learned, we have developped a script that checks the validity of thoses annotations. This script called `check_ - annotations.py`, is used to verify the quality, completeness, consistency and coherence of the annotations.

Files are regroup under the following categories:

- **Correct and Complete Files:** Files that have no missing keys, and that have all the keys with correct values.
- **Broken Files:** Files that are not valid JSON files, and cannot be loaded as such.
- **Incorrect Files:** Files that have contradictory information in their annotation file.
- **Missing key Files:** Files that have missing keys in their annotation file. A typical example is a JSON file with `"KEY_E": ""`. This means that the key E is missing, and that the key E address is not present in the annotation file, which is a problem for the machine learning since it means that we cannot label correctly the key E.
- **Incomplete key Files:** Files that have incomplete keys in their annotation file. A typical example is a JSON file with `"KEY_E": "689e549a80ce4be95d8b742e36a229bf"`. This means that the key E is present, but that the key E address is not present in the annotation file, which is a problem for the machine learning since it means that we cannot label correctly the key E.

The script is used to validate each JSON file using the following process:

---

**Algorithm 1** Json Annotation Validation

---

```
1: procedure ValidateJson(json_data)
2:   Initialize errors = Dictionary{}           ▷ Serve as collection for counted errors
3:   Initialize mandatory_json_keys = ['HEAP_START', 'SSH_STRUCT_ADDR', 'SESSION_-
    STATE_ADDR']
4:   Initialize key_names = {}
5:   Initialize incorrect_keys, missing_keys, incomplete_keys = 0
6:   for mandatory_json_key in mandatory_json_keys do ▷ Check if some expected json keys are
    missing
7:     if mandatory_json_key not in json_data or not correct hex address then
8:       errors[mandatory_json_key] = False
9:     else
10:      errors[mandatory_json_key] = True
11:    end if
12:  end for
13:  for json_key in json_data.keys() do           ▷ Get all the keys names, like A, B, C, D, E, F
14:    if json_key.startswith("KEY_") then
15:      key_name = GetLetterOfSSHKeyFromJSONKeyName(json_key)
16:      key_names.add(key_name)
17:    end if
18:  end for
19:  for key_letter in key_names do
20:    base_key = "KEY_" + key_letter
21:    PerformSSHKeyAnnotationValidationAndCompleteness(base_key, json_data) that
    counts incorrect_keys, missing_keys, incomplete_keys
22:  end for
23:  Store counters in errors
24:  return errors
25: end procedure
```

---

The counting error algorithm done on each SSH key annotation by is described in the following:

---

**Algorithm 2** SSH Key Annotation Validation

---

```
1: procedure PerformSSHKeyAnnotationValidationAndCompleteness(base_key, json_data)
2:   Initialize incorrect_keys, missing_keys, incomplete_keys = 0
3:   if length(json_data[base_key]) == 0 then
4:     missing_keys += 1 ▷ missing key
5:   else
6:     is_key_len_present = exists(json_data[base_key_LEN])
7:     is_key_addr_present = exists(json_data[base_key_ADDR])
8:     is_key_real_len_present = exists(json_data[base_key_REAL_LEN])
9:     if not is_key_len_present or not is_key_addr_present or not is_key_real_len_present then
10:      incomplete_keys += 1 ▷ Incomplete keys
11:      Generate and store error message about missing annotations
12:     else if not is_hex_address_correct(json_data[base_key_ADDR]) then
13:      incorrect_keys += 1 ▷ Incorrect address
14:      Generate and store error message about incorrect address
15:     else if json_data[base_key_LEN] is not a number or is negative then
16:      incorrect_keys += 1 ▷ Incorrect length
17:      Generate and store error message about incorrect length
18:     else
19:       Validate key value length based on annotation length
20:       if json_data[base_key_LEN] == 0 then
21:         missing_keys += 1 ▷ missing key
22:       else if length(json_data[base_key]) != json_data[base_key_LEN] * 2 then
23:         incorrect_keys += 1 ▷ contradictory length
24:         Generate and store error message about incorrect key value length
25:       end if
26:     end if
27:   end if
28:   return incorrect_keys, missing_keys, incomplete_keys
29: end procedure
```

---

Note that I have simplified this algorithm. The `is_hex_address_correct` function requires other manipulations to be called, since it checks that the given value is in the range of the heap dump addresses. To do so, it requires handling potentially missing `HEAP_START` annotation, hexadecimal conversion with correct endianness, and other manipulations like determining the size of the heap dump. The full code is available in the `check_annotations.py` file.

The script runs in a few seconds on all the 103595 JSON annotation files, and give the following results:

- **Number of Correct and Complete Files:** 26196 files
- **Number of Broken Files:** 6 files are broken. A direct look at those files shows that they are empty files.
- **Number of Incorrect Files:** 0 files
- **Number of Missing key Files:** 58643 files have missing keys.
- **Number of Incomplete key Files:** 18750 files have incomplete keys.

We can also directly look at the the keys in general:

- **Number of SSH keys:** 546534 keys
- **Number of missing (empty) SSH keys:** 157244 keys
- **Number of incompletely annotated SSH keys:** 37500 keys
- **Number of incorrectly annotated SSH keys:** 0 keys

In total, this means that only 25.3% of the JSON files are actually usable (correct and complete), and can be used for machine learning. This is because we don't have access to the packets that have been used to generate the dataset, and thus we cannot regenerate the annotations. Since the machine learning relies entirely on those annotation, we cannot afford to use partially annotated files.

This is a limitation of the dataset that should be kept in mind when using it for research purposes, and especially for supervised machine learning.

#### 4.1.4 Exploring patterns in RAW heap dump files

Before diving into programming, we need to gain a better understanding of how to retrieve useful information from heap dump raw file. For that matter, we will continue to experiment with simple commands in RAW heap dump files. Note that in the following, number bases are indicated, since endianness and conversions can get confusing.

Let's start by looking back at the RAW file we already presented in Listing 4.2.

##### 4.1.4.1 Detecting potential pointers

The paper „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ indicates that the keys are 8-bytes aligned. In fact, this is the case for the whole heap dump file. This is why we have chosen to split the study of heap dump files in chunks or blocks of 8 bytes. The term *block* in code is always referring to this, unless specified otherwise. The precision is important, since these blocks should not be confused with *memory blocks* like the ones that are allocated by the `malloc()` function in C.

Let's re-open the heap dump file in vim, and let's use the following vim commands to explore the example heap dump file:

- `:%xxd -c 8 5070-1643978841-heap.raw!`: This vim command converts the opened file to a hex dump. The `-c 8` option indicates that we want to display 8 bytes per line.
- `:set hlsearch`: This vim command highlights the search results.
- `:%s/\s\+//g`: This vim command removes all the whitespaces in the file.
- `:%s/\v([0-9a-f]{8}):)\1\` This vim command adds a whitespace after each 8 byte addresses.
- `:%s/\v(([0-9a-f]{8}: )([0-9a-f]{16}))\1\` This vim command adds a whitespace after each heap dump byte line.

To find potential pointers, we can use the following command in vim:

```
:/[0-9a-f]\{12\}0\{4\}
```

Listing 4.9: Vim command to find potential pointers

This is a search that looks for 12 hexadecimal digits followed by 4 zeros. This is because, the maximum possible addresses in the heap dump file have a size of around 12 hexadecimal digits, and because pointer addresses are in little-endian format, meaning that the last 4 bytes of the address are also the Most Significant Bytes (MSB) of the address.

The result is illustrated below:

```
00000000: 00000000000000000000 . . . . .
00000008: 51020000000000000000 Q . . . . .
00000010: 0607070707070303 . . . . .
00000018: 0200000604010206 . . . . .
00000020: 0200000101000107 . . . . .
00000028: 06040100000000203 . . . . .
00000030: 0103010100000000 . . . . .
00000038: 000000000000000002 . . . . .
00000040: 0001000000000000 . . . . .
00000048: 0000010000000001 . . . . .
00000050: 80221a3a34560000 .":4V..
00000058: 007f1a3a34560000 ...:4V..
00000060: f0401a3a34560000 .@.:4V..
00000068: 90321a3a34560000 .2.:4V..
00000070: 608b1a3a34560000 `.:4V..
00000078: 90471a3a34560000 .G.:4V..
```

We have information about the starting address of the heap using "HEAP\_START": "56343a198000". Considering that the example heap dump file contains 135169 bytes, this means that for this given heap dump file, the pointer addresses range from value 94782313037824<sub>10</sub> and 94782313172993<sub>10</sub>. Note that the little-endian hexadecimal representation of the heap end address is 0x01901b3a3456 which is 12 character long, or 6 bytes long.

Note that conversions here can get confusing, since potential pointer strings extracted from the heap dump file are given in little-endian hexadecimal format, but the heap start address from the



JSON annotation file is given in big-endian hexadecimal format.

That way, we can refine the detection of potential pointers by only considering the bytes that are in the range of the heap. Potential pointers are highlighted with "<<<" in the following hex dump:

```
1      # conversion from hex to decimal
2      def hex_str_to_int(hex_str: str) -> int:
3          """
4          Convert a normal (big-endian) hex string to an int.
5          WARNING: HEAP_START in JSON is big-endian.
6          """
7          bytes_from_str = bytes.fromhex(hex_str)
8          return int.from_bytes(
9              bytes_from_str, byteorder='big', signed=False
10         )
11
12     def pointer_str_to_int(hex_str: str) -> int:
13         """
14         Convert a pointer hex string to an int.
15         WARNING: Pointer hex strings are little-endian.
16         """
17         bytes_from_str = bytes.fromhex(hex_str)
18         return int.from_bytes(
19             bytes_from_str, byteorder='little', signed=False
20         )
```

Listing 4.10: Conversions function from hex strings to decimal *int* values.

Using the functions above, we can check which potential pointers are indeed within the heap dump range.

That way, we can refine the detection of potential pointers. In the following, pointers are highlighted with <<< in the following hex dump:

```
1      00000000: 0000000000000000 .....
2      00000008: 5102000000000000 Q.....
3      00000010: 0607070707070303 .....
4      00000018: 0200000604010206 .....
5      00000020: 0200000101000107 .....
6      00000028: 0604010000000203 .....
7      00000030: 0103010100000000 .....
8      00000038: 0000000000000002 .....
9      00000040: 0001000000000000 .....
10     00000048: 0000010000000001 .....
11     00000050: 80221a3a34560000 .".:4V.. <<<
12     00000058: 007f1a3a34560000 ...:4V..
13     00000060: f0401a3a34560000 .@.:4V.. <<<
14     00000068: 90321a3a34560000 .2.:4V.. <<<
15     00000070: 608b1a3a34560000 '..:4V.. <<<
16     00000078: 90471a3a34560000 .G.:4V.. <<<
```

Listing 4.11: 8 bytes per line visualization of a Hex Dump from *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*

One last check we can do, is verify if the potential pointers are 8-bytes aligned. This can be done by checking if the last 3 bits of the potential address are 0, or using a modulo 8 operation. A simple python function can be used to check that:

```

1      def is_pointer_aligned(pointer: int) -> bool:
2          """
3              Check if a pointer is 8-bytes aligned.
4              """
5          return pointer % 8 == 0

```

Listing 4.12: Python function to check if a potential pointer is 8-bytes aligned

Using this function on the potential pointers we have found so far, we can see that all of them are indeed 8-bytes aligned. This is a good sign for pointer detection, as we now have a range of tests that can be used to detect potential pointers from other potentially random values.

Here is the pseudo-code for the pointer detection algorithm. This algorithm is presented for a full heap dump file:

---

**Algorithm 3** Pointer Detection Algorithm

---

```

1: procedure PointerDetection(heapDumpFile, HEAP_START)
2:   heapStart  $\leftarrow$  hex_str_to_int(HEAP_START)
3:   heapEnd  $\leftarrow$  heapStart + FileSize(heapDumpFile)
4:   position  $\leftarrow$  0
5:   potentialPointers  $\leftarrow$  []
6:   while position < FileSize(heapDumpFile) do
7:     block  $\leftarrow$  Read8Bytes(heapDumpFile, position)
8:     if block  $\neq$  0 then
9:       pointer  $\leftarrow$  pointer_str_to_int(block)
10:      if heapStart  $\leq$  pointer  $\leq$  heapEnd then
11:        if is_pointer_aligned(pointer) then
12:          Append(pointer, potentialPointers)
13:        end if
14:      end if
15:    end if
16:    position  $\leftarrow$  position + 8
17:  end while
18:  return potentialPointers
19: end procedure

```

---

This pseudo-code outlines the steps for detecting potential pointers in the heap dump file. It starts by reading the heap dump file 8 bytes at a time. For each 8-byte block, it checks if the block is non-zero and within the heap range. If so, it checks if the potential pointer is 8-bytes aligned using the `is_pointer_aligned` function we described before. If all conditions are met, the potential pointer is added to the list of potential pointers. The algorithm returns this list at the end.

#### 4.1.4.2 Detecting potential keys

Encryption key prediction is the main focus of the present thesis. As such, we will now focus on how to detect potential keys in heap dump files. The paper „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ introduces 2 algorithms for key detection. The first one is a brute force approach that consists in trying all the possible keys in the heap dump file.

The second one is a more sophisticated approach that uses a set of rules to detect potential keys.

The first brute-force algorithm is given by:

---

**Algorithm 4** SSH keys brute-force algorithm from „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ [2]

---

```

1: procedure FindIVAndKey(netPacket, heapDump)
2:   ivLen  $\leftarrow$  16                                ▷ Based on the encryption method
3:   keyLen  $\leftarrow$  24                                ▷ Based on the encryption method
4:    $i \leftarrow \text{sizeof}(\text{cleanHeapDump})$ 
5:    $r \leftarrow 0$ 
6:   while  $r < i$  do
7:     pIV  $\leftarrow \text{heapDump}[r : r + \text{ivLen}]$ 
8:      $x \leftarrow 0$ 
9:     while  $x < i$  do
10:      pKey  $\leftarrow \text{heapDump}[x : x + \text{keyLen}]$ 
11:       $f \leftarrow \text{decrypt}(\text{netPacket}, \text{pIV}, \text{pKey})$ 
12:      if  $f$  is TRUE then
13:        return pIV, pKey
14:      end if
15:       $x \leftarrow x + 8$                                 ▷ The IV is 8-bytes aligned
16:    end while
17:     $r \leftarrow x + 8$                                 ▷ The key is 8-bytes aligned
18:  end while
19: end procedure

```

---

Algorithm 1 outlines the brute-force approach for finding the Initialization Vector (IV) and the key. Initially, the lengths ivLen and keyLen are set based on the encryption method used for the heap. The algorithm then takes the first ivLen bytes from the heap dump to serve as the potential IV ( $pIV$ ). Subsequently, keyLen bytes are extracted from the heap dump, starting from the first byte, to act as the potential key ( $pKey$ ). The algorithm iterates through this potential key until it reaches the end of the heap dump. If decryption of the network packet is unsuccessful, the process is repeated by reading the next potential IV and the subsequent potential key [2].

This algorithm is fairly straightforward, and can be implemented in a few lines of code. However, it is also very inefficient, as it tries all the possible keys in the heap dump file. It also needs some encrypted network packets to be able to test the keys, which are not included in the dataset. As such, we will not implement this algorithm.

This is why the authors of the paper have also developed a more sophisticated algorithm that uses a set of rules to detect potential keys.

No pseudo-code is given for the second algorithm, but the paper „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ gives a description of the algorithm. It rely on the high-entropy assumption of encryption keys. The algorithm is inspired by image processing techniques, and can be described as follows:

This Preprocessing Algorithm serves as a crucial step in adapting the heap data for machine learning models. The algorithm begins by reshaping the raw heap data into an  $N \times 8$  matrix  $X$ , since keys are 8-bytes aligned [2]. Here,  $N \times 8$  is the size of the original heap data in bytes. It then calculates the discrete differences of the bytes in both vertical and horizontal directions, storing the results in matrix  $Y$ . The algorithm employs a logical AND operation on these differences to

---

**Algorithm 5** Image-processing inspired Preprocessing Algorithm, as described in „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“ [2]

---

```

1: procedure Preprocessing(heapData)
2:   Reshape heapData into  $N \times 8$  matrix  $X$ 
3:   for  $i = 0$  to  $N - 1$  do
4:     for  $j = 0$  to  $7$  do
5:        $Y[i][j] = |X[i][j] - X[i][j + 1]| \& |X[i][j] - X[i + 1][j]|$ 
6:        $Z[i] = \text{count}(Y[i][k] == 0) \geq 4$ 
7:       if  $i < N - 1$  then
8:          $R[i] = Z[i] \& Z[i + 1]$ 
9:       end if
10:    end for
11:  end for
12:  Extract 128-byte slices from  $R$  for training
13: end procedure

```

---

identify high-entropy regions, which are likely candidates for encryption keys. Each 8-byte row in  $Y$  is examined for randomness, and if at least half of its bytes differ from adjacent bytes, it is marked as a potential part of an encryption key. The algorithm then filters out isolated rows that are unlikely to be part of an encryption key, resulting in an array  $R$ . Finally, 128-byte slices are extracted from  $R$  for training the machine learning model. This preprocessing step not only adapts the data for machine learning but also narrows down the search space for potential encryption keys, thereby enhancing the efficiency of the subsequent steps.

However, this algorithm is not as efficient as it could be. It rely on using a kind of sliding window, which is not easily parallelizable. Also, the entropy-inspired computation is not as straightforward as it could be. That why we propose a new algorithm that is more efficient and more easily parallelizable.

In order to perform some ML techniques, and because the keys we are looking for can have a range of possible lengths (16, 24, 32, or 64 bytes), we shift the focus from detecting the full key, to just be able to predict the address of the key. That way, we can deal with keys of different sizes, and we can also use the same algorithm to detect the IV. This is why we will focus on detecting potential keys addresses, and not the full keys.

We thus introduce a new algorithm for narrowing the search space for potential keys. This algorithm is inspired by the paper „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“, but is more efficient and more easily parallelizable, as it focuses on producing pairs of blocks of 8 bytes with high entropy. It uses directly the Shannon entropy formula, with each entropy computation being independent from the others.

---

**Algorithm 6** Entropy Based Detection of Potential Key blocks

---

```

1: procedure EntropyDetection(heapData)
2:   Pad heapData with 0s to be 8-bytes aligned
3:   Reshape heapData into  $N \times 8$  matrix  $X$ 
4:   for  $i = 0$  to  $N - 1$ , do
5:      $\text{entropy}[i] = \text{ShannonEntropy}(X[i])$  ▷ Independents, compute in parallel.
6:   end for
7:   Add  $\text{entropy}$  2 by 2 pairs into  $\text{entropy\_pairs}$  ▷ Keep block indexes in resulting tuples.
8:   Sort  $\text{entropy\_pairs}$  by entropy as  $\text{sorted\_pairs}$ 
9:   return SortedPairs(sorted_pairs)
10: end procedure

```

---

The *Entropy Based Detection of Potential Key blocks* algorithm takes a raw heap dump, represented by the variable `heapData`, as input. The data is first padded with zeros to align it to 8-byte blocks and then reshaped into an  $N \times 8$  matrix  $X$ . The Shannon entropy is computed for each 8-byte block in parallel, resulting in an array `entropy`. Subsequently, the entropy values of adjacent blocks are summed to form pairs, which are stored in `entropy_pairs` along with their block indexes. These pairs are then sorted by their entropy sums to produce `sorted_pairs`. The idea of using pairs of blocks instead of a single block or more than two blocks is related to the minimum key size, which is 16 bytes. This means that we need at least 2 blocks to be able to detect a potential key. The algorithm returns sorted pairs, so that we can easily extract the ones with the highest entropy sums. Given the index of a block, its actual memory address can be recomputed using the `HEAP_START` address available in annotations.

Using this algorithm, let's continue to explore our example heap dump file from Listing 4.2. We will use the following python function to compute the Shannon entropy of a given block of 8 bytes:

```

1  def get_entropy(data: bytes):
2      """
3      Computes the entropy of a byte array, using Shannon's formula.
4      """
5
6      if len(data) == 0:
7          return 0.0
8
9      # Count the occurrences of each byte value
10     _, counts = np.unique(data, return_counts=True)
11
12     # Calculate the probabilities
13     prob = counts / len(data)
14
15     # Calculate the entropy using Shannon's formula
16     entropy = -np.sum(prob * np.log2(prob))
17
18     return entropy

```

Listing 4.13: Python function to compute the Shannon entropy of a given block of 8 bytes

This function used np array function for efficient computation. We can now use this function to compute the entropy of each block of 8 bytes in the heap dump file. We can then sort the pair of blocks by their entropy, and keep the ones with the highest entropy.

When applied to the file *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*, the algorithm produced 16896 entropy pairs, with 631 pairs having the maximum entropy sum. Another test using the index to address conversion and the JSON annotation file also indicate that all of the 6 key addresses are within the 631 pairs with the highest entropy sum.

This allows to reduce significantly the search space for potential keys, to already less than 4% of the original heap dump file, which is significantly better than the 30% reduction obtained by the preprocessing algorithm described in the paper SmartKex [2], but less than the 2% reduction obtained by the ML-based processing algorithm described in the paper [2]. However the same paper indicated that it was tested only for Key A and Key C, whereas this algorithm is tested for all the keys. Keep in mind that this is just an example for a single random file in the dataset, as a way to introduce the subject. In-depth experiments will be done in the dedicated chapter on Machine Learning.

Indeed, it is important to mention that we can rely on the JSON annotation files for providing

labelling for key address prediction. Using this, we do not need to decrypt the network packets to be able to train our ML models. This is a huge advantage, and is also required since we don't have the encrypted network packets in the dataset. Since we don't have those, and since the keys are already given, that is why we will focus on key address prediction, and not on key prediction.

#### 4.1.4.3 Detecting potential data structures

Since the dataset contain whole heap dump file, we can also try to detect potential data structures in those heap dumps. This can be done by looking for patterns in the heap dump file, in a similar fashion as we have done for potential pointers.

Since OpenSSH is written in C, we can expect to find some C data structures in the heap dump files. C uses the `malloc` function to allocate memory. This function is used to allocate memory for a given data structure. This function takes as input the size of the data structure to allocate, and returns a pointer to the allocated memory. Looking at the code <sup>2</sup> for `malloc`, we can see the following:

```
1      /* Include extra space to record the size of the allocated block.  */
2      size += MALLOC_HEADER_SIZE;
```

Listing 4.14: Malloc implementation for uclibc

A `malloc` call is expected to allocate a memory block of the size of the data structure, plus 8 bytes (). This is because the `malloc` function also stores some metadata about the allocated memory block. This metadata is stored in the bytes that are allocated in addition to the data structure size. This is why we can expect to find some 8-bytes aligned blocks in the heap dump file, that are not pointers, but that are the result of a `malloc` call. Detecting and using those *malloc headers* is how we will also try to detect potential data structures in the heap dump file.

We don't know which implementation of `malloc` has been used for the OpenSSH programs used to generate the dataset. However, we can expect it to be similar to the one used in `uclibc`, a lightweight C library that is often used in embedded systems.

In Linux on a `x86_64` architecture, the `malloc` function typically uses a block (chunk) header to store metadata about each allocated block. This header is placed immediately before the block of memory returned to the user. The exact layout can vary depending on the implementation of the C library (e.g., `glibc`, `musl`), but generally, it contains the following:

- **Size of the Block:** The size of the allocated block, usually in bytes. This size often includes the size of the header itself and may be aligned to a multiple of 8 or 16 bytes.
- **Flags:** Various flags that indicate the status of the block, such as whether it is free or allocated, or whether the previous block is free or allocated. These flags are often stored in the least significant bits of the size field, taking advantage of the fact that the size is usually aligned, leaving the least significant bits zeroed.

Since the `malloc` header respects the endianness of the system, we can expect to find the `malloc` header in little-endian format in the heap dump file. Using `vim` on *Training/basic/V\_7\_-*

---

<sup>2</sup>The dataset has been produced on a `x86_64` architecture, so as an example code, we have looked at `uClibc` for a `malloc` implementation: <https://git.busybox.net/uClibc/tree/libc/stdlib/malloc/malloc.c>

*8\_P1/16/5070-1643978841-heap.raw*, we can use the following command to find some potential malloc headers:

```
1 :/[0-9a-f]\{4\}0\{12}
```

Listing 4.15: Vim command to find potential malloc headers

This gives something like the following:

```
00000000: 00000000000000000000 . . . . .
00000008: 510200000000000000 Q . . . . .
00000010: 0607070707070303 . . . . .
```

Figure 4.2: Attempt at malloc header detection in *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*, at heap start.

Indeed, after a first zero block of 8 bytes, we expect a first data structure to be allocated at the start of the heap. Here this data structure is of size  $5102000000000000_{16LE}$  (little-endian hex format) or  $593_{10}$  bytes. The fact that it is an odd number is due to the LSB being set to 1, to indicate that the block is allocated (flag). This means that the real size of the structure is actually  $593_{10} - 1_{10} = 592_{10}$ . This value is 8-byte aligned.

Since we know that the allocator allocates chunks after chunks, we can expect the next chunk to be allocated at the address  $5102000000000000_{16LE} + 592_{10} + 8_{10} = 5882193a34560000_{16LE} =$ . Note that we need to add 8 to the size to account for the malloc header block.

In vim, since the address start at 0, we have to look at  $592_{10} + 8_{10} = 258_{16}$ . Let's have a look there:

```
00000250: 000000000000000000 . . . . .
00000258: 210000000000000000 ! . . . . .
00000260: 7373686400000000 sshd . . .
00000268: 000000000000000000 . . . . .
00000270: 000000000000000000 . . . . .
00000278: 210000000000000000 ! . . . . .
```

Figure 4.3: Attempt at malloc header detection in *Training/basic/V\_7\_8\_P1/16/5070-1643978841-heap.raw*, at index  $592_{10} = 250_{16}$ .

There, we can see a zero block, followed by what we can expect to be another malloc header at address  $258_{16}$ . By doing the same process, we can thus propose an algorithm to detect the malloc headers, and thus the structures in the heap dump file.

---

**Algorithm 7** Malloc Header Detection Algorithm

---

```
1: procedure MallocHeaderDetection(heapDumpFile)
2:   position  $\leftarrow$  0
3:   while position < FileSize(heapDumpFile) do
4:     block  $\leftarrow$  Read8Bytes(heapDumpFile, position)
5:     if block  $\neq$  0 then
6:       size  $\leftarrow$  ConvertToSize(block) - 1  $\triangleright$  -1 due to the flag
7:       Assert size mod 8 = 0  $\triangleright$  Check if the size is 8-bytes aligned
8:       position  $\leftarrow$  position + 8 + size  $\triangleright$  Leap over data structure. + 8 for the header.
9:     else
10:      position  $\leftarrow$  position + 8
11:    end if
12:  end while
13: end procedure
```

---

The idea behind the malloc header detection algorithm is simple. We start at the beginning of the heap dump file, and we look for the first non-zero block. We then assume that the next block is a malloc header. We convert it to a size, and then leap over the data structure and the header. The process is repeated until reaching the end of the heap dump file.

## 4.2 Hardware and software architecture

Throughout this thesis, we have used a variety of hardware and software architectures.

### 4.2.1 Hardware development and testing environment

In this section, as a reference for the reader, we will describe shortly the hardware development environment. All environments are running some Linux x86\_64 distribution.

At the start of the project, around the end of 2022, the project started on a old laptop *HP EliteBook Folio 1040 G2*, running Ubuntu 22.04 LTS (Jammy Jellyfish) with the following specifications:

- **CPU:** 5th Generation Intel Core i7-5600U 2.6 GHz (max turbo frequency 3.2-GHz), 4 MB L3 Cache, 15W
- **GPU:** Intel HD Graphics 5500
- **RAM:** 8GB DDR3L SDRAM (1600 MHz, 1.3v)

This device was used for the first experiments, and for the development of the first programs. However, it was not powerful enough to run the experiments on the whole dataset, and especially working on ML part. As such, we have moved to a more powerful machine, a *TUXEDO InfinityBook Pro 16 - Gen7* with the following specifications:

- **CPU:** 12th Gen Intel i7-12700H (20) @ 4.600GHz



- **GPU:** NVIDIA Geforce RTX 3070 Ti Laptop GPU
- **RAM:** 64GB DDR5 4800MHz Samsung

For the Operating System, we have switched from Fedora 37 to NixOS 23 (Tapir). This change was motivated by the fact that NixOS is a Linux distribution that uses a purely functional package management system [38]. This means that the operating system is built by the Nix package manager, using a declarative configuration language. This allows to have a reproducible development environment, and to easily switch between different development environments. This has proved to be very useful in many areas like work environment isolation, on work collaboration with Clément Lahoche, and for software deployment to the server.

Unfortunately, the *TUXEDO InfinityBook Pro 16 - Gen7* laptop was not powerful enough to run the experiments on the whole dataset. Running the python script would have taken more than a week for some simple ML experiments to run on the whole dataset, and even reprogrammed and better optimized programs in Rust that we have tested would take more than 10 hours to process the dataset. Small bash and python scripts have been run on this laptop, but all the main experiments have been run on the server.

In that context, we were provided a development server towards the end of the thesis, around august 2023. This server is a *AS-4124GS-TNR* with the following specifications:

- **CPU:** 2x AMD EPYC 7662 (256) @ 2.000GHz
- **GPU:** NVIDIA Geforce RTX 3090 Ti
- **RAM:** 512GB DDR4 3200MHz

This server is running Ubuntu 20.04.6 LTS and is used to run the ML experiments on the dataset, as it is much more powerful than the laptop. This server has been provided by the Department of Computer Science of Universität Passau, and especially the Chair of Data Science of Prof. Dr. Michael Granitzer. We would like to thank them for their support.

#### 4.2.2 Software, languages and tools

In Computer Sciences, it doesn't take long to realize that testing hypotheses, diving deeper in problems and finding solutions to them is a very iterative process that requires a lot of experimentation. As such, the development of scripts and programs has been a substantial part of this thesis, from the very beginning to the very end. In this process, we have used a variety of tools and programming languages, such as Rust, Python, Bash, or Nix just to name the programming language used.

In this section, as a reference for the reader, we will describe the software architectures, languages and tools that have been used throughout this thesis.

Throughout the project, we have come to use a range of programming languages. Initial tests have been done using shell and bash command and simple scripts. However, as the project grew, we quickly moved to more powerful programming languages.

Python version 3.11 has been the main language for high level data science and ML development. This new version of python features many improvements over the previous version, and especially

in terms of performance. Better error messages, exception groups, improved support for f-strings, support for TOML configuration files, variadic generics, improved support for asyncio, and new modules for working with cryptography and machine learning are just some of the new features of this new version of python. While relatively new, this is why we have decided to use this version of python for the development of the ML part of the project.

Although Python is a popular and powerful language, it is not the most efficient language. As such, we have used Rust for some parts of the project, especially when no high level library is needed and when performances are critical to be able to parse efficiently the dataset. Rust is a systems programming language that runs blazingly fast, prevents segfaults, and guarantees thread safety. It is a very powerful language, and is especially useful for low-level programming. We have used it for the development of the algorithms that are used to extract the data from the dataset.

#### 4.2.2.1 Packaging and deployment

We made an extensive use of git repositories for version control, with GitHub as a main platform for hosting the repositories. An ever growing number of script and programs have been developed for this thesis. As such, we have needed a way to easily deploy those programs on different machines.

Rust comes with a handful of tools for managing packages and dependencies. Cargo is Rust's build system and package manager. Cargo downloads your Rust project's dependencies, compiles your project, makes executables, and runs the tests. It is a powerful tool that allows to easily manage Rust projects. However, it is not the best tool for deploying programs on different machines.

On Python's side of things, things are a bit more complicated. For a long time, we have relied on virtual environments using the conda package manager. However, it is heavy to use, and it doesn't allow to easily export an environment from one linux distribution to the other. This was especially a problem when development was done on a Fedora system but the server was running Ubuntu.

An example is the library `pygraphviz`. This library relies on third parties system libraries, that have different names depending on the linux distribution:

- **Ubuntu:** `sudo apt-get install graphviz graphviz-dev` is needed before a correct install of the Python `pygraphviz` library.
- **Fedora:** `sudo dnf install graphviz graphviz-devel` is needed before installing `pygraphviz`

Although it can be seen as a minor issue, this is just one example among dozens of libraries. This is real problem with conda. This is why we have decided to use Nix for managing python packages and dependencies. Nix is a purely functional package manager [35]. It allows to easily manage packages and dependencies, and to easily deploy programs on different machines as it guarantees reproducible builds. It is also very useful for development, as it allows to easily create isolated environments for development. This is why we have used Nix for managing the python packages and dependencies. Gradually, Nix has become a superset of other package managers like pip, conda, or cargo.

Any Nix project comes with either a *shell.nix* or a more modern *flake.nix*. Those files are used to describe the project, and to list all necessary dependencies. Since we are developping on NixOS,

the integration of Nix with the operating system is very good, and can be easily setup.

Nix is however really straightforward to install on any other distribution through the use of a single script available online. It can be install in as little as one command.

## 4.3 Converting memory dumps to graphs

Now that we have a decent understanding of the dataset as well as the low level memory dump format, we can start to think about how to convert the memory dumps into graphs. As a recall, we want to be able to convert a memory dump into a graph representation that can be used for machine learning, since we want to be able to create a memory modelization as a basis for efficient embedding and feature engineering later. This is inherently due to the imbalanceness of the dataset, as we want to add more information to each memory block that just its raw bytes. The goal is to have a graph representation of the memory dump that can be used for efficient machine learning.

This section will be dedicated to describe the methods developped around this problem. We will first discuss the design of the graph representation, then we will discuss the implementation of the graph construction process. Finally, we will discuss the semantic embedding of the graph, which is the process of adding more information to the graph.

### 4.3.1 Initial work from Python to Rust

#### 4.3.1.1 First steps

Initially, we have been working and manipulating the code provided by SmartKex<sup>3</sup> for key detection. Our first explorations of the dataset quickly gave birth to some Jupyter Notebooks, which were used to explore the dataset and to understand the code, like `search_in_heap_mem.ipynb`. Rapidly, we decided to rebuild a complete Python 3.11 version of the code. This was done for several reasons:

- The original code has no type hinting, which makes it hard to read and understand.
- We wanted to explore the dataset and learn by doing.
- The original code was not designed to be used as a library, but rather as a standalone script.
- The original code was just a few hundred lines of code and was not designed to be easily extensible, nor to be able to handle a large number of memory dumps.

We decided to build a memory graph representation at that moment because we wanted to be able to add more information to the memory blocks than just their raw bytes. This new program was called `ssh_key_discover`, and relied on a number of Python libraries to work, like `graphviz`. This was a all-in-one library, composed of 2 sections, `mem_graph` and `ml_discovery`. The first one is dedicated to building the memory graph, while the second one is dedicated to the data science and machine learning part.

---

<sup>3</sup>SmartKex GitHub repository: <https://github.com/smartvmi/Smart-and-Naive-SSH-Key-Extraction>

This initial program was already capable of handling several data processing pipelines, including machine learning pipelines with models like Random Forest, a grid search for hyperparameter optimization, and a cross validation pipeline, several balancing strategies and of course, a memory graph representation and semantic embedding. As an early development version, this program was not optimized for performance, and just loading a given heap dump file and its annotation, and then building the memory graph representation, could take from 30 seconds to a minute (on the TUXEDO machine), depending on the size of the heap dump file. As we are dealing with more than  $10^6$  files, a rapid estimation of the time needed just for the semantic embedding of the memory graph representation was above a month. In this regard, this initial program was just used on a bunch of files as a way to develop the semantic embedding model, algorithm and start working on feature engineering and machine learning.

This optimization issue was clearly not acceptable, and we decided to rewrite the graph part in Rust, which is a compiled language, and thus, several orders of magnitude faster than Python. This was also a good opportunity to learn Rust, which is a language that is gaining more and more popularity, especially in the security community. This new program was called mem2graph. Switching from Rust to Python and doing a proper use of multithreading allowed us to reduce the time needed to build the memory graph representation from 30 seconds to less than 1 second. In our case, and comparing using only the TUXEDO laptop, this represents an estimated minimum of a 130x speedup. But this is even much better on the server, where the multithreading can really be leveraged. This was a huge improvement, and allowed us to build the memory graph representation for the whole dataset in a just a few hours.

### 4.3.2 Memory Graph Representation

Now, let's describe the memory graph representation. The goal is to be able to represent a memory dump as a graph. This modelization makes sense since the heap dump can be considered as having data structures as nodes, being connected by pointers acting as arrows. This is a very natural way to represent a memory dump. However in our cases, and since the goal is to make predictions on raw bytes, we will not use the data structures as nodes, but rather the memory blocks directly. This is because we want to be able to make predictions on raw bytes, and not on data structures.

The main idea behind the following memory graph representation is to be able to represent the whole content of a heap dump file as a directed graph that contains the address as well as the byte block of every element. This last point is especially important since we want to be able to make predictions on raw bytes, and not on data structures. This is why we will not use the data structures as nodes, but rather the memory blocks directly.

In the

### 4.3.3 Graph Construction

### 4.3.4 Exporting the Graph

Here are the constants define for memory format conversion in the Rust code:

```
1 pub const BLOCK_BYTE_SIZE: usize = 8; // 64-bit, ex: C0 03 7B 09 2A 56 00 00
2 pub const PTR_ENDIANNESS: Endianness = Endianness::Little;
3 pub const MALLOC_HEADER_ENDIANNESS: Endianness = Endianness::Little;
```

Listing 4.16: Rust memory format constants in *mem2graph*

### 4.3.5 Semantic Embedding

## 4.4 Feature engineering

## 4.5 ML for key prediction

The chapter has been an overview of the dataset, development environment, and tools used for this thesis. In the next chapter, we will dive deeper into the graph memory representation and associated algorithms and programs.

# 5 Results

Describe the experimental setup, the used datasets/parameters and the experimental results achieved

## **6 Discussion**

# 7 Conclusion

The evolving landscape of cybersecurity necessitates robust techniques for safeguarding digital communications. OpenSSH, a pivotal element in this landscape, is a popular implementation of the Secure Shell (SSH) protocol, which enables secure communication between two networked devices. The protocol is widely used in the industry, particularly in the context of remote access to servers. Using digital forensic techniques, it is possible to extract the SSH keys from memory dumps, which can then be used to decode encrypted communications thus allowing the monitoring of controlled systems. At the crux of this Masterarbeit is the development of machine learning algorithms to predict SSH keys within these heap dumps, focusing on using graph-like-structures and vectorization for custom embeddings. With an interdisciplinary approach that fuses traditional feature engineering with graph-based methods as well as memory modelization for inductive reasoning and learning inspired by recent developments in Knowledge Graph (KG)s, this research not only leverages existing machine learning paradigms but also explores new avenues, such as Graph Convolutional Networks (GCN).

## 7.1 Summary of Results

## 7.2 Outlook on Future Work

The present report and related Masterarbeit work have introduced a range of new algorithms and implementations.

In order to integrate efficiently the machine learning part into the *mem2graph* program, more work is needed. Incorporating the machine learning directly after the memory graph construction would greatly improve the efficiency, for instance in the objective of developping a real-time OpenSSH memory forensics tool.

The pipeline and algorithms introduced specifically for the OpenSSH memory dump dataset could be applied to other memory dump datasets. The *mem2graph* program is designed to be modular and extensible, and the algorithms are designed to be as generic as possible. The *mem2graph* program could be used to generate memory graphs for other memory dump datasets, and the machine learning algorithms could be applied to other memory dump datasets. Only some minor algorithm changes are needed, related to the architecture used to generate or extract the heap dump. This would be a major step towards a generic memory forensics tool for machine learning assisted extraction.

Indeed, and as illustrated by the background section, there exist so many ML architectures and solutions that it is impossible to cover them all. The present work has focused on the most popular and promising architectures, but there are many more that could be explored. The programs developped to support ML pipelines are designed to be modular and extensible, and could be used to explore other ML models.



**.1 Code**

**.2 Math**

**.3 Dataset**

# Acronyms

<b>AI</b>	Artificial Intelligence.	21
<b>DEL</b>	Directed Edge-labelled Graphs.	20
<b>ESS</b>	Estimated Security Strength.	6
<b>GCN</b>	Graph Convolutional Networks.	3
<b>GNN</b>	Graph Neural Network.	35, 36
<b>GRU</b>	Gated Recurrent Units.	34
<b>KG</b>	Knowledge Graph.	3, 16, 20, 21
<b>KNN</b>	K-Nearest Neighbors.	23
<b>LDA</b>	Linear Discriminant Analysis.	23
<b>LLM</b>	Large Language Model.	27
<b>ML</b>	Machine Learning.	3, 21, 23, 31, 32
<b>NLP</b>	Natural Language Processing.	25
<b>PCA</b>	Principal Component Analysis.	23
<b>RDF</b>	Resource Description Framework.	20
<b>RNN</b>	Recurrent Neural Networks.	33
<b>SMOTE</b>	Synthetic Minority Over-sampling Technique.	11
<b>SSH</b>	Secure Shell Protocol.	ii, 3, 4, 6
<b>SVM</b>	Support Vector Machine.	23
<b>t-SNE</b>	t-distributed Stochastic Neighbor Embedding.	23
<b>VMI</b>	Virtual Machine Introspection.	10

# References

- [1] Stewart Sentanoe and Hans P. Reiser. „SSHkex: Leveraging virtual machine introspection for extracting SSH keys and decrypting SSH network traffic“. en. In: *Forensic Science International: Digital Investigation* 40 (2022), p. 301337. doi: 10.1016/j.fsidi.2022.301337. url: <https://linkinghub.elsevier.com/retrieve/pii/S2666281722000063>.
- [2] Christofer Fellicious et al. „SmartKex: Machine Learning Assisted SSH Keys Extraction From The Heap Dump“. In: arXiv:2209.05243 (Sept. 2022). arXiv:2209.05243 [cs]. doi: 10.48550/arXiv.2209.05243. url: <http://arxiv.org/abs/2209.05243>.
- [3] M. Baushke. *Key Exchange (KEX) Method Updates and Recommendations for Secure Shell (SSH)*. RFC 9142. Updates: 4250, 4253, 4432, 4462; Errata exist. Internet Engineering Task Force (IETF), Jan. 2022.
- [4] Nicole Perlroth, Jeff Larson, and Scott Shane. „N.S.A. Able to Foil Basic Safeguards of Privacy on Web“. In: *The New York Times* (Sept. 2013). url: <https://www.nytimes.com/2013/09/06/us/nsa-foils-much-internet-encryption.html>.
- [5] James Ball, Julian Borger, and Glenn Greenwald. „Revealed: how US and UK spy agencies defeat internet privacy and security“. In: *The Guardian* (Sept. 2013). Published at 11.24 BST. url: <https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security>.
- [6] Aris Adamantiadis. *OpenSSH introduces curve25519-sha256@libssh.org key exchange !* Retrieved 2023-09-05. Nov. 2013. url: <https://www.libssh.org/2013/11/03/openssh-introduces-curve25519-sha256libssh-org-key-exchange/>.
- [7] OpenSSH. *OpenSSH 5.7 release notes*. Retrieved 2022-11-13. Jan. 2011. url: <https://www.openssh.com/txt/release-5.7>.
- [8] Stewart Sentanoe, Benjamin Taubmann, and Hans P. Reiser. „Sarracenia: Enhancing the Performance and Stealthiness of SSH Honeypots Using Virtual Machine Introspection“. In: *Lecture Notes in Computer Science*. Vol. 11252. Conference paper. Nov. 2018. url: [https://link.springer.com/chapter/10.1007/978-3-030-03638-6\\_16](https://link.springer.com/chapter/10.1007/978-3-030-03638-6_16).
- [9] OpenSSH. *OpenSSH 6.5 release notes*. Retrieved 2022-11-13. Jan. 2014. url: <https://www.openssh.com/txt/release-6.5>.
- [10] OpenSSH. *OpenSSH 7.0 release notes*. Retrieved 2022-11-13. Aug. 2015. url: <https://www.openssh.com/txt/release-7.0>.
- [11] Christine Solnon. „Théorie des graphes et optimisation dans les graphes“. In: *INSA de Lyon* ().
- [12] Douglas Brent West et al. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River, 2001.
- [13] OpenSSH. *OpenSSH 7.2 release notes*. Retrieved 2022-11-13. Jan. 2016. url: <https://www.openssh.com/txt/release-7.2>.
- [14] OpenSSH. *OpenSSH 8.2 release notes*. Retrieved 2022-11-13. Feb. 2020. url: <https://www.openssh.com/txt/release-8.2>.
- [15] OpenSSH. *OpenSSH 8.8 release notes*. Retrieved 2022-11-13. Sept. 2021. url: <https://www.openssh.com/txt/release-8.2>.
- [16] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. „A survey of feature selection and feature extraction techniques in machine learning“. In: *2014 Science and Information Conference*. Aug. 2014, pp. 372–378. doi: 10.1109/SAI.2014.6918213.

- [17] Sinan Ozdemir and Divya Susarla. *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing Ltd, 2018.
- [18] C E Shannon. „A Mathematical Theory of Communication“. en. In: *The Bell System Technical Journal* 27 (Oct. 1948), pp. 379–423.
- [19] F. Pedregosa et al. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [20] Richard Boddy and Gordon Smith. *Statistical methods in practice: for scientists and technologists*. John Wiley & Sons, 2009.
- [21] Michael I Jordan and Tom M Mitchell. „Machine learning: Trends, perspectives, and prospects“. In: *Science* 349.6245 (2015), pp. 255–260. url: <https://www.science.org/doi/full/10.1126/science.aaa8415>.
- [22] Todd G Nick and Kathleen M Campbell. „Logistic regression“. In: *Topics in biostatistics* (2007). Publisher: Springer, pp. 273–301.
- [23] S. B. Kotsiantis. „Decision trees: a recent overview“. In: *Artificial Intelligence Review* 39.4 (Apr. 2013), pp. 261–283. issn: 0269-2821, 1573-7462. doi: 10.1007/s10462-011-9272-4. url: <http://link.springer.com/10.1007/s10462-011-9272-4> (visited on 08/30/2023).
- [24] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix. „Hyperparameters and Tuning Strategies for Random Forest“. In: *WIREs Data Mining and Knowledge Discovery* 9.3 (May 2019), e1301. issn: 1942-4787, 1942-4795. doi: 10.1002/widm.1301. arXiv: 1804.03515[cs, stat]. url: <http://arxiv.org/abs/1804.03515> (visited on 08/30/2023).
- [25] Yann LeCun et al. „Gradient-Based Learning Applied to Document Recognition“. In: *proc of the IEEE* (1998).
- [26] Dai Quoc Nguyen et al. „A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network“. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-2053. url: <https://doi.org/10.18653/v1/n18-2053>.
- [27] Daixin Wang, Peng Cui, and Wenwu Zhu. „Structural Deep Network Embedding“. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1225–1234. isbn: 9781450342322. doi: 10.1145/2939672.2939753. url: <https://doi.org/10.1145/2939672.2939753>.
- [28] Michael Schlichtkrull et al. „Modeling relational data with graph convolutional networks“. In: *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer. 2018, pp. 593–607. url: <https://arxiv.org/pdf/1703.06103.pdf>.
- [29] Liang Yao, Chengsheng Mao, and Yuan Luo. „KG-BERT: BERT for knowledge graph completion“. In: *arXiv preprint arXiv:1909.03193* (2019). url: <https://arxiv.org/pdf/1909.03193.pdf>.
- [30] Ye Liu et al. „Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning“. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 7. 2021, pp. 6418–6425. url: <file:///home/onyr/Downloads/16796-Article%20Text-20290-1-2-20210518.pdf>.
- [31] Aditya Grover and Jure Leskovec. „node2vec: Scalable feature learning for networks“. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864. url: <https://dl.acm.org/doi/pdf/10.1145/2939672.2939754>.

- [32] Jian Tang et al. „Line: Large-scale information network embedding“. In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 1067–1077. url: <https://dl.acm.org/doi/pdf/10.1145/2736277.2741093>.
- [33] Petar Velickovic et al. „Graph attention networks“. In: *stat* 1050.20 (2017), pp. 10–48550. url: <https://personal.utdallas.edu/~fxc190007/courses/20S-7301/GAT-questions.pdf>.
- [34] Keiron O’Shea and Ryan Nash. „An Introduction to Convolutional Neural Networks“. In: *CoRR* abs/1511.08458 (2015). arXiv: 1511.08458. url: <http://arxiv.org/abs/1511.08458>.
- [35] Eelco Dolstra. *The purely functional software deployment model*. Utrecht University, 2006. url: <https://edolstra.github.io/pubs/phd-thesis.pdf>.
- [36] OpenScience ASAP. *Was ist Open Science?* Accessed: 2023/09/19. 2023. url: <http://openscienceasap.org/open-science/> (visited on 09/19/2023).
- [37] DGE Gomes et al. „Why don’t we share data and code? Perceived barriers and benefits to public archiving practices“. In: *Proc Biol Sci* 289.1987 (Nov. 30, 2022), p. 20221113. doi: 10.1098/rspb.2022.1113. eprint: Epub2022Nov23.
- [38] Eelco Dolstra and Andres Löh. „NixOS: A Purely Functional Linux Distribution“. In: *SIGPLAN Not.* 43.9 (Sept. 2008), pp. 367–378. issn: 0362-1340. doi: 10.1145/1411203.1411255. url: <https://doi.org/10.1145/1411203.1411255>.
- [39] Jamie Ludwig. „Image convolution“. In: *Portland State University* (2013). url: [https://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Ludwig\\_ImageConvolution.pdf](https://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Ludwig_ImageConvolution.pdf).
- [40] Vic Degraeve et al. „R-GCN: the R could stand for random“. In: *arXiv:2203.02424 preprint* (2022). url: <https://arxiv.org/pdf/2203.02424.pdf>.
- [41] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. „Deepwalk: Online learning of social representations“. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 701–710. url: <https://dl.acm.org/doi/pdf/10.1145/2623330.2623732>.
- [42] Will Hamilton, Zhitao Ying, and Jure Leskovec. „Inductive representation learning on large graphs“. In: *Advances in neural information processing systems* 30 (2017). Ed. by I. Guyon et al. url: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf).
- [43] Zonghan Wu et al. „A comprehensive survey on graph neural networks“. In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24. url: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9046288>.
- [44] Zonghan Wu et al. „Beyond low-pass filtering: Graph convolutional networks with automatic filtering“. In: *IEEE Transactions on Knowledge and Data Engineering* (2022). url: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9806316>.
- [45] Federico Monti et al. „Geometric deep learning on graphs and manifolds using mixture model cnns“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5115–5124. url: [https://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Monti\\_Geometric\\_Deep\\_Learning\\_CVPR\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2017/papers/Monti_Geometric_Deep_Learning_CVPR_2017_paper.pdf).
- [46] Franco Scarselli et al. „The graph neural network model“. In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80. url: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4700287>.
- [47] Joan Bruna et al. „Spectral networks and locally connected networks on graphs“. In: *arXiv preprint arXiv:1312.6203* (2013). url: <https://arxiv.org/pdf/1312.6203.pdf>.

- [48] Quoc V Le et al. „A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks“. In: *Google Brain* 20 (2015), pp. 1–20. url: <https://ai.stanford.edu/~quocle/tutorial2.pdf>.
- [49] Thomas N Kipf and Max Welling. „Semi-supervised classification with graph convolutional networks“. In: *arXiv preprint arXiv:1609.02907* (2016). url: <https://arxiv.org/pdf/1609.02907.pdf>.
- [50] Junyoung Chung et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. Dec. 11, 2014. arXiv: 1412.3555[cs]. url: <http://arxiv.org/abs/1412.3555> (visited on 08/23/2023).
- [51] J. Laaksonen and E. Oja. „Classification with learning k-nearest neighbors“. In: *Proceedings of International Conference on Neural Networks (ICNN'96)*. International Conference on Neural Networks (ICNN'96). Vol. 3. Washington, DC, USA: IEEE, 1996, pp. 1480–1483. isbn: 978-0-7803-3210-2. doi: 10.1109/ICNN.1996.549118. url: <http://ieeexplore.ieee.org/document/549118/> (visited on 08/30/2023).
- [52] Sepp Hochreiter and Jürgen Schmidhuber. „Long short-term memory“. In: *Neural computation* 9.8 (1997). Publisher: MIT Press, pp. 1735–1780. (Visited on 08/23/2023).
- [53] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. url: [https://books.google.de/books?hl=en&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=deep+learning&ots=MNV2eosBRS&sig=jN2QwFikq3g\\_YqU3hJVPEP0XI4&redir\\_esc=y#v=onepage&q=deep%20learning&f=false](https://books.google.de/books?hl=en&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=deep+learning&ots=MNV2eosBRS&sig=jN2QwFikq3g_YqU3hJVPEP0XI4&redir_esc=y#v=onepage&q=deep%20learning&f=false).
- [54] Qiang Wu and Ding-Xuan Zhou. „Analysis of Support Vector Machine Classification“. In: *Journal of Computational Analysis & Applications* 8.2 (2006).
- [55] Jianlong Zhou et al. „Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics“. In: *Electronics* 10.5 (Mar. 4, 2021), p. 593. issn: 2079-9292. doi: 10.3390/electronics10050593. url: <https://www.mdpi.com/2079-9292/10/5/593> (visited on 09/11/2023).
- [56] Jose Manuel Gomez-Perez, Ronald Denaux, and Andres Garcia-Silva. „Understanding Word Embeddings and Language Models“. In: *A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP*. Cham: Springer International Publishing, 2020, pp. 17–31. isbn: 978-3-030-44830-1. doi: 10.1007/978-3-030-44830-1\_3. url: [https://doi.org/10.1007/978-3-030-44830-1\\_3](https://doi.org/10.1007/978-3-030-44830-1_3).
- [57] OpenSSH. *OpenSSH 9.0 release notes*. Retrieved 2022-11-13. Apr. 2022. url: <https://www.openssh.com/txt/release-9.0>.
- [58] Unknown. *How does SSH use both RSA and Diffie-Hellman?* Asked 8 years, 9 months ago; Modified 8 years, 9 months ago; Viewed 22k times; Accessed: 2023-09-21. 2023. url: <https://security.stackexchange.com/questions/76894/how-does-ssh-use-both-rsa-and-diffie-hellman>.
- [59] T. Ylonen and C. Lonvick. *The Secure Shell (SSH) Protocol Architecture*. RFC 4251. Updated by: 8308, 9141. Network Working Group, Jan. 2006.
- [60] T. Ylonen and C. Lonvick. *The Secure Shell (SSH) Transport Layer Protocol*. RFC 4253. Updated by: 6668, 8268, 8308, 8332, 8709, 8758, 9142; Errata Exist. Network Working Group, Jan. 2006.
- [61] Girish Venkatachalam. „The OpenSSH Protocol under the Hood“. In: *Linux Journal* 156 (Apr. 2007). url: <https://www.ecb.torontomu.ca/~courses/coe518/LinuxJournal/elj2007-156-OpenSSH.pdf>.
- [62] Oscar M. Guillen et al. „Towards post-quantum security for IoT endpoints with NTRU“. In: *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. 2017, pp. 698–703. doi: 10.23919/DATE.2017.7927079.

- [63] Paulo Nunes de Souza and Pavel Gladyshev. „Inference of Endianness and Wordsize From Memory Dumps“. In: *European Conference on Cyber Warfare and Security*. Academic Conferences International Limited. 2017, pp. 619–627.
- [64] P. McLaren et al. „Decrypting live SSH traffic in virtual environments“. In: *Digital Investigation* 29 (2019), pp. 109–117. url: <https://www.sciencedirect.com/science/article/abs/pii/S1742287619300647>.
- [65] Tatu Ylonen. *Portable OpenSSH*. Github repositor. Accessed on 25.08.2023. 1995. url: <https://github.com/openssh/openssh-portable/>.
- [66] José Tomás Martínez Garre, Manuel Gil Pérez, and Antonio Ruiz-Martínez. „A novel Machine Learning-based approach for the detection of SSH botnet infection“. In: *Future Generation Computer Systems* 115 (Feb. 2021), pp. 387–396. doi: 10.1016/j.future.2020.09.004. url: <https://www.sciencedirect.com/science/article/pii/S0167739X20303265>.
- [67] SSH Communications Security. *SSH Annual Report 2018*. Annual Report. Accessed: 2023-08-30. SSH Communications Security, 2018. url: [https://info.ssh.com/hubfs/2021%20Investor%20documents/SSH\\_Annual\\_Report\\_2018\\_final.pdf](https://info.ssh.com/hubfs/2021%20Investor%20documents/SSH_Annual_Report_2018_final.pdf).
- [68] W. Yurcik and Chao Liu. „A first step toward detecting SSH identity theft in HPC cluster environments: discriminating masqueraders based on command behavior“. In: *CCGrid 2005. IEEE International Symposium on Cluster Computing and the Grid, 2005*. Vol. 1. May 2005, 111–120 Vol. 1. doi: 10.1109/CCGRID.2005.1558542.
- [69] *Weak CRC allows packet injection into SSH sessions encrypted with block ciphers*. Accessed: 2023-08-30. Nov. 2001. url: <https://www.kb.cert.org>.
- [70] Core Security Technologies. *SSH Insertion Attack*. <https://www.coresecurity.com/core-labs/advisories/ssh-insertion-attack>. Archived from the original on 2011-07-08. 2023.
- [71] US CERT. *SSH CBC vulnerability. Vulnerability Note VU#958563 - SSH CBC vulnerability*. <https://www.kb.cert.org/vuls/id/958563>. Archived from the original on 2011-06-22. 2011.
- [72] Spiegel Online. *Prying Eyes: Inside the NSA's War on Internet Security*. Spiegel Online. Archived from the original on January 24, 2015. 2014. url: <https://www.spiegel.de/international/germany/inside-the-nsa-s-war-on-internet-security-a-1010361.html>.
- [73] Olivier Bilodeau et al. *Operation WINDIGO*. en. Mar. 2014, p. 69. url: [https://web-assets.esetstatic.com/wls/2014/03/operation\\_windigo.pdf](https://web-assets.esetstatic.com/wls/2014/03/operation_windigo.pdf).
- [74] Pooneh Nikkhah Bahrami\* et al. „Cyber Kill Chain-Based Taxonomy of Advanced Persistent Threat Actors: Analogy of Tactics, Techniques, and Procedures“. In: *Journal of Information Processing Systems* 15.4 (Nov. 2019), pp. 865–889. doi: 10.3745/JIPS.03.0126. url: <http://xml.jips-k.org/full-text/view?doi=10.3745/JIPS.03.0126>.
- [75] Sanjay Madan and Monika Singh. „Classification of IOT-Malware using Machine Learning“. In: *2021 International Conference on Technological Advancements and Innovations (ICTAI)*. Nov. 2021, pp. 599–605. doi: 10.1109/ICTAI53825.2021.9673185.
- [76] Connor Hetzler, Zachary Chen, and Tahir M. Khan. „Analysis of SSH Honeytrap Effectiveness“. en. In: *Advances in Information and Communication*. Ed. by Kohei Arai. Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland, Mar. 2023, pp. 759–782. isbn: 9783031280733. doi: 10.1007/978-3-031-28073-3\_51.
- [77] ppacher. *honeyssh*. <https://github.com/ppacher/honeyssh>. GitHub repository. 2017.
- [78] Aidan Hogan et al. „Knowledge Graphs“. In: *ACM Comput. Surv.* 54.4 (July 2021). issn: 0360-0300. doi: 10.1145/3447772. url: <https://doi.org/10.1145/3447772>.

- [79] Aidan Hogan et al. „Knowledge Graphs (Extended)“. In: *ACM Computing Surveys* 54.4 (May 2022). arXiv:2003.02320 [cs], pp. 1–37. doi: 10.1145/3447772. url: <http://arxiv.org/abs/2003.02320>.
- [80] Marvin Hofer et al. „Construction of Knowledge Graphs: State and Challenges“. In: *arXiv preprint arXiv:2302.11509* (2023). url: <https://doi.org/10.48550/arXiv.2302.11509>.
- [81] Google. „Introducing the Knowledge Graph: Things, not strings“. In: *Google Blog* (May 2012). Accessed: 2023-06-16. url: <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.

## Additional bibliography

- [82] Auguste Kerckhoffs. „La cryptographic militaire“. In: *Journal des sciences militaires* (1883), pp. 5–38.
- [83] Dimitrios Georgoulas et al. „Botnet Business Models, Takedown Attempts, and the Darkweb Market: A Survey“. en. In: *ACM Computing Surveys* 55.11 (Nov. 2023), pp. 1–39. doi: 10.1145/3575808. url: <https://dl.acm.org/doi/10.1145/3575808>.
- [84] Paul Groth et al. „Knowledge Graphs and their Role in the Knowledge Engineering of the 21st Century“. In: *Dagstuhl Reports* 12.9 (2022). Report from Dagstuhl Seminar 22372. Specific usage: pp. 60–72, Subsection "3.2 A Brief History of Knowledge Engineering: A Practitioner's Perspective", pp. 60–120. doi: 10.4230/DagRep.12.9.60.
- [85] Lisa Ehrlinger and Wolfram Wöß. „Towards a Definition of Knowledge Graphs“. In: (2016), pp. 1–4.
- [86] Frederick Edward Hulme. *Proverb Lore: Many Sayings, Wise Or Otherwise, on Many Subjects, Gleaned from Many Sources*. E. Stock, 1902, p. 188.
- [87] Michelle Venables. *An Introduction to Graph Theory*. Accessed: 2023-06-12. 2019. url: <https://towardsdatascience.com/an-introduction-to-graph-theory-24b41746fabe>.
- [88] Gianluca Fiorelli. *Best of 2013: No 13 - Search in the Knowledge Graph era*. Accessed: 2023-06-12. 2013. url: <https://www.stateofdigital.com/search-in-the-knowledge-graph-era/>.
- [89] Jackson Gilkey. *Graph Theory and Data Science*. Accessed: 2023-05-25. 2019. url: <https://towardsdatascience.com/graph-theory-and-data-science-ec95fe2f31d8>.
- [90] M.S. Jawad et al. „Adoption of knowledge-graph best development practices for scalable and optimized manufacturing processes“. In: *MethodsX* 10 (2023), p. 102124. issn: 2215-0161. doi: <https://doi.org/10.1016/j.mex.2023.102124>. url: <https://www.sciencedirect.com/science/article/pii/S2215016123001255>.



## **Eidesstattliche Erklärung**

Hiermit versichere ich, dass ich diese Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe und alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, als solche gekennzeichnet sind, sowie, dass ich die Masterarbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt habe.

Passau, September 22, 2023

---

Rascoussier, Florian Guillaume Pierre