On GDPR Compliant Data Processing

Supriya Adhatarao

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Supriya ADHATARAO

Thèse dirigée par **Claude CASTELLUCCIA**
et co-encadrée par **Cédric LAURADOUX**, INRIA

préparée au sein du **Laboratoire Institut National de Recherche en Informatique et en Automatique**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Sur le traitement de données conforme au RGPD

On GDPR Compliant Data Processing

Thèse soutenue publiquement le **22 juillet 2021**,
devant le jury composé de :

Monsieur CLAUDE CASTELLUCCIA
DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES, Directeur de thèse
Madame MARINE MINIER
PROFESSEUR DES UNIVERSITES, UNIVERSITE DE LORRAINE, Rapporteure
Monsieur PHILIPPE ELBAZ-VINCENT
PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES, Président
Monsieur PASCAL LAFOURCADE
MAITRE DE CONFERENCE HDR, UNIVERSITE CLERMONT AUVERGNE, Rapporteur
Monsieur GUILLAUME PIOLLE
MAITRE DE CONFERENCE, CENTRALESUPELEC, Examinateur

Université Grenoble Alpes

THESIS

for the degree of

DOCTOR of GRENOBLE ALPES UNIVERSITY

Specialty: MATHEMATICS AND COMPUTER SCIENCE

Ministerial Order: May 25, 2016

Presented by,

Supriya Srikant Adhatarao

Thesis supervised by **Claude Castelluccia,** Research Director, Inria Grenoble - Rhône-Alpes and co-supervised by **Cédric Lauradoux,** Tenured Researcher, Inria Grenoble - Rhône-Alpes

Prepared at the laboratories: **PRIVATICS INRIA Grenoble - Rhône-Alpes** in **the Doctoral School MSTII (Mathématiques, Sciences et Technologies de l'Information et de l'Informatique)**

On GDPR Compliant Data Processing

Defended on 22/07/2021
Jury:

M. Philippe ELBAZ-VINCENT
Professor, Univ. Grenoble Alpes, President
Mme. Marine MINIER
Professor, Université de Lorraine, Reviewer
M. Pascal LAFOURCADE
Associate Professor, Université Clermont Auvergne, Reviewer
M. Guillaume PIOLLE
Associate Professor, CentraleSupélec (Rennes campus), Examiner
M. Claude CASTELLUCCIA
Research Director, Univ. Grenoble Alpes, PhD Supervisor
M. Cédric LAURADOUX
Tenured Researcher, Univ. Grenoble Alpes, PhD Co-supervisor

Abstract

The General Data Protection Regulation (GDPR) in Europe aims to give back the control of the personal data to its owners. This requires a legal basis for data processing, *in order for processing to be lawful, personal data should be processed on the basis of the consent of the data subject concerned or some other legitimate basis, laid down by law (Recital 40 GDPR)*. In order to comply with the GDPR regulations both data subjects and data controllers need to know what exactly constitutes to personal data. According to the law, personal data is any information that relates to an identified or identifiable individual [11]. At first glance, this definition of personal data seems simple but there resides many ambiguities around this definition. In this context, we first examine the definition of personal data under GDPR and its territorial scope. The broad objective of this work is to leverage difficulties faced by both data subjects and data controllers when a subject access request is made. In this regard, we mainly consider two kinds of identifiers: *name and IP address*. We describe the legal and technical ambiguities faced by data subjects to exercise their rights and by data controllers to assess the status of these identifiers. In the context of personal data, we also show different kind of information shared by users. We examine what information shared by users constitutes to build a profile and how the profile information can be used to identify a specific individual and how it could be exploited by an adversary to target an individual. The later part of this dissertation focuses on the documents shared by users and the personal data leaked within these documents. In our work, we mainly focus on the analysis of PDF files. We have analyzed various hidden information present in PDF files and show how an adversary can exploit it to target an author and also an organization. We then demonstrate that there is a strong need for the enforcement of PDF file sanitization. Finally, *we focus on the forensics problem: is it possible to determine how a PDF file has been created using the file itself?* We describe a novel approach to detect the software that has been used to produce a PDF file. Our detection tool is based on coding style: given patterns that are only created by certain PDF producers. This tool has many applications in offensive security and as well as in incident response.

Résumé

Le règlement général sur la protection des données (RGPD) en Europe vise à redonner le contrôle des données personnelles à leurs propriétaires. Cela nécessite une base juridique pour le traitement des données, *pour que le traitement soit licite, les données personnelles doivent être traitées sur la base du consentement de la personne concernée ou sur une autre base légitime, prévue par la loi (considérant 40 du RGPD)*. Afin de se conformer à la réglementation du RGPD, les personnes concernées et les responsables du traitement doivent savoir en quoi consistent exactement les données à caractère personnel. Selon la loi, une donnée personnelle est toute information qui se rapporte à une personne physique identifiée ou identifiable [11]. À première vue, cette définition des données à caractère personnel semble simple, mais de nombreuses ambiguïtés subsistent autour de cette définition. Dans ce contexte, nous examinons d'abord la définition des données personnelles dans le cadre du RGPD et sa portée territoriale. L'objectif général de ce travail est de tirer parti des difficultés rencontrées tant par les personnes concernées que par les responsables du traitement des données lorsqu'une demande d'accès est formulée. À cet égard, nous considérons principalement deux types d'identifiants: le nom et l'adresse IP. Nous décrivons les ambiguïtés juridiques et techniques auxquelles sont confrontées les personnes concernées pour exercer leurs droits et les responsables du traitement pour évaluer le statut de ces identifiants. Dans le contexte des données personnelles, nous montrons également différents types d'informations partagées par les utilisateurs. Nous examinons quelles informations partagées par les utilisateurs constituent un profil et comment les informations du profil peuvent être utilisées pour identifier un individu spécifique et comment elles pourraient être exploitées par un attaquant pour cibler un individu. La dernière partie de cette thèse se concentre sur les documents partagés par les utilisateurs et les données personnelles divulguées dans ces documents. Dans notre travail, nous nous concentrons principalement sur l'analyse des fichiers PDF. Nous avons analysé diverses informations cachées présentes dans les fichiers PDF et montré comment un attaquant peut les exploiter pour cibler un auteur et aussi une organisation. Nous démontrons ensuite qu'il existe un fort besoin d'appliquer la nettoyage des fichiers PDF. Enfin, nous nous concentrons sur le problème criminalistique: est-il possible de déterminer comment un fichier PDF a été créé à partir du fichier lui-même? Nous décrivons une nouvelle approche pour détecter le logiciel qui a été utilisé pour produire un fichier PDF. Notre outil de détection est basé sur le style de codage: des motifs donnés qui ne sont créés que par certains producteurs de PDF. Cet outil a de nombreuses applications dans la sécurité offensive ainsi que dans la réponse aux incidents.

Acknowledgement

With great pleasure and immense gratitude, I would like to convey, my sincerest acknowledgement and appreciation to all those people who have contributed to the wonderful learning experience during my PhD. These three and a half years of my PhD in Privatics team will be very memorable. I had an opportunity to work on several interesting topics that I liked and more importantly, I had the opportunity to meet remarkable people. Everyone in the team created a friendly environment that I enjoyed during my work which contributed to be a memorable period. Everything, including my thesis, would not have been the same without these amazing people.

First and foremost, I would like to sincerely thank my PhD adviser: Dr. Cedric Lauradoux, whose continuous support, encouragement, patience and guidance along with his expertise were vital to my PhD thesis. I am grateful for the multiple opportunities that he let me explore for my research and for introducing me to newer topics and collaborations. His knowledge and passion for the research work has influenced my work greatly. I am deeply grateful for all the encouragement and support that I have received from him throughout my PhD.

I would like to convey my deepest admiration and gratitude to my PhD director Dr. Claude Castelluccia for generously contributing his invaluable time, support and guidance in my pursuit of PhD. I was fortunate to learn many things from him and I will always be grateful for everything he has done.

To my team leader Dr. Vincent Roca, I would like to convey my heartfelt gratitude and thank him for always encouraging and helping. He created a friendly environment for all the PhD students and was always caring and encouraging. I was allowed to discuss and seek help whenever needed. Your immense support, concern, advice and discussions are very invaluable and contributed a lot during my PhD, thank you.

I am thankful to Helen Pouchot for helping me in managing the administrative aspects of the PhD and for all the concern. My special thanks to all the non-permanent members and friends in the team.

I am greatly thankful to the reviewers and examiners of my thesis: Prof. Marine Minier, Prof. Philippe Elbaz-Vincent, Dr. Pascal Lafourcade and Dr. Guillaume Piolle for making time to review my work and for participating in my thesis defense.

I wish to convey my deepest and heartfelt gratitude to my parents Srikant Adhatarao and Surekha Adhatarao without whom I wouldn't have got the opportunity to pursue higher education. Along with giving me freedom, they have always encouraged me to pursue my dreams and goals. I can never thank them in words for everything they have done for me, I am eternally grateful to you both. I would also like to thank my sister Dr. Sripriya Adhatarao and my brother M.Sc. Sreeprasad Adhatrao for their never ending encouragement and support, I will never be able to thank them properly. I am extremely lucky to have such a beautiful and loving family whose support and encouragement is the reason this thesis was ever possible. I would also like to thank all my friends for their support and encouragement.

Finally, I want to thank anyone and everyone who directly and indirectly helped me and assisted me during my PhD.

Contents

1 Introduction

Today we live in a global digital economy, we create data very differently and the data volumes are exploding. Digital activities in our lives has reached new heights, more and more people are spending time doing a number of things online than ever before. Globally it has been noted that in January 2020, the number of people around the world using the internet has grown to 4.54 billion and there are 3.80 billion social media users[1]. Users are increasingly becoming interested in online activities and sharing the content they desire. Many of the users share their personal data online and are unaware or know very little about their online privacy and plausible risks associated with it.

Both the legal and technical definition of personal data is very complicated to understand. There lies many ambiguities around the definition of personal data and hence it looks like a maze. Since it is so complicated to understand, many users tend to do whatever they want and share a lot of content online. Privacy concerns have spiked in last few years due to the implementation of cookie banners and many data scandals. Incidents of data misuse have alarmed many users and forced them to rethink their relationships to social media and the security of their personal information. One such alarming incident was the Facebook–Cambridge Analytica data scandal[2]. In this scandal the personal data of millions of Facebook users was obtained without their consent by British consulting firm `Cambridge Analytica`. The firm exploited the private information of over 50 million Facebook users to influence the American presidential election of 2016. This example and others have steadily deteriorated public trust and resulted in many users wondering if they have lost control over their personal data.

According to a study conducted by the Pew Trust, 80% of social media users report being concerned about businesses and advertisers accessing and using their social media posts[3]. Internet privacy is becoming a growing concern these days for people of all ages.

[1] `https://thenextweb.com/growth-quarters/2020/01/30/digital-trends-2020-every-single-stat-you-need-to-know-about-the-internet/`
[2] `https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal`
[3] `https://sopa.tulane.edu/blog/key-social-media-privacy-issues-2020`

The growing privacy concerns have prompted advocacy for tighter regulations. As of January 2021, over 130 jurisdictions [21] have enacted on the data privacy laws to increase or improve information privacy and security. In addition, they have placed companies and organizations responsible for safeguarding personal data under greater scrutiny.

In Europe, in regards to protect user's personal data the General Data Protection Regulation (GDPR) was implemented. The EU Commission describes the GDPR as: *an essential step to strengthening citizens' fundamental rights in the digital age [which] provides tools for gaining control of one's personal data.* It aims to protect fundamental rights and freedoms of natural persons and more specifically, their right to the protection of personal data [82] (Article 1(2) GDPR).

The primary objectives of the GDPR is to give control back to citizens and residents over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU. It gives users different kinds of rights on their data: *the right to be informed, right of access, right to rectification, right to erasure, right to restrict processing, right to data portability, right to object, rights related to automated decision-making and profiling*[4].

With the increase in online user activities, predicting user behavior, disclosure of users' privacy [6, 5, 64, 104], third-party tracking on the web [96, 4, 22, 35], browser fingerprinting [3, 113, 112, 62], industry regulations such as the EU Cookie Law [30] and W3C Do-Not-Track [114] etc.. have been some of the most active research topics.

1.1 Motivation

This section discusses the high level open research problems and challenges that are the focus of this dissertation.

Systéme Intelligent D'Enseignement en Santé (SIDES) is a national web platform shared by the 53 medicine schools in France to validate and automatically correct exams using tablets. The platform is also used to prepare 70000 students to the Epreuves Classantes Nationales informatisées (ECNi) through an auto-evaluation module. SIDES 3.0 is a proposition of evolution as many partners wanted to add new functionalities like, personalized recommendations and coaching, augmented interfaces and correction (for the self-evaluation mode). One of the many goals of SIDES is to allow researchers to use the data for scientific purposes. Hence,

[4] `https://www.termsfeed.com/blog/gdpr-8-user-rights/`

data protection and privacy are critical issues for SIDES. Personalization implies to manipulate data to create profiles and then modify the content provided to the students. It is therefore important to ensure that students privacy is respected in order to prevent any abuses such as tracking or discrimination. In regards to these issues, initially our research started on the analysis of the way data is collected and proceeded by SIDES.

During our ongoing research, we realized that this situation is same for almost all the online users, it is important to ensure that users privacy is respected in order to prevent any abuses such as tracking or discrimination. On a border view, in regards to privacy and personal data of online users, some of our research questions were:
(i) What are the different personal information shared by online users?
(ii) Do they know the sensitivity associated to the information shared?
(iii) How to educate them about online privacy?

In regards to these questions, our research work, firstly focuses on the definition of personal data under GDPR, different identifiers that constitute to be personal data and different scenarios a piece of information is viewed as personal data etc.. We started working on the most basic personal information shared by online users, there `name`. Our research questions involved, the scenarios where the users share there name (either completely or partially). Is name always a personal data? What GDPR says about name being the personal data? Different options provided by GDPR to exercise the rights, difficulties faced by data controllers to identify a data subject using his/her name etc..

After the analysis of names, we worked on an identifier used by all the devices that are connected to the internet: *IP address*. An IP address is an online identifier [54, 50] used for identifying devices online and to route information between websites and its end-user devices. Almost every website collects and processes (maybe under legitimate motivations) IP addresses of their users, mostly for quality of services or security purposes. Work done by Mishra *et al.* [75] showed that IP addresses can be effectively used to track Internet users, suchlike cookies [60] or browser fingerprinting [62].

Personal data has been constantly broadened to include IP address [84, 85, 55, 19]. In the last years, the Court of Justice of the European Union (CJEU) was asked several times [27, 36, 18] to determine if an IP address is a personal data or not. The question asked to the court was: *Is an IP address an information related to an identified or identifiable natural person (Article 4 of the GDPR)?* The decisions taken by the courts can be summarized as follows: *IP addresses are personal data under certain conditions.* At first sight, this position looks like a perfect compromise for Internet users and websites: users are comforted with the fact that it is not possible to abuse IP addresses to track them and websites can find exceptions to

keep collecting IP addresses without being bound to the legal requirements set forth by the GDPR.

This question is fundamental for online privacy because organizations processing IP addresses will need to apply the GDPR. Therefore, determining if IP addresses can identify a user or not can have a huge impact on the protection of personal data of individuals and on the safeguards required to process IP addresses. Understanding the legal status of IP addresses is complex. In Europe, the General Data Protection Regulation (GDPR) is supposed to have leveraged the legal status of IP addresses as personal data, but recent decisions from the European Court of Justice undermine this view so our next research involved analysis of IP addresses.

In the later part of the dissertation, we started working on a totally different research topic. Our second research topic involves work on the documents shared online. Our research involved some question such as:
(i) Is metadata present within files personal data subject to the GDPR?
(ii) Can metadata and other hidden information be used to identify an individual?
(iii) Are hidden information in files highly sensitive data? etc..

Among all the available file formats, at some point or another, all of us have used PDF files to share documents digitally. Since PDF maintains the original fonts, images, graphics as well as the exact layout of the file, it can be shared, viewed and printed by anyone regardless of the Operating System, original design application or fonts. Over time, the format has evolved to support more features like security, searchability and description by metadata etc.. For several years, this format has been the most popular format to share documents. In fact, it has become the industry standard for sharing professional documents online. It is an active research topic, many works in the past involve analysis of PDF file security [103, 100, 17, 115, 68], privacy [99, 57, 42] and PDF file sanitization [13, 42, 39]. Even though this format is constantly evolving to provide better service and security, many vulnerabilities have been found in the past in PDF viewers: 1090 vulnerabilities were found by December 2020 according to https://cve.mitre.org. This motivated us in exploring other aspects related to the PDF file format, different hidden information, their forensics and sanitization.

1.2 Contribution

Instant connectivity has changed the way we live and work for the better, but this convenience comes at a huge price: our privacy. That's why it has become more important than ever for

all of us to take responsibility for protecting our Internet privacy and personal information shared online. This dissertation provides the following contributions to address the research problems discussed in Section 1.1.

1.2.1 Personal data

The term *personal data* is the entryway to the application of the GDPR. Only if a processing of data concerns personal data, the General Data Protection Regulation applies (Article. 4 (1)). Personal data are any information which are related to an identified or identifiable natural person. A person is identifiable if he/she can be directly or indirectly identified, especially by reference to an identifier such as a name, an identification number, location data, an online identifier etc..

Our research mainly involves analysis of two identifiers **Name** and **IP address**. We first consider different naming system and determine when a name of an individual is personal data. Then we considered different scenarios where the uniqueness of a name directly identifies an individual and different context where name is not enough to identify someone and additional quasi identifiers like date of birth, gender, place etc.. are used. Our analysis is studied on the European naming system. We show different scenarios and difficulties faced while applying privacy rules such as GDPR to identify a individual using his/her full name. We also consider the territorial scope of GDPR and analyze the difficulties faced while processing the data of users belonging to different countries.

Our study further continues to explain the uniqueness of names for French citizens. INSEE[5] maintains the statistics of the registered names of all the French people. We used the dataset available on the INSEE website for both first names and last names and developed a tool to show the uniqueness of French names. Our tool shows the uniqueness of a name w.r.t to the department a user belongs to. This tool is a step towards educating online users on the sensitivity of their name and the privacy implications of sharing it online.

To further analyze different ways a user shares his/her name online, we got access to the dataset of Fun Inria MOOC users. We studied usernames of 20,661 unique participants. Our experiments show that users use personal information such as first name, last name, department code, postal code, date of birth, country code etc.. while creating there usernames. Using existing online tools we predicted 731 users ethnicity and nationality . Using gender prediction tool, we predicted 2436 users gender. Using the postal code information used in usernames, we were able to find the city of at least 15 users. Apart from sharing sensitive

[5] `https://www.insee.fr/fr/accueil`

information within the usernames, many users also tend to use same usernames across different websites. We found many online tools that could be used to link online usernames across several websites and compromise their security and privacy (Section 2.3.3). Our research w.r.t usernames also showed that websites consider only password as the sensitive information and not the username.

After the analyses of identifier *name*, we analyzed the identifier *IP address*. Understanding the legal status of IP addresses is difficult. Legislation on data protection vary in each country. In Europe, IP addresses are considered personal data under the GDPR and also by the European Data Protection Board.

We have attempted to determine if IP addresses are personal data using the tools that are available to an Internet user. We have visited 109 websites of public and private organizations. We have analyzed their privacy policies and then we have sent them subject access requests. Our requests were very specific: we ask the websites to provide all the data concerning the IP addresses we have used when visiting their websites. The result obtained is clear: companies display that they consider IP addresses as personal data but it has no implication for them. All our requests were denied, it is not possible to access data related to an IP address.

The findings on privacy policies of 124 organizations shows that, most private companies (with the exception of Ebay and Tripadvisor) mention the collection of IP addresses in their privacy polices. On SAR responses, from websites of 68 companies and 45 DPAs to whom we have sent a SAR request on IP addresses, 0 websites provided data on the IP address and at least 6 companies (Trip Advisor, Microsoft, gumgum.com, rubiconproject.com, pubmatic and lyst.co.uk) mention the processing of IP addresses in their response but did not provide data on the IP addresses as they are dynamic and also might be shared by other users.

The most interesting answer given to us consists in saying that a given IP address can identify several users. This answer is legitimate because of the network topology and also because so far Internet users are unable to prove that they have used a given IP address during a certain period of time. Therefore, they cannot submit valid subject access requests to the websites. Our study shows that the current status of IP addresses has favored websites to be detriment of Internet user's rights. To fix this situation, we propose to change how IP addresses are allocated to Internet users.

1.2.2 PDF files

Portable Document Format (PDF) is used to publish documentation or exchange printable documents. The PDF format is one of the most popular way to exchange documents on Internet. Over 1.6 million PDF files have been published on arXiv[6] and 0.8 million have been uploaded on the open archive HAL[7]. A PDF file contains all the information (authors identity, organization etc..) that its authors have decided to provide to the readers. But it also contains metadata which are not often provided by the authors or they are not aware that metadata are present. The metadata of a PDF file can provide information on the authors and on the authoring process used to create the file. By inspecting a PDF file, it is possible to recover the Operating System used by the authors and also different software (references, pictures...) used to create the file as well as many other information. Tools like `exiftool` or `pdfimages` are available to extract these data and anybody can now analyze all the information contained in a PDF file as shown in the past [13, 42, 39, 74]. Metadata are often considered as a threat to privacy. Users are often unaware of the metadata and that they can contain more information than the actual content they are describing. Our work takes on the metadata of PDF files, we have conducted a large scale study of PDF files published by the security agencies (39664 files) and the scientific community (555865 files).

Sanitization is the obvious choice to deal with the hidden data before sharing or publishing a document. However, our work shows that most organizations are unaware that they need to sanitize their files. We investigated how the hidden data of PDF files can be exploited and also if they are sanitized. We focus on answering two questions:
(i) What can be done using the hidden data of PDF files?
(ii) Are there any organizations that sanitize their PDF files?
To answer these questions, we collected PDF files from security agencies and scientific community and analyzed them.

For scientific community we have conducted a large scale study of 555865 PDF files published. It was now possible for us to extract the metadata and analyze them. We observed that 99.6% of the PDF files included metadata information on the tool used to create PDF file. 23% of the file contain other valuable information on the authoring process, it included information on the organization of the authors, information on the Operating System, the tool managing the references or the software used to create the file itself. We realized that the metadata field names were already giving a lot of information. During our analysis, we also observed that PDF sanitization was not popular. We only found one PDF file poorly sanitized (out of

[6] `https://arxiv.org/`
[7] `https://hal.archives-ouvertes.fr/`

555865). PDF sanitization is an important question. Recently, several conferences like ISCA 2020[8] ask authors to clean their PDF file before submitting. The conference chairs of these conferences are clearly aware that PDF files can contain invisible re-identifying information and want to enforce sanitization.

For security agencies, we have crawled the websites of 75 security agencies of 47 countries and collected 39664 PDF files. For the majority of the files (76%), we were able to recover the authoring process: we identify the PDF producer tool and the Operating System (OS) used by the file's authors. Collecting and analyzing PDF files from the same source over several years can reveal the habits of a given employee. It is possible to learn if he/she update/change (or not) their software regularly. This kind of information is particularly interesting for a hacker to target an individual with bad software habits. By analyzing the PDF files published by several employees of the same agency, it is possible to learn the software policies of an agency. We found at least 19 security agencies in our dataset who are using the same software over a period of 2 years or more. We have observed that, only 8% of all the PDF files have been properly sanitized. We have identified 7 agencies which sanitize their PDF files before publishing. However, our analysis shows that the sanitization method used was weak for 65% of the sanitized PDF files. It was possible to recover sensitive information from these files. Only 3 agencies are reaching a satisfying sanitization level.

Further research on PDF files involved, identifying how a PDF file has been created. We focus on the following forensics problem: is it possible to determine how a PDF file has been created using the file itself? This problem is important because PDF files are extremely popular: many organizations publish PDF files online and malicious PDF files are commonly used by attackers.

A PDF producing tool detector has applications in offensive security and in incident response. In offensive security, it can be used to determine which software is used by an author or by an organization to create and view PDF files. The attackers can find vulnerabilities corresponding to the PDF viewer identified. Then, the attacker can craft and send malicious PDF files to the organization thanks to the knowledge obtained from PDF files. In incident response, a PDF producing tool detector is valuable to understand how a malicious PDF file has been created. It is a useful step toward an attack attribution. The most simple approach to design a PDF producing tool detector consists to look at the file metadata. By default, PDF producer tools put many information in the field Creator and Producer of the file's metadata. It is possible to find the name of the producer tool and its version as well as details on the Operating System.

[8] `https://www.iscaconf.org/isca2020/submit/guidelines.html`

Unfortunately, metadata are not a reliable source of information: they can be easily modified using tools like exiftool or using sanitization tools like Adobe Acrobat.

We have designed a robust PDF producing tool detector based on the coding style of the file. The PDF standard [56] defines the language that is supported by PDF viewers. Developers of PDF producer tool have their own interpretation of the PDF language. Therefore, it is likely that their coding style is reflected on the output of their PDF producer tool. There are coding style elements [58] in PDF files which can be used to identify the producer tool. To observe the coding style of PDF files, we have created a dataset of 900 PDF files using 11 popular PDF producer tools. We have compared different files to identify the pattern in each section of the PDF files. We created 192 rules in regular expression engine to identify these patterns and detect the PDF producer tool. Then, we tested the efficiency of our detection tool on PDF files of security agencies and scientific community pre-prints. We were able to detect PDF files created by `LibreOffice` and `PDFLaTeX` tool with an accuracy of 100%. PDF files created by `Microsoft Office Word` and `Mac OS X Quartz` were detected with an accuracy greater than 90%. More generally, it correctly detected the producer tool of 74% of the PDF files in our dataset.

1.3 Outline of the dissertation

The remainder of the dissertation is structured as follows. In Chapter 2, we describe ambiguity around personal data. We mainly focus on two identifiers: Name and IP address. Chapter 3 covers the analysis on the PDF files, mainly the hidden information found in PDF files, their sanitization and forensics using coding style. Chapter 4 summarizes the dissertation and provides future prospects and the impact of the research performed.

2 Personal Data

Abstract: Personal data is an important concept for any one around the world and especially for computer scientists. It is mandatory for organizations to determine if they are collecting and processing personal data. If it is the case, they have to enforce regulations on the personal data like the GDPR. Unfortunately, the existing definition of personal data is vague and difficult to grasp. To illustrate the difficulty to determine if some data is personal or not, we have studied two important cases: full names and IP addresses. By law both full name and IP address are obviously included in the definition of personal data. However, the reality is far more complicated and many elements need to be taken into consideration.

Privacy laws have never been as important as they are today, nowadays data travels the world through borderless networks. The concept of personal data has become popular as information technology and Internet has made it easier to collect personal data leading to a profitable market [52]. The definition of personal data varies across different countries and since the online market is based on the personal data, it is complicated to regulate this market. For instance, EU has the General Data Protection Regulation (GDPR) and e-privacy directives [107] to define personal data while U.S. relies on a `combination of legislation, regulation and self-regulation` rather than government intervention alone. US has approximately 20 industry or sector-specific federal laws and more than 100 privacy laws at the state level[1]. This all looks like a maze.

After brexit[2], currently EU is comprised of 27 countries and all these countries are governed by common economic, social and security policies. General Data Protection Regulation (GDPR), is one of the most strict regulation on the personal data. The territorial scope of the GDPR is determined by Article 3 and it reflects the legislator's intention to ensure comprehensive protection of the rights of data subjects in the EU. It also establishes scopes in terms of data protection requirement, for companies active on the EU markets.

> Pursuant to Article 3, Section 2 [44, 46] of the GDPR, the processing of personal data is defined for the data subject and also the data controller. The term **data**

[1] `https://i-sight.com/resources/a-practical-guide-to-data-privacy-laws-by-country/`
[2] `https://www.gov.uk/transition`

subject refers to any individual person who can be identified, directly or indirectly, via an identifier such as a name, an ID number, location data or via factors specific to the person's physical, physiological, genetic, mental, economic, cultural or social identity. In other words, a data subject is an end user whose personal data can be collected (Article 4.1 of the GDPR). And the term **data controller** refers to the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law (Article 4 (7) of the GDPR). In short, the individual or legal person who determines the purposes for which and the means by which personal data is processed.

There are two main questions in regards to the territorial scope of GDPR:
i) Who is concerned by the GDPR regulation?
ii) Who has to comply with the GDPR regulation?
Answers to both these questions have been established in Article 3 GDPR. It not only provides the territorial scope but also extraterritorial effects.

i) Who is concerned by the GDPR regulation?
The data subjects are concerned, Article 3 covers the processing of data of data subjects. The territorial scope of GDPR obviously protects individuals (data subjects) who are in the EU and its extraterritorial scope also protects individuals living outside the EU when their data is processed by a branch in the Union.

ii) Who has to comply with the GDPR regulation?
Data controllers need to comply. The GDPR provides uniform rules between companies located inside and outside Europe in Article 3 of the legislation.

1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not.

2. This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union.

3. This Regulation applies to the processing of personal data by a controller not established in the Union, but in a place where Member State law applies by virtue of public international law.

Due to the above mentioned extra territoriality of the GDPR, any global business either has to become compliant for all of its users/customers or be able to accurately identify EU residents and enable compliant systems to handle only that subset of the customer base. Building and maintaining two separate information systems is not practical or cost effective and the downside risk of making a mistake is too large to make it acceptable. It has therefore become normal practice for businesses to apply GDPR compliant information systems to all its users, regardless of location[3].

2.1 What is personal data in Europe?

Defining personal data is already a challenge. In general, the concept of personal data encompasses any information by which a person identifies himself or can be identified by others. However, it is not always clear what really falls under the category of information "based on which a person can be identified". *Personal data is any information relating to an identified or identifiable natural person (referred to as data subject): an identifiable natural person is one who can be identified, directly or indirectly.* In particular by reference to an identifier such as a name, an identification number, location data, an online identifier or association to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Article 4(11) of the GDPR [94]). Personal information is not strictly objective information or facts. Subjective information, including opinions, judgments or estimates is also considered personal data [43].

The concept of *personal data* was set by GDPR as: ***Any information relating to an identified or identifiable natural person (referred to as data subject): an identifiable natural person is one who can be identified, directly or indirectly.*** There has often been confusion around what is personal data and hence the definition of personal data under the GDPR is very broad, far more than most of the other country's current or previously existing personal data protection laws.

Prior to the GDPR, Data protection authorities formed the Article 29 Working Party (Art. 29 WP). It was the independent European working party that dealt with issues relating to the protection of privacy and personal data, it ceased to exist as of 25 May 2018 (entry into

[3] `https://www.gdpreu.org/the-regulation/who-must-comply/`

application of the GDPR). Currently the European Data Protection Board (EDPB)[4] functions as an independent European body, which contributes to the consistent application of data protection rules throughout the European Union and promotes co-operation between the EU's data protection authorities. The opinion [10] decomposes the definition of personal data into four building blocks: *any information, relating to, an identified or identifiable and natural person*.

The term **any information** clearly signals the broad concept of personal data. This means that both objective and subjective information about a person in whatever capacity may be considered as "personal data". For instance, it covers "objective" information, such as the presence of a certain substance in one's blood. It also includes "subjective" information such as opinions or assessments.

The term **relating to** plays a crucial role in determining the scope of the concept, especially in relation to objects and new technologies. This covers information that may have a clear impact on the way in which an individual is treated or evaluated. For instance the data registered in one's individual file in the personnel office are clearly `related to` the person's situation as an employee. So are the data on the results of a patient's medical test contained in his medical records.

The term **identified or identifiable** focuses on the conditions under which an individual should be considered as `identifiable` and especially on `the means likely reasonably to be used` by the controller or by any other person to identify that person. The particular context and circumstances of a specific case play an important role in this analysis. The opinion also deals with `pseudonymised data` and the use of `key-coded data` in statistical or pharmaceutical research. For instance, the Working Party provides a scenarios where *asylum seekers hiding their real names in a sheltering institution have been given a code number for administrative purposes. That number will serve as an identifier, so that different pieces of information concerning the stay of the asylum seeker in the institution will be attached to it and by means of a photograph or other biometric indicators, the code number will have a close and immediate connection to the physical person, thus allowing him to be distinguished from other asylum seekers and to have attributed to him different pieces of information, which will then refer to an "identified" natural person.*

The term **natural person** deals with the requirement that `personal data` are about `living individuals`. The opinion also discusses the interfaces with data on deceased persons, unborn children and legal persons. For instance, Information relating to dead individuals is therefore

[4] `https://edpb.europa.eu/about-edpb/about-edpb_en`

in principle not to be considered as personal data subject to the rules of the Directive, as the dead are no longer natural persons in civil law.

Under the GDPR, any information relating to an natural person currently encompasses both direct and indirect identifiers. Direct identifiers are those information that explicitly identifies a person (such as a name, a social security number or biometric data). Indirect or quasi identifiers refer to the information that can be combined with some additional data to identify a specific person (such as the combination of gender, birth date, geographic indicator, medical records or financial information etc..).

In the rest of this chapter we mainly consider two identifiers: name and IP address. Based on the context, these identifiers behave both as direct and indirect identifiers. With various examples, we describe scenarios where `name and IP address` information constitutes to be personal data and identify a specific person.

2.2 Are names personal data?

The word `name` comes from old English `nama`[5] and it is traditional for individuals to have a personal name. A name is basically used for the identification of an individual, it can be a word or set of words by which a person or thing is known, addressed or referred. A person's full name is usually used to identify that person for legal and administrative purposes. Although sometimes we see that, it may not be the full name by which an individual is commonly known. Some use only a part of their full name or some are known by their titles, initials, nicknames, aliases or other formal or informal designations.

As long as there has been languages, there have been names. Naming differentiate one person from others and mostly every society has its own naming system. To demonstrate this with an example, let us consider French naming system. First names in France are usually chosen by the parents of the child. Nowadays there are no legal a priori constraints on the choice of the first name, but back in 18th century choice of given names was originally restricted by law to only the tradition of naming children after a small number of popular saints[6]. Later in 1966 a new law allowed a limited number of mythological, regional or foreign names. And finally in 1993, French parents were given the freedom to name their children without any constraint.

[5] `https://www.etymonline.com/word/name`
[6] `https://en.wikipedia.org/wiki/The_New_Zealand_Herald`

When we consider the last name the use of last name or surname in France, like in much of Europe, it didn't become necessary until the 11th century. Coming from the medieval French word `surnom`, which translates as `above-or-over name`, it became necessary to add a last name to distinguish between individuals with the same first name[7]. Initially, it was easy to adopt any last name people wished until 1474 and then the king of that period decreed that all the last name changes had to go through his approval. From then on, all name changes were recorded which also helped in making it easier to trace family history. French last names can be grouped into different types: Patronymic/Matronymic surnames, Occupational surnames, Descriptive surnames, Regional surnames, Alias surnames or Dit Names and French Names With Germanic Origins etc.. We describe three of these surnames in order to show different ways these surnames are formed.

Patronymic and Matronymic Surnames: A very common method to construct the surnames was to use the names of parents. Patronymic surnames are usually based on the father's name and matronymic surnames are based on the mother's name. Traditionally it was a practice to only use father's name but when the father's name was unknown mother's name was used. Patronymic and matronymic surnames are mostly direct derivations of the parent's given name (Louis Robert, for "Louis, son of Robert"). **Occupational Surnames**: It was also very common to distinguish individuals by referring to their jobs or trades (Thomas Boulanger for "Thomas, the baker"). Several common surnames based on the occupations found as French surnames include Caron (cartwright), Fabron (blacksmith), Berger (shepherd), Charpentier (carpenter) and Chevrolet (goat farmer) etc.. **Descriptive Surnames**: Some surnames were constructed based on the individuals appearance, unique qualities and also from their nicknames or pet names (Jacques Legrand, for Jacques, "the Big"). Some common examples include Petit (small), Brun (someone with brown hair or a brown complexion) and LeBlanc (blonde hair or fair complexion) etc..

Like France, all the other countries around the globe have their own system of naming individuals. And hence many a times our names identify us both as individuals and also as the members of a particular group or community. One might think that someone's name is as clear an example of personal data as it gets, it is literally what defines a person. But it's not always that simple, to illustrate this we have considered first and last names in the following sections.

[7] `https://www.thoughtco.com/french-surname-meanings-and-origins-1420788`

Country	Top first name
Spain	Lucía (female) Marc (male)
Germany	Mia (female) Ben (male)
Denmark	Ida (female) William (Male)
Sweden	Alice (Female) Oscar (Male)
France	EMMA (female) GABRIEL (Male)

Table. 2.1: Popular first names in Europe.

2.2.1 First names

We have considered most popular first names in EU. Table 2.1 shows the most commonly used first names[8] across EU. Statistics for the number of people owning these names were not available. The countries in Europe have their own naming system and from Table 2.1 we can notice that popular names differ across countries. Since EU comprises many countries, naming differences across countries makes fewer people owning common first names. This difference in naming system makes the application of GDPR regulations easier for many people in EU.

Are first names unique in France?

INSEE maintains the records of all the first and last names of French citizens. **INSEE**[9] is a Directorate-General of the Ministries for the Economy and for Finances and is located in offices throughout the French territory. It was created by the Budget Law on 27 April 1946. Each year, INSEE estimates the population of the regions and departments (metropolitan France and DOM) as of January 1st. These annual population estimates are broken down by sex, age and other informations. These databases are available for download and it provided us with the list of french names.

Our first download from INSEE includes the statistics for the first names[10].This dataset, i.e, the national data file contains the first names given to children born in France (excluding Mayotte) between 1900 and 2017. The dataset provides the sexe (gender), preusuel (first name), annais

[8] https://en.wikipedia.org/wiki/List_of_most_popular_given_names#Europe
[9] https://www.insee.fr/fr/accueil
[10] https://www.insee.fr/fr/statistiques/2540004

(year of birth), dpt (department) and nombre (total number of people) information. Table 2.2 provides an example to the sample entry in the INSEE file, there are 4 people born in 2016, owning the name LOU in department 94 and their gender is male (sexe-1). Total number of entries in our dataset for first names are 35,73,026.

sexe	preusuel	année	dpt	nombre
1	LOU	2016	94	4

Table. 2.2: First name sample from INSEE database.

Using the list of first names, we computed top 10 male and female names registered from 1900 to 2017 as shown in Table 2.3.

Popular male first names	
JEAN (1918735)	PIERRE (890565)
MICHEL (820224)	ANDRÉ (711946)
PHILIPPE (538288)	RENÉ (516471)
LOUIS (513403)	ALAIN (506839)
JACQUES (482560)	BERNARD (469278)
Popular female first names	
MARIE (2234359)	JEANNE (554164)
FRANÇOISE (401499)	MONIQUE (399843)
CATHERINE (394676)	NATHALIE (382841)
ISABELLE (377884)	JACQUELINE (372550)
ANNE (364946)	SYLVIE (364645)

Table. 2.3: Top 10 popular first names and number of people owning these names in INSEE dataset.

We also computed the rare names and there are at least 1042 people (486 males and 556 females) using rare names in our dataset. Each of these rare names are used by just 20 people across complete France (distributed in different regions). Since the dataset includes information on the gender, year of birth and number of people owning the name in particular department, when the first names are rare or unique, using these quasi identifiers it is possible to identify a specific person.

2.2.2 Last names

After the first name analysis, we have considered most popular last names in EU. Table 2.4 shows the most popular surnames[11] across five EU countries. We can see that Garcia is the most popular surname and it is used by nearly 1.3 million (3.5% of Spain's population) people.

[11] https://europeisnotdead.com/europeans-surnames/

To put the results of last names in EU into a prospective, we consider the last names in France to illustrate there uniqueness.

Country	Surname	# people
Spain	Garcia	1,378,000
Germany	Müller	320,000
Denmark	Jensen	288,050
Sweden	Johansson	265,000
France	Martin	240,000

Table. 2.4: Popular last names in Europe.

Are last names unique in France?

Using the statistics available for last names in France, we computed their uniqueness. We downloaded the last name statistics from INSEE[12]. The last name dataset includes all the registered last names from the year 1891 to 2000. Different identifiers in the dataset are NOM(last name), DEP(department), année (year), nombre(count). Table 2.5 shows an entry in the INSEE file. The last name ABRY in department 1 was owned by 2 people from 1891-1900 and so on. . . We observed that the department information is not available for nearly 25% of the last names. Total number of entries of last names in the INSEE file are 521517.

NOM	DEP	1891-1900	1901-1910	1911-1920	...	1981-1990	1991-2000
ABRY	01	2	15	13	...	6	4

Table. 2.5: Last name sample from INSEE dataset.

Using this dataset we computed the most popular names used in France as shown in Table 2.6. Since these last names are very popular, they do not constitute personal data *at all times*. In this case, to identify a specific person, quasi identifiers like first name, department, gender, date of birth etc.. are used.

Popular last names	
DUBOIS (108619)	LEFEBVRE (91459)
FONTAINE (70000)	MULLER (64309)
BOYER (61672)	MARIE (48635)
PAYET (37410)	GRONDIN (25104)
HOARAU (23868)	HOAREAU (17719)

Table. 2.6: Top 10 popular last names in INSEE dataset.

[12] https://www.insee.fr/fr/statistiques/3536630#consulter

After computing the popular last names, we computed the last names used by only one person in France. We found that at least 1090 people can be solely identified based on their unique last name in France. Table 2.7 shows the year when these last names were registered, we can observe that from 1961 to 200 more and more unique last names have been registered. The department information for these 1090 names is not given in the dataset. Using the information in Table 2.7, one can also determine the age group of the person with the unique last name. These unique last names are personal data that leads to the identification of an individual solely using their last name in France and may be in the world.

Year	Number of unique last names
1891-1900	42
1901-1910	39
1911-1920	59
1921-1930	65
1931-1940	52
1941-1950	62
1951-1960	49
1961-1970	96
1971-1980	132
1981-1990	189
1991-2000	305
Total	1090

Table. 2.7: Number of people identified by their unique last names in France.

No doubt that unique names make the application of GDPR easier for data controllers to verify a data subject but our analysis of last name and first name in France also show that, unique last names and first names impose higher privacy risks and these individuals are easier to identify. In the following section we demonstrate how the privacy risks are higher when the full names are unique.

2.2.3 Uniqueness of a full name

A person's full name is probably the most obvious example of personal information. We have seen from the analysis of first and last names that a name can be shared by more than one person in the same region, department, country etc.. In such cases, technological advancements highlight the difficulties in sustaining GDPR in its entirety both EU and outside. For instance, let us consider the definition of extra territoriality scope of GDPR, it clearly says:

> Data controllers which have activities in an establishment in the European Union have to enforce the GDPR if they process personal data regardless of whether the processing takes place in the Union or not and regardless of the nationality of the data subjects. It is true wherever the data subjects are located. This is the first effect of extraterritoriality.

In regards to this regulation, if we consider a country like China, by itself, it still lacks rules and regulations in user information management and it does not have a good supervision system. Recently in article [89], it has been demonstrated that China gradually builds a data privacy system through the legal transplantation of both the EU and the U.S. reference models. Since China contributes to nearly 18.6% of the world's population, there exists a possibility that many people share the same first name and last name.

For instance, the last name `Garcia` is the most popular in EU and it is used by nearly 1.3 million (3.5% of Spain's population) people. When we compare this number with that of China, it is way lesser than the most popular surname `Wang` which is used by more than 100 million people in China. This comparison shows the complications in the application of GDPR to identify a data subject.

China was one of the first countries in the world to adopt last names, dating back more than 5000 years. With 1.37 billion citizens, China has the world's largest population, but has one of the smallest last name pools. According to the Ministry of Public Security[13], only about 6000 surnames are in use. Majority of the population constituting to almost 86%, share just 100 of last names from 6000. According to the national name report released by the Ministry of Public Security, `Wang, Li, Zhang, Liu` and `Chen` remained the top five common surnames in China in 2019 and 2020. According to government figures[14], these five surnames are shared by more than 433 million people or 30% of the total population registered in China.

Many people in China share the same full name[15]. For instance, full name `Zhang Wei`, `Wang Wei`, `Wang Fang` and `Li Wei` are the most popular names shared in China as of 2019 (Table 2.8).

When the organizations situated in EU deal with data subjects located in China, the question raised would be: *Is it possible to efficiently and fairly apply GDPR and fulfill request of a data subject in China?*

[13] `http://www.china.org.cn/english/government/130485.htm`
[14] `http://gat.ah.gov.cn/public/7081/40201012.html`
[15] `https://www.theworldofchinese.com/2014/07/the-most-popular-names-in-china-not-a-john-smith-in-sight/`

Names #	individuals (2019)
Zhang Wei	294.282
Wang Wei	287.101
Wang Fang	271.550
Li Wei	266.037

Table. 2.8: Most popular full names in China.

When a data controller situated in EU gets a subject access request from a individual in China, there exists a possibility of first and last name collisions. In such cases the data controllers would simply ask their birth date to the subjects alongside their full names to remove any ambiguity to the data subject's identity. This looks like an ideal solution as full names in these cases do not identify an individual. But this situation is much more complicated w.r.t Chinese naming system.

In the article [2], authors show how collisions in names affect the GDPR and subject access right. They showed that the naming system in China can creates opportunities for impersonation attacks and for denial of access. They illustrate this situation with an example of the name *Wang Wei*. Between 2007 and 2019, there were around 5533 Wang Wei born in China. Assuming that the births are equally distributed over the years, it gives 461 birth per year on an average. Due to the Pigeonhole principle, people named *Wang Wei* having the same birth date exists and makes the verification process tedious for the data controller. This situation is similar for many names shared by Chinese people. Applying GDPR and fulfilling a subject access request to a data subject located in China with common names is more difficult.

To put naming into a perspective, let us consider the situation in EU. We considered first names and last names in EU (no statistics is available for full names). Compared to China, very less % of the population share common names in EU. In the rest of this section, we provide an example of french naming system to show the uniqueness of names w.r.t to a specific country in EU.

2.2.4 Visualization of the Uniqueness of French first & last names

Using the data from INSEE, we created a tool to visualize the uniqueness of French names. Along with the uniqueness of a name in France, our tool also shows the impact of privacy associated to the names in France.

Figure. 2.1: Result for the uniquness of first name ADEN in Department 38.

The first name dataset from INSEE includes entries till year 2017 and provides more information (gender, department etc..) while the dataset for last name is available only till year 2000 (statistics are not yearly but for decades) and department details for 25% of the last names are missing. Hence our tool provides 2 different types of results: One for the uniqueness of a first name in a particular department and second to check the uniqueness of last name in complete France.

In order to check the uniqueness of first name, the user needs to enter a first name and the department number (ex: 38, 75 43...). Figure 2.1 shows the output for the first name `ADEN` in department 38. We can see in the figure 2.1 that the name `ADEN` is in the red color within the wordcloud. This means that `ADEN` name is a rare and very few people in department 38 own this name. The individual with this name should be careful before sharing the first name along with other details on Internet as it has higher privacy issues compared to some name like `ELIO, Luca etc.. that are very common in this region`. Similarly Figure 2.2 shows another result for the name `ADAMS` in department 971, we can see that this name is very rare in this department and risks associated are higher.

Using our tool, an individual can also check the uniqueness of a particular last name in complete France. Figure 2.3 shows the results for the last name `ACUTI`, this is a rare last name in France and hence has higher privacy impact.

The individuals owning unique names are more impacted when they share there names online along with some other information. For instance, if an user with unique/rare name creates an email or username and also provides some additional information within their username then it is easy to distinguish him/her from others and this impacts his/her security and privacy.

Figure. 2.2: Result for the uniquness of first name ADAMS in Department 971.

Figure. 2.3: Uniqueness of last name in France.

To illustrate this scenario, we have performed some experiments on the impact of sharing name and other personal information online (using usernames). Our analysis and results are described in the next section.

2.3 Are usernames personal data?

Due to the advancement in technology, Internet has become more central in our day-to-day lives. As the number of users are increasing in the digital world, the need to have a personalized account is increasing too. We as individuals interact with a number of online websites and services, sometimes these websites and services require an user account (username and password) to access. The online accounts created by the individuals become their online identity while accessing websites and services. Unlike the actual full name, online users can have different usernames around websites. One username could be shared by several users on different websites and one user can have multiple usernames on one or more websites.

During our analysis, we observed that, the awareness to protect the password is widely spread and instructions to create a strong and safe password is enforced by several online websites. Many websites impose creation of password with a combination of alphabets, symbols, numbers... in order to create a strong password. Password meters are also used by several websites to show the strength of the chosen password, so that the user takes care to create a strong password. However, we observed that no emphasis is provided to create a strong username.

We analyzed several websites to check if creation of username has been taken into consideration to be strong and safe. Our analysis showed that websites providing options to have user accounts, do not provide any information for the creation of safe usernames. Username carries information that reflect an individual's characteristics, tastes, habits etc., this is especially true while using social media and gaming sites where users interact with one another. According to Warren and Brandeis (1890), disclosure of private information and the misuse of it can damage people's feelings and cause considerable damage in people's lives [29]. We studied usernames to check different kind of information used. In our work, we identified personal information from usernames and show how using personal information within username can lead to compromising the privacy of online users.

2.3.1 Guidelines to create usernames

Many websites require users to create unique account. Some websites have introduced tighter security policies around usernames. In some cases, websites determine if a username is unique and unless it is, user cannot set up an account. These measures are taken in order to protect user information. If the desired username is already taken, then these websites suggest to add favorite things to the username to make it unique and available. For instance one of the suggestions found online says: *Your username is your identity online. Whether you're posting on forums, editing a wiki, playing games or doing any other online activities that involves interacting with others, your username will be the first thing other people see. People will make assumptions about you based on the name you choose, so pick wisely!*[16]. After reading several suggestions posted on websites, the common suggestions provided are grouped as follows:

1. Know that your username represents you, your username is going to be the first thing people see when they interact with you online.

2. Make sure that you like your username, because you'll be seeing it a lot.

3. Tap into your interests and consider things around you to include in the username.

4. Cross the language barriers. Look up words in other languages.

5. Keep it short and simple. If you're going to be typing in your username on a regular basis, you'll appreciate a shorter name.

6. Try a name generator. There are a variety of random name generators available online.

7. Don't choose a username that gives clues to your passwords and don't use the same username and password combination, especially on financial accounts.

8. You may want to break down your Internet usage into two different categories: professional and personal interest. You can then use one username for all of your professional websites and then use one username for all of your personal interest sites. This will make it easier for you to remember your usernames.

[16] `https://www.wikihow.com/Create-a-Username`

9. Stay anonymous, avoid using any personal information like your first name or last name or your birthdate when creating your username. Use a variation of your name so that it is easier for you to remember but difficult for others to associate it to you.

Some of these suggestions let users know that they need to be careful while creating usernames but some suggestions (mainly first four) can lead a user to create username that will compromise his/her privacy. A study by NordPass revealed the top 200 most popular usernames[17]. The point behind this research was to show the tendency to use real names as a username. After the research on usernames, NordPass security expert Chad Hammond offered some tips for users to create secure usernames. These tips include:

1. Don't just use your name as a username.
2. Avoid using the beginning of your email address as your username.
3. Your username should be simple enough to remember but hard to guess
4. Never use easy-to-guess numbers with your usernames (for example, address or date of birth).
5. Don't use your Social Security number or ID number as your username.
6. If you're struggling, try an online username generator.

During our analysis we found that, websites that let user create a personalized account do not provide any instructions mentioned above to create a secure username (unlike for passwords). Many users are unaware that the information leaked in their username can compromise their security and privacy. We observed that some sites especially username generator sites provide the awareness to have a secure username[18] but these sites need to be explicitly searched by a user and many users do not look for such information.

2.3.2 Issues with username

Username can be a rich source that discloses one or more information like user's first name, last name, age, date of birth, department they live in, country, their profession, gender and some

[17] https://nordpass.com/blog/all-time-most-popular-usernames/
[18] https://leapfrogservices.com/why-usernames-are-important-and-how-to-choose-good-ones/

personal traits... Username can be used to create a accurate profile of a user's demographic category, personal preferences etc.. These information are enough for a hacker to start a social engineering attacks.

Pervious work done in [105] shows how using some personal information an individual can be identified. Her research found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on three quasi identifiers (5-digit ZIP, gender, date of birth). About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only place, gender, date of birth, where place is basically the city, town or municipality in which the person resides. And even at the county level, (county, gender, date of birth) are likely to uniquely identify 18% of the U.S. population.

On january 1st 2014, Snapchat[19] was hit by a data breach and attackers downloaded 4.6 million usernames along with the phone numbers. As many people used their own names or surnames as their usernames, they were easy for the hackers to identify. Cyber-criminals having access to username and other information can launch social engineering attacks. Previous work in [88], showed the possibilities to link online profiles using only the usernames. It is possible that, if a user is registered on several websites with a single username, the accounts could be linked and some information could be gathered. This work was done to link the profiles of Google and eBay accounts. These previous attacks and studies show that, hackers can misuse the username information for identity theft, unauthorized access, misuse of personal information and stalking and profiling.

Figure 2.4 shows different types of informations that can be obtained on a username. We considered a random username `john38` and show different information that can be predicted on this username. Details like gender, place, DOB... information can be inferred from usernames.

2.3.3 Linking online usernames across several websites

Users join multiple social media platforms and their profiles across these platforms can be linked using different methods [92] to obtain their interests, locations, content and friends. Altogether, this information can construct a person's social profile. Assuming that we already know the username(s), we found different but very similar tools with which one can trace an username across multiple platforms like *Twitter, Github, Facebook, Pinterest, Instagram*.... Out of many available tools, we used eight tools and describe our observation below.

[19] `https://www.zdnet.com/article/predictably-snapchat-user-database-maliciously-exposed/`

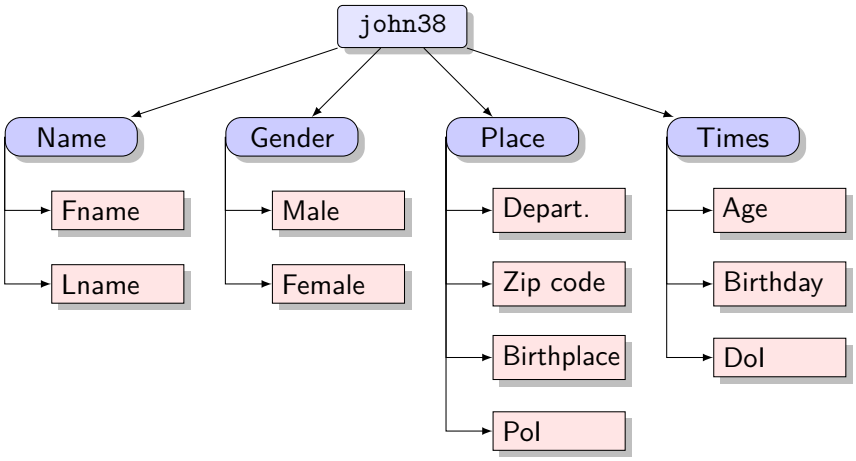

Figure. 2.4: User information that could be extracted from username.

WhatsMyName[20] - This tool allows to enumerate usernames across many websites. It provides list of all the websites and also links to check the account (without logging in) and an option to download the website names with link associated to username. One can also refine their search as gaming, coding, dating... For example, if someone is a gamer, he/she may want to check only the gaming sites to see if their desired username is available or not.

Namecheckup[21] - This tool was built to allow users to check the availability of social media usernames and domain names across many platforms. Search result display the grey colored boxes, these are the platforms where the username is taken. To access the account one has to simply click on a greyed box of interest and proceed with verification steps to login.

Sherlock[22] - a powerful command line tool provided by Sherlock Project, that can be used to find usernames across many social networks. It also allows to make requests over a proxy, a list of proxies and even the TOR network.

Instant Username Search[23] - Instant Username Search helps users find out if their username is taken on more than 100 social media sites.

namech_k[24] - a tool to see if desired username or vanity url is still available at dozens of popular Social Networking and Social Bookmarking websites.

[20] https://whatsmyname.app/
[21] https://namecheckup.com/
[22] https://sherlock-project.github.io/
[23] https://instantusername.com/#/
[24] https://namechk.com/

checkusernames[25] & **Knowem**[26] - these two applications allows to check for the use of ones brand, product, personal name or username instantly on over 500 popular and emerging social media websites.

social-searcher[27] - allows to search for content in social networks in real-time and provides deep analytics data. Users can search without logging in for publicly posted information on Twitter, Google+, Facebook, Youtube, Instagram, Tumblr, Reddit, Flickr, Dailymotion and Vimeo.

namecheckr[28] - This tool checks domain & social username availability across multiple networks.

All these tools and many other tools available online were developed to help users and businesses to check availability of usernames across different platforms. But hackers can also use these tools to gather information and target victims.

Privacy problems associated to online users has become a major concern among Internet users over the past decade. The emergence of social networks has even increased these concerns. People create accounts in social networks to be connected to their friends, family and for several other reasons. They also share messages, pictures, videos and their thoughts on certain subjects etc.. they perceive this as a great deal in terms of social interaction, friendship, carrier and other opportunities.

A lot of information are voluntarily shared on online social networks and many people rest assured that different social network accounts on different platforms won't be linked as long as they don't grant permission to these links. However, according to Diane Gan, information gathered online enabled "target subjects to be identified on other social networking sites such as Foursquare, Instagram, LinkedIn, Facebook and Google+, where more personal information was leaked [40]". This is concerning, as attackers/burglars can use social media networks to target their victims.

2.3.4 Inferring personal attributes from usernames

To analyze the usernames, INRIA Privacy MOOC's Fun platform provided us with the dataset. The dataset consists of two sessions. Session 1 was opened from January 2018 to March 2018

[25] https://checkusernames.com/
[26] https://knowem.com/
[27] https://www.social-searcher.com/
[28] https://www.namecheckr.com/

Username	YoB	Gender	Education	City	Country	Goals
claire75	1994	f	m	Paris	FR	Merci...
MS24	1985	m	jhs	None	NL	PhD

Table. 2.9: MOOC Dataset - Information provided by users.

with 13,211 participants and Session 2 was opened from November 2018 to January 2019 with 9,095 participants. There were 1645 participants who attended both sessions. After the intersection of participants from two sessions, the final dataset (from now on addressed as MOOC dataset) consists of 20,661 unique participants. Table 2.9 shows the example of different fields in the dataset. MOOC users have provided this information during the creation of their accounts, we do not know the ground truth about the data provided, users might have filled or provided some invalid information.

During our analysis we observed that usernames can include some personal data that could be used to identify an individual directly, indirectly or some times it include the quasi identifiers that provide more information that helps to identify an individual. We made predictions and extracted some personal data on the MOOC users using their usernames. All the results obtained on the username were then validated using the different information provided by these users.

Nationality and gender prediction using username

Using the existing tools for ethnicity and nationality prediction `Ethnea` [108] , we predicted the nationality or cultural interest from the usernames (e.g., ikizen25. kizen is 94% Japanese name). Using the gender prediction tool, genderizer [29] we predicted the gender. Table 2.10 shows the predictions for nationality, we managed to find 731 user's nationality correctly using the tool `Ethnea`. Table 2.11 shows the predictions for the gender, we can see that just using username, 2436 user's gender was correctly predicted.

Nationality Prediction - Ethnea Tool				
Nationality	#users	Correct predictions	Wrong prediction	Unknown
French	910	687	142	81
Others	2792	44	2627	119

Table. 2.10: Nationality prediction using usernames.

[29] `https://genderize.io/`

Genderize.io API				
Gender Prediction	#users	Correct predictions	Wrong predictions	Unknown
Male	1711	1099	388	224
Female	1991	1337	467	187

Table. 2.11: Gender Preiction using username - Genderize.io.

Religion Prediction	
Religious names	#users
Christian	129
Jewish	38
Christian/Jewish	13
Muslim	32

Table. 2.12: Religious details using top 10 religious names.

Religious association to the usernames

Some names like Mohammad, Abdul are usually associated to Muslim families and David, Alice to Christian families. In order to predict the association of names to different religions, we choose top 10 `Cristian, Jewish and Muslim` names and developed a tool using the information available on Wikipedia[30]. Table 2.12 shows our predictions for 3 different religious groups (Note: we do not claim that these users are followers of some particular religion, there names are associated to some religion). In our dataset we found 212 users whose names could predict their religious background or their ancestral family practices.

Prediction of places

From the information provided by the MOOC users, we noticed that more than 70% of the users registered in MOOC are from France, it was interesting for us to work on dataset targeting french users. We noticed that several usernames in our dataset include numbers along with name. As shown in Figure 2.4, a 2 digit number could either be age, date of birth, postal code, department etc.. We used 2 digit and 5 digit numbers present in the username to find the department/city of the user.

We observed that some of these numbers are the actual postal code and these users belong to these regions too, the results obtained are as shown in Table 2.13. Using the department code (2 digit number) may be less risky than using the complete postal code (5 digit number). We were able to find the city of 15 MOOC users using the 5 digit numbers. All these predictions

[30] `https://en.wikipedia.org/wiki/List_of_biblical_names`

were evaluated using the information provided by the user. Some of these regions have very less population, for instance one of the region has, just 0.3% of the total french population (Ex-59680 - 0.2% population, 66400 - 0.1% population etc..). such information can restrict the anonymity with respect to the region the user belongs to and increases the risk of privacy.

Pattern	#users	Correct predictions	Wrong predictions	Unknown
2digit numbers	1418	387	798	233
5digit numbers	99	15	54	30

Table. 2.13: 2 & 5 digit numbers found in usernames.

Predictions using INSEE files

Since more than 70% of our users in MOOC dataset are French, INSEE statistics can be used to predict information on the users. Using our INSEE dataset of first and last name, we found that at least 3044 users have used their last name in the username and 3661 users have used their first names. We used these names to predict the gender of the user and results are shown in Table 2.14. Using the INSEE dataset we were able to correctly predict the gender of at least 2408 user.

Gender			
INSEE Data	#users	Correct predictions	Wrong prediction
Male	1497	1428	69
Female	1770	980	790

Table. 2.14: Gender prediction using INSEE statistics.

Our methods to extract personal information using online tools and INSEE files could be used by any attackers to perform social engineering attacks. Our analysis on usernames shows that online users are either unaware of the information leaked in their usernames or they think that usernames could not be exploited to target them.

2.4 Are IP addresses personal data?

Internet is a huge network of billions of connected devices communicating with each other on a daily basis to send and receive data. Computers and other devices are connected to a network using either wired or wireless connections. IP addresses are a backbone of Internet. They are used to identify devices and almost every web servers process them in their logbook. Therefore, determining if IP addresses are personal data or not is crucial because it can have a huge impact on the individuals privacy and on the measures required to process IP addresses.

Our work on IP address focuses on the following questions:
i) Is IP address qualified as personal data?
ii) Do companies and DPAs treat IP address as personal data? and
iii) How do companies and DPAs respond to SAR with IP address?

In the following sections, we first explain what is an IP address, different types and distribution. Then we discuss the legal status of IP address, what GDPR says, when is IP address personal data. . . Finally we illustrate how IP-based SAR requests are handled.

2.4.1 Types and distribution of IP addresses

An IP address (Internet Protocol address) is an identifying number for a network hardware that is connected to the internet, it is a `digital address` of a device. Since the devices on internet exchange data, they need to have a digital identity to communicate. Hence, each connected device includes an digital address known as IP address. The laptop, smartphone, smart lights, baby monitor, tablets, thermostat and many other devices that connect to internet has an IP address. As such, in many cases, it is possible to assume a strong connection between a device and its user [101]. Not only the devices connected to the internet but even the websites have there own unique IP address (Google, Amazon, Apple etc..).

IP addresses are used by electronic communications service providers to help identify a subscriber [65]. In the words of the CJEU, IP addresses are *series of digits assigned to networked computers to facilitate their communication over the internet* [27] (parag 15).

Representation of IP address

IP addresses are in binary number format in the form of 0s and 1s. To represent the numbers in human readable format, they are expressed either in the decimal form (IPv4) or hexadecimal form (IPv6). Table 2.15 shows examples of IP addresses both in decimal and hexadecimal form.

IPv4 address	131.412.276.201
IPv6 address	2DAB:FFEF:0000:4DAE:01BA:00FF:DE72:2C2A

Table. 2.15: Representation of IP addresses (examples- IPv4 and IPv6).

Distribution of IP addresses

Internet Assigned Numbers Authority (IANA) [53] is responsible for management of IP addresses. IANA distributes large blocks of addresses to Regional Internet Registries (RIRs) [93]. Currently there are 5 RIR across the globe (AfriNIC, APNIC, ARIN, LACNIC and RIPE NCC). These RIRs are responsible for the management and distribution of IP addresses in there particular regions. Two versions of Internet Protocol, IPv4 and IPv6 are popularly used. The IPv4 (Internet Protocol version 4) defines a 32-bit address space. Distribution of IP addresses using IPv4 is upto 2^32 (around 4.3 billion addresses). However, increasing use of internet and increase in the number of devices on internet has lead to the depletion of available IPv4 addresses. To overcome the issue of depletion, IPv6 (Internet Protocol version 6) has been brought into existence with 128-bit address space and upto 2^128 unique addresses (around 3.403*10^38 addresses).

The representation of IP address using IPv4 and IPv6 slightly differs (Table 2.15). IPv4 consists of a string of decimal numbers that are separated by dots. Each IP address is separated into four segments by three dots. Whereas IPv6 consists of hexadecimal numbers separated by colon. Each IP address is separated into eight segments by seven colons [26] [50]. Compared to IPv4, IPv6 was designed to provide more number of available address spaces, security, simplified packet processing in the router, efficiency and less time consumption.

Types of IP addresses

IP addresses are divided into two main types: Static and Dynamic IP address. As the name suggests static IP address is fixed. Static IP address is manually allocated by Internet Service Provider (ISP), once allocated static IP address will not change. Where as a dynamic IP address is automatically assigned by Dynamic Host Configuration Protocol (DHCP) and it frequently changes when the user connects to the internet.

2.4.2 IP-based SAR

Website tracking is the activity of monitoring its user's movements, interests and behavior on the Internet. Tracking is done using cookies, mouse movements, IP addresses and other website trackers and tools. In this section we briefly explain why users might be interested to submit a IP-based SAR request and the challenges faced due to current IP address allocation schemes.

Why IP-based SARs are important?

Users access website services as an external user, registered user or both. It is very likely that the tracking occurs whether or not they are connected to their registered account. *Is there a way to access the data tracked for the activities of an external user?* For instance, we used *Google website* as both registered and external user and then downloaded the data from google takeout. The downloaded data contains IP addresses and other information about each access as a registered user. But no information was available about the access made as an external user. An user can either think that he/she is not tracked as an external user or no information is available. This leads to question like:

(i) How can we expose tracking when people are not connected or registered?
(ii) Can an user use the right to access under GDPR to get the data from websites?
(iii) Which are the different identifiers that can be used by users to identify themselves to the data controllers?

Studies in the past [41, 116] already showed that IP addresses are used by websites to identify their users and also to track their activities. And hence, IP address seems to be a obvious choice of identifier that users can provide to the data controllers to access the data collected as an external user. This is why we think that IP-based SAR are very important to understand the extent of online tracking.

Why IP address allocation does not help?

When an user visits a website on his/her browser, packets are exchanged between the user device and the server hosting the website. Source and destination of those packets are identified by the IP addresses. These IP addresses can be either IPv4 or IPv6. Studies have shown that IP addresses of the user device provides many user data and geographical location [83, 67].

In our study it is important to understand how IP addresses are allocated to user devices and which IP addresses are seen by the websites. This helps to understand if a data controller should provide data w.r.t to an IP address and the way IP-based SAR needs to be processed. Currently due to the network topology used, many users can share the same IP address. This makes it impossible for the data controller to authenticate a user and fulfill the IP-based SAR request.

The impacts of PETs

In the recent years, there is a growing concern towards online privacy. Virtual private network (VPNs) and Tor[31] are privacy enhancing technologies which can be used to hide the IP addresses used by people to visit websites. This supports the idea that IP addresses are sensitive information for Internet users. When users submit a IP-based SAR using VPN/tor IP addresses, it is impossible for the data controller to identify a user as many people share the same IP addresses using VPN and tor.

2.4.3 Legal Ambiguity- Is IP address qualified as personal data?

An IP address is an online identifier [54, 50] used for identifying devices online and to route information between websites and its end-user devices. IP addresses can be effectively used to track Internet users [75], suchlike cookies [60] or browser fingerprinting [62]. Almost every website collects and processes (maybe under legitimate motivations) IP addresses of their users, mostly for quality of services or security purposes. *IP addresses have a tracking potential but are they considered as personal data by law?* The answer has surprisingly deep consequences for online privacy and on the way websites operate.

- **If the answer is *yes***, then personal data processing must be fair and lawful and ensure that IP addresses are only "collected for specified, explicit and legitimate purposes and not processed for purposes incompatible with the purposes for which they were originally collected" (Article 5(1)(a)(b) GDPR). The processing of IP addresses would then have to be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed" (as stated in Article 5(1)(c) GDPR). Accordingly, if IP addresses are personal data, websites have to change all their mode of operations to comply with the GDPR and so far, it is unclear if it is always possible to implement the necessary changes. Therefore, websites prefer to consider that IP addresses are not personal data.

- **If the answer is *no***, the burden relies on the side of Internet users. Empirical results and the current legal reasoning about IP addresses are not convergent. Notably, *empirical studies* have demonstrated [67, 75] that users can get assigned over time a set of IP addresses which are unique and stable. According to these finding, if IP addresses are not personal data, then there is a major flaw in the GDPR because it would provide a mechanism to websites to track their users without any control.

[31] `https://www.torproject.org/`

"Personal data" is "*any information relating to an identified or identifiable natural person* (Article 4(11) GDPR). "Identified" means when a person, within a group of persons, is "distinguished" from all other members of the group [33]. "Identifiable" person is one who, although has not been identified yet, is possible to be identified in the future, either directly or indirectly. Such possibility to be identified can be made in two ways:

i) by reference to identifiers: as a name, an identification number, location data, an on-line identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Article 4(11) GDPR [94]); and
ii) by "*all the means likely reasonably to be used either by the controller or by any other person to identify the said person*" (Recital 26 GDPR).

Does this definition of personal data apply to IP addresses? The answer to this question is very important and has been debated for years by both computer scientists and lawyers. We explain the current state of this debate.

Let us consider a case study wherein Alice has subscribed to an Internet Service Provider (ISP) called Bob. *Is Alice's IP address personal data for Bob?* There is a consensus [28, 18] acknowledging that Alice's IP address is personal data for Bob. Indeed, Bob stored the identification data of all his subscribers (including Alice's) and assigned them their IP addresses. Therefore, Bob can identify Alice from her IP address.

Now, Alice visits Eve's website. *Is Alice's IP address personal data for Eve?* The answer to this question divides the data protection scholarship [95, 117]. This community diverges in the understanding of "dynamic" or "temporary" IP addresses as personal data[32] [61, 72, 66, 97, 48, 117, 59, 49].

In 2011, an official report of the Publications Office of the EU [91] studied the case law on the circumstances in which IP addresses are considered personal data; it showed that from 49 decisions regarding the IP address's status in 13 EU Member States, 41 decisions ruled (either explicitly or implicitly) that IP addresses should be considered personal data and 8 ruled against this interpretation. Currently, three stances prevail, as summarized in Table 2.22, which presents a non-exhaustive list of the legal positions from courts and stakeholders (EDPS and 29WP) on the legal status of IP addresses.

1. IP addresses can *per se* identify a person;

[32] Such divergence does not happen in the case of "static" or "fixed" IP, which are 'invariable and allow continuous identification of the device connected to the network' [19] (parag 36).

2. IP addresses do not suffice alone and thus additional information are needed to enable the identification of a person;

3. IP addresses only configure personal data- if additional data is obtained by lawful means.

Decisions	IP address are personal data		
	Alone	With added info.	Added info obtained by legal means
EDPS	•	-	-
Berlin district court	•	-	-
29 WP	•	-	-
CJEU (Breyer)	-	•	•
ECHR	•	-	-
Munich court	-		•
Paris appeal court	-	-	•

Table. 2.16: Summary of the legal positions concerning the status of IP addresses as personal data.

IP addresses *alone*

In a case held at the European Court of Human Rights (ECHR) [31], the decision of the court evidences that it regarded dynamic IP address as information based on which the offender at stake could be identified. In Germany, both the Berlin district court and an appellate court decided that IP addresses are personal data; they added that "determinability" of a person should account for both legal and illegal means to obtain additional data [91].

Article 29 Working Party [1] declared that IP addresses should be treated as personal data by both ISPs and search engines (even if they are not always personal data) and adds that *unless an ISP or a search engine are in a position to distinguish with absolute certainty that the data correspond to users who cannot be identified, they will have to treat all IP information as personal data, to be on the safer side.*

In 2008, the European Union Data Protection Supervisor (EDPS) referred that for IP addresses to count as personal data, there is no requirement that the controller knows the surname, first name, birth date, address (among other) of the individual whose activity it was monitoring. It further stated that an IP address which is showing special behaviour in terms of the transactions one can follow, then in a reasonable world, that is an individual [51].

IP address with additional information

The Court of Justice of the European Union (CJEU) determined in Breyer case [19] that dynamic IP addresses (temporarily assigned to a device), per se, is not information relating to

an "identified" person, due to the fact that "*such an address does not directly reveal the identity of the person who owns the computer from which a website was accessed or that of another person who might use that same computer*". IP addresses can constitute personal data, provided that the relevant person's identity can be deduced from a combination of the IP address and additional data.

This additional information [34, 19] can consist of *e.g.* name, login details, email address, username (if different from the email address), subscription to a newsletter or other account data, in the course of logging in and using the website; cookies containing a unique identifier [32], device fingerprinting or similar unique identifiers. By holding added data, the website can tie it with the visitor's IP address and therefore this visitor would be identifiable[19]. This argument explains why everyone would agree that Alice's IP address is a personal data for Bob (as ISP), because he knows both Alice's name and her IP address. However, if Eve has access to additional information to identify Alice, then Alice's IP address is personal data for Eve.

IP address with lawfully obtained additional information

Pursuant to this view, an IP address will only be personal data when a website has legal means to lawfully obtain access to sufficient additional data held by a third-party in order to identify a person. Respectively, IP addresses will not consist of personal data when such added data is obtained in a way that is prohibited by law, because ISP have to meet its own legal obligations before it just hands over the data. As ISPs are generally *prohibited* from disclosing information about a customer to a third party, the only means wherein an ISP is forced to disclose IP addresses data consist of consent, court order, by law enforcement agencies or national security authorities [34]. The Paris appeal court, in two rulings, stated that the processing of IP addresses does not constitute personal data unless a law enforcement authority obtains a user identity from an ISP [84, 85].

The Munich district court, in 2008, held that dynamic IP address lack the necessary quality of "determinability" to be personal data, which means that it cannot be easily used to determine a person's identity, without a significant effort and by using "normally available knowledge and tools." The court recalls that ISPs are not legally permitted to hand over the information identifying an individual, without a proper legal basis (only when ordered by a court) [76]. The CJEU in Breyer case [19] concluded that a dynamic IP address constitutes personal data if the website operator has "legal means" for obtaining access to additional information held by the ISP that enables the website publisher to identify that visitor and there is another party (such as an ISP or a competent authority) that can link the dynamic IP address to the identity

of an individual. *Legal means* could consist, for example, bringing criminal proceedings in the event of denial-of-service attacks to obtain identifying information from the ISP.

Empirical studies

Against this background, empirical studies have already demonstrated in [67, 75] that an user can get assigned, over time, a set of IP addresses which are unique and stable. Mishra *et al.* [75] found that the retention period of an IP address was, on average, 9.3 days. 2% of user's IP addresses did not change for more than 100 days and 70% of users (amounting to 1569) had at least one IP address constant for more than 2 months. Therefore, it is possible to discriminate Internet users based on their sets of IP addresses. Cycles and patterns of IP addresses were also observed in [75] in a user's browsing history. These cycles have the potential to be abused to infer traits of user behaviour, as well as mobility traces. Accordingly, we observe that if IP addresses are not considered personal data, then there is a *hole in the GDPR*.

To summarize, Court decisions (both at national and CJEU level) and stakeholder positions so far diverge on the legal status of IP addresses. Conversely, empirical studies make evident that even dynamic IP addresses are afforded with uniqueness and stability features and thus could be a relatively reliable and robust way to identify a user visiting a website. This ambiguity triggers uncertainty and confusion to all the organizations handling IP addresses.

2.4.4 How organization handle IP address?

Defining what is or not personal data is important question for anybody who collects and processes the data of a user or customer. Depending on the answer to this question, an organization applies certain regulations in order to process personal data, like the GDPR in the European Union. At a first sight, it seems easy because the GDPR provides a definition of personal data, but in practice, it is very difficult to interpret despite some clarifications made by European data protection authorities (DPAs [33]).

We have already seen that, there exists many ambiguities around IP addresses, it is confusing to understand if they are personal data or not and when they are personal data. Our research focuses on finding answers to some questions like:

[33] https://gdprhub.eu/index.php?title=Category:DPA

(i) How organizations treat IP addresses, do they consider them personal data or not?
(ii) How they respond to IP-based subject access request?

Targeted websites

We have chosen 124 organizations and performed experiments to check how they handle IP addresses. We have visited two groups of websites: 74 websites maintained by private companies and 50 websites maintained by public organizations from which most are national Data Protection Authorities. First, we analyzed the privacy policies of these websites to check if they mention the processing of IP addresses. Then, we have submitted subject access requests based on the IP addresses used to visit the corresponding websites.

- **Private companies websites**: Firstly, we have chosen 22 websites of private companies that are considered the most visited around the world[34] in 2021. Article 29 Working Party (29WP) stated that *devices with a unique identifier (through the cookie) allows the tracking of users of a specific computer even when dynamic IP addresses are used. In other words, such devices enable data subjects to be 'singled out', even if their real names are not known.* Hence, our next choice was a list of 52 websites of companies that set cookies on their user's browser. The computation of these cookies depends on the IP address of the user. Table 2.17 provides the names of all the 74 private companies we have considered in our work.

- **Public organizations websites**: We have chosen 50 websites from Data Protection Authorities (DPAs), the website of the European Data Protection Board (EDPB) and the website of the European Data Protection Supervisor (EDPS). DPAs are independent public organizations that supervise through their investigative and corrective powers, the application of the data protection law in each EU country. They all have a website, therefore it was logical to investigate how do they consider IP addresses. We have considered all the DPAs listed by the GDPR hub[35]. We have also visited the website of the EDPB [36] and the EDPS [37]. Table 2.18 provides a list of 48 DPAs whose websites we visited during our experiments.

[34] https://ahrefs.com/blog/most-visited-websites/
[35] https://gdprhub.eu/index.php?title=Category:DPA
[36] https://edpb.europa.eu/
[37] https://edps.europa.eu/about-edps_en

Popular companies			
Google	Youtube	Amazon	LinkedIn
Zoom	Yahoo	Ebay	Pinterest
Twitch	Roblox	Bitly	Fandom
Microsoft	Netflix	Euronews	Facebook
Reddit	Indeed	Wikipedia	Twitter
Tripadvisor	Apple		
Companies that set cookies using user's IP address			
1000.menu	adbox.lv	addthis.com	admanmedia.com
bigcommerce	britishairways.com	caranddriver.com	dongao.com
dsar.everydayhealth.com	forever21.com	gismeteo.ua	grainger.com
assets.new.siemens.com	kuleuven.be	louisvuitton.com	lifepointspanel.com
nesine.com	mylu.liberty.edu	my-personaltrainer.it	okta.com
pgatour.com	pubmatic.com	point2homes.com	russianfood.com
spiceworks.com	smartadserver.com	sprint.com	start.me
turktelekom.com.tr	urbanfonts.com	vans.com	warriorplus.com
worldpopulationreview.com	yandex.com.tr	yandex.kz	zoho.com
adswizz.com	jpnn.com	constantcontact.com	duda.co
gumgum.com	iheart.com	lyst.co.uk	mckinsey.com
officedepot.com	orange.es	rubiconproject.com	sinoptik.ua
sunstar.com.ph	trafficjunky.net	wikimedia.org	wikiquote.org

Table. 2.17: List of 74 Private companies (22 popular companies and 52 companies that set cookies using IP address).

Lithuania	Spain	Estonia	Germany	Romania	Netherlands
Bavaria, public sector	Croatia	Bavaria	Greece	Belgium	Luxembourg
Mecklenburg-Vorpommern	Portugal	Cyprus	Saarland	France	Thuringia
Baden-Württemberg	Sweden	Denmark	Hungary	Bremen	Berlin
Schleswig-Holstein	Ireland	Austria	Saxony	Slovakia	Iceland
Rhineland-Palatinate	Latvia	Lower Saxony	Poland	Norway	Liechtenstein
North Rhine-Westphalia	Malta	Hesse	Bulgaria	Finland	Hamburg
Saxony-Anhalt	UK	Czech Republic	Italy	Slovenia	Brandenburg

Table. 2.18: List of 50 public organizations (48 DPAs, EDPB and EDPS).

What privacy policies say?

Our first step towards analyzing how organizations handle IP addresses of a user was to check what their privacy policies say about IP addresses. Every organisation located in the European Union or the one that is processing data from European residents must publish a privacy policy on their website [47]. Website users are able to access necessary information in the privacy policies on the purposes of data processing, types of data collected, rights of the data subjects etc.. (Articles 13 to 22 of the GDPR).

Websites collect and process different user information, if they collect IP addresses of a user, they must also inform their users about the collection, purpose and processing. The 29WP [33] stressed that when determining the nature of personal data, it is crucial to evaluate the "*purpose pursued by the data controller in the data processing*". Therefore, IP addresses need to be mentioned in the privacy policy of a website if they are collected/processed. Accordingly, we have visited the privacy policies section of each website considered in our study. Our findings are summarized in Table 2.19 (Table 2.20 and 2.21) provides names of these companies and DPAs based on their category.).

We have identified three different ways to handle IP addresses in the websites privacy policy:

(i) We observed that IP addresses are processed by the visited websites as these are mentioned explicitly in the website's privacy policy. (ii) Privacy policies can also mention that IP addresses are processed but then they are anonymized. The technique used to anonymize them is never mentioned or detailed. Still, it can be acknowledged that a website considers an IP address as personal data because anonymization consists in the process of turning personal data into data that does not relate to an identified or identifiable person any longer (Recital 26 of the GDPR). (iii) Privacy policies also state that IP addresses are not collected. Optimistically, one can acknowledge that websites consider that IP addresses are personal data. They prefer to mention explicitly that they are not processing IP addresses because they are personal data.

The 29WP [33] stated that, it is crucial to evaluate the "*purpose pursued by the data controller in the data processing*". Accordingly, we analysed the purposes for processing IP addresses described in the consulted privacy policies. We noticed that the purpose of the collection and processing of IP addresses include the following: *enhancing the user experience and security*. Enhancing the user experience refers to the identification of the location of the user, personalizing and improvement of products, customization of services and trend analysis or the website administration. Some organizations collect IP addresses for security reasons to protect their business against fraudulent behavior or in case of legal process relating to a

criminal investigation or alleged or suspected illegal activity. In fact, these mentioned purposes require some degree of user personalization. As the 29WP refers [33], *'to argue that individuals are not identifiable, where the purpose of processing is precisely to identify them, would be a sheer contradiction in terms. Therefore, the information should be considered as relating to identifiable individuals and the processing should be subject to data protection rules'*. As such, we reason that all these purposes potentially enable the collection of data that conducts to the identification of a user without unnecessary or disproportional effort.

Description	# Private	# Public	Personal data
Process	60	12	Yes
Anonymize	1	8	Yes
Do not collect	0	3	Yes
Do not mention	10	23	Unknown
No page found	3	4	Unknown

Table. 2.19: Analysis of privacy policies of 124 organizations.

Summary: We were unable to find the privacy policies of the seven websites. 23 (46%) public websites and 10 (13%) private websites do not mention IP addresses in their privacy policies. It is unclear whether these websites do not store IP address or whether they do not take care to mention them in their privacy policies. Using our analysis, we found that most private websites (82%) considers that IP addresses are personal data and only 46% of the public websites share this view. Majority of the private websites mention explicitly the processing of IP addresses. Whereas, the situation is slightly different for public websites: 8 of them anonymize IP addresses and 3 of them explicitly state that they do not collect IP addresses.

Exercising IP-based SAR

We observed that it is possible to conclude that IP addresses are personal data according to the privacy policies of websites (Section 2.4.4). The GDPR provides the right to a data subject to access the personal data that a company collects. But do websites accept to answer a request based on an IP address? A subject access right request (SAR) is a request sent by a data subject to a data controller to exercise his/her right to access their data (Article 15 of the GDPR). Several studies [79, 110, 14, 15, 71, 16, 86, 25] have used SAR as a methodological tool to assess the transparency of certain data processing, the strength of their authentication procedure or their readiness to comply with the GDPR. During our analysis, we have submitted IP-based subject access requests to private and public organizations. In this section we explain our methodology implemented to access the websites of organizations and how we delivered the IP-based SAR request to organizations

Process IP address			
Google	Youtube	Amazon	LinkedIn
Zoom	Yahoo	Indeed	Pinterest
Twitch	Roblox	Bitly	Fandom
Netflix	lifepointspanel.com	admanmedia.com	worldpopulationreview.com
yandex.kz	jpnn.com	turktelekom.com.tr	grainger.com
yandex.com.tr	orange.es	gumgum.com	lyst.co.uk
sinoptik.ua	zoho.com	constantcontact.com	sprint.com
addthis.com	pubmatic.com	wikiquote.org	iheart.com
start.me	my-personaltrainer.it	nesine.com	pgatour.com
spiceworks.com	forever21.com	adswizz.com	wikimedia.org
Reddit	Apple	Wikipedia	Twitter
Microsoft	Euronews	caranddriver.com	officedepot.com
vans.com	britishairways.com	duda.co	trafficjunky.net
okta.com	point2homes.com	mckinsey.com	rubiconproject.com
smartadserver.com	dsar.everydayhealth.com	bigcommerce	
Anonymize IP address before processing			
gismeteo.ua			
Do not mention IP address in the privacy policy			
louisvuitton.com	mylu.liberty.edu	sunstar.com.ph	kuleuven.be
urbanfonts.com	1000.menu	Ebay	Tripadvisor
warriorplus.com	assets.new.siemens.com		
Privacy policy page not found			
adbox.lv	russianfood.com	dongao.com	

Table. 2.20: 74 Private companies categorized based on their privacy policy.

Process IP address				
Slovakia	Lower Saxony	Schleswig-Holstein	Iceland	Brandenburg
Rhineland-Palatinate	Saxony-Anhalt	Bremen	Poland	Hesse
North Rhine-Westphalia	Slovenia			
Anonymize IP address before processing				
Germany	Liechtenstein	Bavaria	Lower Saxony	Bremen(LFDI)
LFD (Saxony-Anhalt)	EDPB	EDPS		
Do not collect/process IP address				
Schleswig-Holstein	Saarland	Brandenburg		
Do not mention IP address in the privacy policy				
Estonia	Belgium	Croatia	Cyprus	France
Sweden	Ireland	Austria	Saxony	Latvia
Czech Republic	Greece	Hamburg	UK	Malta
Baden-Württemberg	Hungary	Finland	Thuringia	Norway
Mecklenburg-Vorpommern	Denmark	Luxembourg		
Privacy policy page not found				
Lithuania	Spain	Romania	Netherlands	

Table. 2.21: 50 Public Organizations (48 DPAs, EDPB and EDPS) categorized based on their privacy policy.

Experimental setup and exercising SAR: We have visited the websites using two methods. We first visited them directly using the IP address provided by our ISP to our device. Then, we visited websites through Tor Network[38]. For the websites, using Tor means that we are using the IP address of the TOR exit node. It is likely that many devices and users use the same exit node and therefore the same IP address. Even though we have used two different networks to access the websites, due to the fact that our request was always denied, we did not mention the use of different IP addresses in our further discussions.

All the visited websites could be viewed as an external user, but some of them could be also accessed as a registered user. We have always visited the websites as an *external user*.

SAR template: We have devised a generic subject access request to all the companies and public websites. The full text of the subject access request letter appears in Listing 2.1. We used this letter to send a SAR to all the organizations. We have complied with all the requests for additional information (like the copy of an ID) to authenticate the SAR made by these websites. Each recipient of our SAR was then permitted one month time to respond to our request as mandated by the GDPR.

[38] `https://www.torproject.org/`

```
Dear Data Controller,

I am hereby requesting a copy of all my personal data held and/or undergoing
    processing, according to Article 15 of the GDPR. Please confirm whether or not
    you are processing  personal data concerning me. In case you are, I am hereby
    requesting access to the following information:  All personal data concerning
    me that you have stored. This includes any data derived about me, such as
    opinions, inferences, settings and preferences.

Please make the personal data concerning me, which I have provided to you,
    available to me in a structured, commonly used and machine-readable format,
    accompanied with an intelligible description of all variables.

I am including the following information necessary to identify me:
Name - (first name last name)
IP addresses used to access are as follows-
XX.XX.XX.XXX
XXXX:XXXX:XXXX:XXXX::XXX:XX

Yours sincerely,
first name

(As laid down in Article 12(3) GDPR, you have to provide the requested information
     to me without undue delay and in any event within one month of receipt of the
    request. According to Article 15(3) GDPR, you have to answer this request
    without cost to me.)
```

Listing 2.1: Generic SAR template used for IP-based SAR. request

Challenges: We faced several challenges while exercising the SAR for both private companies and public organizations. We visited 74 websites of *private companies* and we submitted SAR to only 62 of them. Four websites (addthis.com, dongao.com, nesine.com and sunstar.com.ph) published wrong e-mail addresses and 6 websites (Google, Youtube, Amazon, LinkedIn, grainger.com and constantcontact.com) did not provide a contact e-mail to submit a SAR if not logged into a user account. Two websites (sprint.com and iheart.com) allow a SAR only for USA citizens. For the *websites of public organizations*, we have visited 50 websites and submitted SAR to only 47 of them. Websites of the Irish and Dutch DPAs did not provide an e-mail address to submit a SAR, while the Italian DPA has provided a wrong e-mail address.

The responses received for our SAR were grouped into 5 different categories. Table 2.22 shows the categories and the number of organizations related to each of these categories. From 109 organizations to which we submitted a SAR, only 62 thereof have responded (36 private

companies and 26 public organizations). Figure 2.5 provides the list of private companies and public organizations belonging to each category. In the following sections we depict the obtained responses per category along with it's legal analysis.

Answer's category	Private	Public
No reply	26	21
No, we have nothing about you	7	23
No account was found	20	0
No, it does not allow to identify you	4	0
No, others	5	3

Table. 2.22: # of websites categorized based on their responses obtained for our IP-based SAR.

No reply from the website: 47 websites (26 private and 21 public) did not reply to the SAR. It is unknown why companies and DPAs did not answer to a SAR request. One could be tempted to conclude that they did not have a process in place to respond to subject access requests. This is particularly surprising (if not shocking) for DPAs. However, the GDPR mandates that the data controllers have an explicit *obligation to facilitate the exercise of data subject rights* (Articles 12(1) and 28(3)(e)), including facilitating SAR requests. Recital 7 recalls that each person should have control of their own personal data and a no-reply to a SAR consists in an obstruction to such control. Recital 59 thereto emphasises that *"modalities should be provided for facilitating the exercise of the data subject's rights"*.

No, we have nothing about you: 30 websites (7 private and 23 public) answered they do not have any data matching our request. We compared these answers with the privacy policies of these websites (see Table 2.23). Some websites (6 private companies and 4 DPAs) are not consistent. Their privacy policies mention the processing of IP addresses, though they fail to find any data relating to the IP addresses included in the SAR request.

Description	# Private	# Public	Consistency
Process	6	4	No
Anonymize	0	5	Yes
Do not mention	1	12	Yes
Do not collect	0	2	Yes

Table. 2.23: Consistency between the websites privacy policies and their response to our SAR.

The answers from the remaining websites are consistent with their privacy policies because they either do not mention IP addresses, do not collect IP addresses or anonymize them. In this latter case, a website is indeed unable to recover data corresponding to our IP addresses.

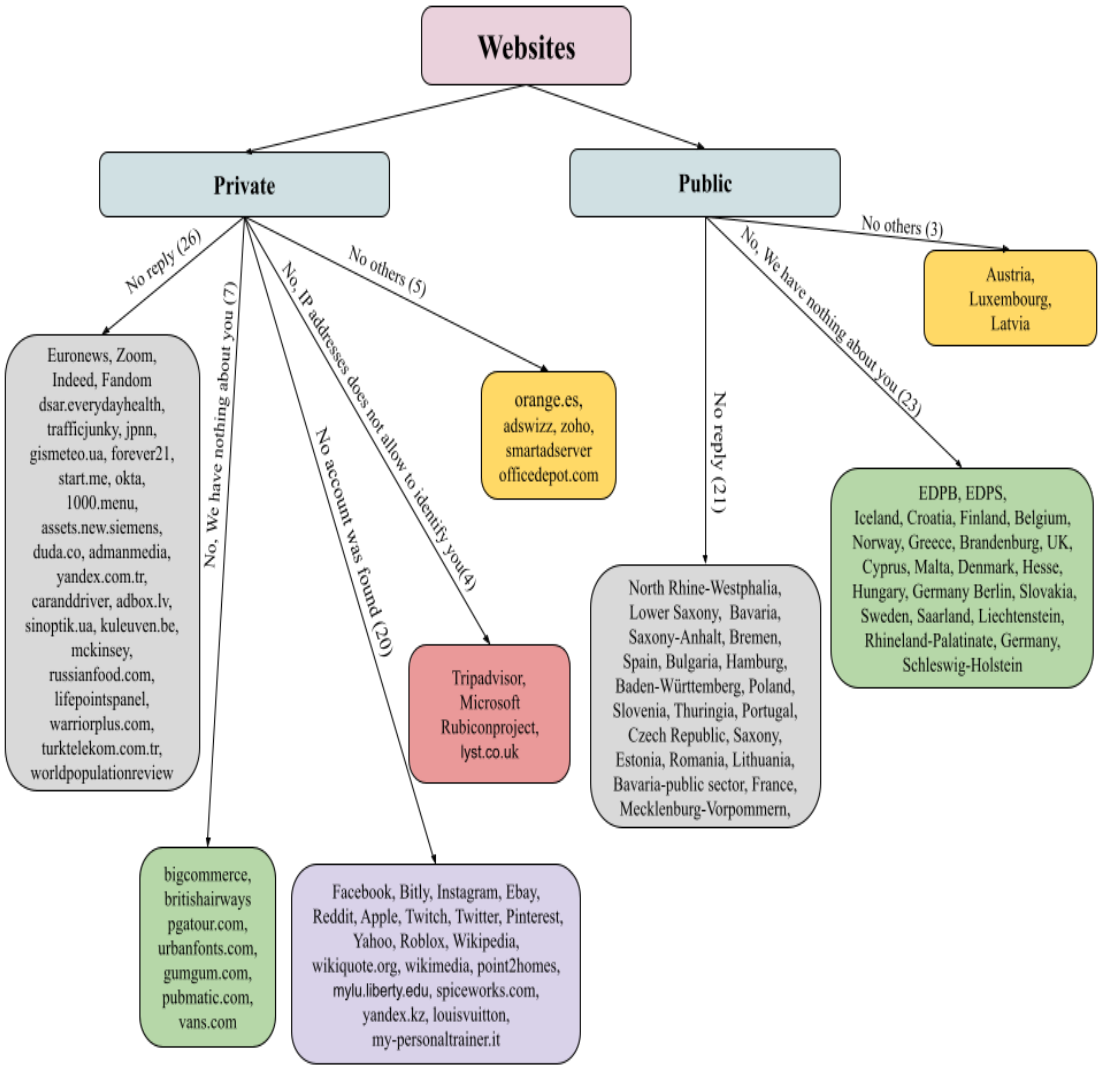

Figure. 2.5: Responses obtained by websites of private and public organizations for a IP-based SAR request.

No account was found: 20 websites explained that they cannot process the SAR request because they did not find any account corresponding to our request in their system. We expected that websites would use our IP addresses to query their information system and then extract the corresponding information associated to these given IP addresses. But it appears that many of them have queried their information system using the email address used to submit the subject access request. As they have found nothing corresponding to this email address, they reply that they have no data associated to this account or that we need to login or provide our login information to complete our request. For instance, *Roblox* replied to us that: *You must verify ownership of any associated Roblox account. We must have a Roblox user name in order to proceed with a GDPR request.*

This kind of answer shows how websites process subject access requests. They have completely ignored the IP addresses provided in our requests to focus only on the email address used to send the request. Thus, their procedure was not able to treat our request.

It looks like, their implementation is (maybe) motivated by the need to authenticate the request: *The controller should use all reasonable measures to verify the identity of a data subject who requests access, in particular in the context of online services and online identifiers* (Recital 64 of the GDPR). They have email addresses in their information systems and they want to ensure that they provide data only to a legitimate user.

Internet users need to provide many personal data, like their last name, first name, gender, location, phone number, email addresses etc.. to create an account on a website. Even if it is possible to lie for some of these fields, it is tempting to visit websites without creating an account: it is faster and also a privacy-preserving choice. However, it implies that the websites can do whatever they want with the user's data as there is no way for a user to exercise the rights provided by the GDPR. This kind of answer is therefore problematic because it creates a legal loophole.

No, it does not allow to identify you: Four websites (Tripdvisor, Microsoft, lyst.co.uk and rubiconproject.com) have taken into account the IP addresses provided in our request. However, according to their answers, they cannot process the request, since IP addresses are shared and dynamic. For instance, *rubiconproject.com* states: *we do process data associated with IP addresses XX.XX.XX.XX and XX.XX.XXX.XXX, our searches suggest that these addresses are associated with multiple different devices across multiple territories. This indicates that these IP addresses are used by multiple different users etc. . .*

These websites have considered that an IP address can be used by multiple users at the same time. In other words, they are unable to distinguish, in their own information system, the data belonging to us from the data of other users using the same IP addresses.

In such situation, if a website is able to demonstrate that it is not in a position to identify a concrete user, it can deny the request (Article 11(2)).

Additionally, there is a *risk* for the websites to disclose the information of other users and whenever such a personal data breach risk exists (Article 4 (12)), a website can deny the request.

Recital 63 of the GDPR: *"That right should not adversely affect the rights or freedoms of others"* One may argue that these websites could have collected or requested *more information* from the user in order to identify the requester. In effect, Recital 64 of the GDPR states that the controller should use all reasonable measures to verify the identity of a data subject who requests access, in particular in the context of online services and online identifiers. This vague concept of "reasonable measures" might result in data controllers implementing weak or irrelevant identity verification means upon receiving a request.

Furthermore, a website is not obliged to collect additional information to identify the data subject for the sole purpose of complying with the GDPR subject access rights (Article 11 (2) and Recital 57 of the GDPR). This argument excludes this possibility of acquiring additional data.

No, Others: 8 websites (5 private and 3 public) did not provide any useful information in their response to the SAR. Generally, these websites requested additional data to identify the requester. This implies that IP addresses alone were not a sufficient piece of data to identify the requester, which means that IP addresses alone are not considered as personal data.

Such request for added data is in line with the GDPR wherein a controller – having reasonable doubts concerning the identity of a person making the request – may request the provision of additional information necessary to confirm the identity of the data subject (Article 12(6)).

However, there seems to be no substantive reason for a data subject to reveal the real identity to these websites through the requested documents: original signed letter (Luxembourgian DPA), SAR in the official language of the DPA (Austrian DPA) or a INE number, *i.e.*, a statistical number (orange.es). These documents requested by these DPAs must be *proportional* or *necessary* to the website's knowledge of the data subject. Moreover, the *minimization principle* mandates that personal data shall be adequate, relevant and limited to what is necessary in

relation to the purposes for which they are processed (Article 5 (1) (c)). Recital 39 specifies further that *personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means.* The "necessity" or "proportionality" requirement that both these provisions note, refers to both quantity and also to the quality of personal data. It is then clear that these DPAs should not process excessive data if this entails a disproportionate interference in the data subject's rights and hence, a privacy invasion.

Summary: In Table 2.24, we can contrast the consistency between what companies and organizations publish on their privacy policies and the way they respond to the IP-based SAR. There is only one private company and just 17 public organizations that are consistent. After analyzing all the responses received by websites, we can see that IP-based SAR requests are not handled properly.

		# of websites (private)	No reply	No, We have nothing about you	No account was found	No, it does not allow to identify you	No, others	# of websites (public)	No reply	No, We have nothing about you	No account was found	No, it does not allow to identify you	No, others
Process	60	19	6	17	3	5		12	7	6	0	0	0
Anonymize	1	1	0	0	0	0		8	3	4	0	0	0
Do not collect	0	0	0	0	0	0		3	0	3	0	0	0
Do not mention	10	4	1	3	1	0		23	8	10	0	0	3
No page found	3	2	0	0	0	0		4	3	0	0	0	0
Total	74	26	7	20	4	5		50	21	23	0	0	3
		Private companies						**Public Organizations**					

Table. 2.24: Consistency between the websites privacy policies and their response to our SAR. (SAR sent to 62 private companies and 47 public organizations)

Responses obtained on IP based SAR: All the reponses received for IP based SAR are briefly described in Table 2.25, 2.26 and 2.27.

2.4.5 Extension to IP-based SAR studies

All the visited websites of 124 organizations could be viewed as an external user, but some of them could be also accessed as a registered user. To observe if it would have changed the response obtained to our subject access requests, we have created an account on 20 private popularly used websites[39] (Google, Youtube, Amazon, LinkedIn, Reddit, Zoom, Yahoo, Ebay, Pinterest, Wikipedia, Twitter, Twitch, Roblox, Bitly, Fandom, Tripadvisor, Microsoft, Apple, Facebook and Indeed) and exercised IP-based SAR.

What privacy policies say?
As already detailed in Table 2.20, 18 of these companies process IP address (Google, Youtube, Amazon, LinkedIn, Reddit, Zoom, Yahoo, Pinterest, Wikipedia, Twitter, Twitch, Roblox, Bitly, Fandom, Microsoft, Apple, Facebook and Indeed) and 2 companies (Tripadvisor and Ebay) did not mention IP address in their privacy policy.

Exercising IP-based SAR as a registered user

We first visited all the 20 websites directly using the IP address provided by our ISP to our device. Then, we visited websites through Tor Network[40]. We observed that, companies impose certain verification techniques when their users location or devices change. There are different kinds of verification (sending a code to a mobile or e-mail or no robot test) imposed by different companies. During our analysis, accessing the websites as a registered user lead to verification when using Tor network. Table 2.28 shows the kinds of verification followed by companies. Some of the websites did not allow the access the registered account even after the verification and hence we could not access Zoom, Fandom, Pinterest, Tripadvisor, Facebook using Tor network. Finally, we used the IP addresses of networks that let us access the websites during further processes.

IP-based SAR response

Table 2.29 shows the number of websites grouped into different categories and Figure 2.6 shows the name of these companies.

[39] https://ahrefs.com/blog/most-visited-websites/
[40] https://www.torproject.org/

Websites	Response
Google Youtube Amazon LinkedIn	No contact details found on the website to make a SAR without an user account.
euronews zoom indeed fandom	No reply
Facebook Instagram Wikipedia	No user account was found with this name
Ucnews	Wrong contact details (email) provided on their website. Hence couldn't contact.
Ebay	Requested a copy of passport, driving license or any other government document with photograph to associate to a account.
Reddit	Without verification that you are the owner of the account in question, we will not be able to proceed.
Apple	Apple gives you the ability to request a copy of data associated with your Apple ID.
Twitch	Please provide current IP from whatsmyip.org and DOB to process request. After providing the details they responded saying they can't find an account.
Twitter	Unfortunately, we aren't able to help you with this issue
Pinterest	No email - I'm sorry that I would not able to offer more help. Unfortunately, there's nothing else I can do on my end.
Yahoo	Unfortunately, we can only provide information associated with a registered account that we have verified account ownership for.
Roblox	You must verify ownership of any associated Roblox account. We must have a Roblox user name in order to proceed with a GDPR request.
bitly	If you don't have a account, please note that we don't maintain sufficient information to verify your identity, we have no reasonable way to respond.
Trip advisor	Unfortunately, we are unable to process your request based on the IP address you have provided. This is on the basis of a security and validation concern. Due to internal protocols we are unable to search, disclose such information in our systems as we are not able to definitively match an IP address directly to an individual user. In other words, IP addresses can be directly or indirectly linked to more than one individual and having regard to that and a careful assessment of the risks of an unauthorized disclosure, we are regrettably unable to proceed. Reasons for this can be, due to members of the same household – each with distinct data protection and privacy rights– using devices with the same IP address. Equally,visitors of our site from a public establishment may also have the same IP address.
Office	Because IP addresses can be dynamic or for shared devices, we do not provide data based on just an IP address.

Table. 2.25: Response from popular companies.

Company name	Response
grainger.com	Not a customer, no way to make a SAR request.
constantcontact.com	No user account, no SAR
louisvuitton.com	We are not able to use your IP addresses, as it is not a search criteria in our systems.
rubiconproject.com	We do process data associated with IP addresses 198.16.70.27 and 89.81.166.110, our searches suggest that these addresses are associated with multiple different devices across multiple territories. This indicates that these IP addresses are used by multiple different users.
officedepot.com	In order to avoid collecting any personal information of EU residents, our website is not available in the EU and we do not ship products to customers in the EU. We are not a data controller as that term is defined in the GDPR.
vans.com	Based on our preliminary investigation, we don't hold any personal information about you, nor the email with which you're writing
pubmatic.com	We can confirm we do not currently hold any data associated with the below IP addresses
orange.es	Requested a ID and after submission did not reply to SAR.
gumgum.com	The only data we collect/process is IP address information. However, that number is anonymized by removing and replacing random numbers which makes it impossible for us to track anything about you.
wikiquote.org wikimedia.org	We do not have any record of an account associated with this email address. (No account no data)
point2homes.com iheart.com	No account found California residents: Log in to your iHeartRadio account to make a CCPA request
lyst.co.uk	IP addresses are shared and it is entirely possible that someone else has used the same IP address as you. We are therefore unfortunately unable to verify your identity and cannot fulfill DSAR request based on an IP address alone.
smartadserver.com-	we do not collect information that could directly identify you, such as e-mail addresses.
urbanfonts.com	I confirm that we do not have any personal data stored concerning you.
pgatour.com	Your request has been processed but there were no records found in our systems based on the information you provided.
dongao.com	E-mail sent to this company was automatically rejected (Reason - Quota exceeded (mailbox for user is full)).
zoho.com	The only data we have about you is this mail communication in our ticket system.
adswizz.com	We do not have the possibility to associate an IP address to the identity of a person because we do not process information like name and surnames or e-mail accounts in our ecosystem. We does not collect personal data about you - other than data needed for the purpose of establishing an internet connection with our website. We may also use cookies to provide you with a better experience when visiting our website or for analytical purposes based on your acceptance.
britishairways.com	undergoing processing, according to Article 15 of the GDPR. With the information you have provided, we have been unable to locate any such data for you.
mylu.liberty.edu	Go to your account and see the dashed board.
spiceworks.com	This email address is not associated with a Spiceworks user account.
bigcommerce	we have performed a review of our ecosystem using your provided personal data. No records were detected/stored.
yandex.kz	Please send me your yandex login
my-personaltrainer.it	We have made some checks, we can't find the references of the email address.

Table. 2.26: Response from companies using IP address to set cookies.

DPA	Response
DSB (Austria)	The input must - in order to be submitted to a formal treatment be entered in the official language of the Austrian Data Protection Authority, i.e. German. You are requested to submit your observations within a period of two weeks and to prove your identity in a suitable form.
CNPD (Luxembourg)	After submitting a ID- We would like to inform you that this document alone is at this stage not sufficient to confirm your identity pursuant to Article 12(6) of the GDPR. We would therefore like to ask you to submit your request by post, with an original signed letter (not a copy) sent to the following address:
Persónuvernd (Iceland)	No personal data about you is being processed or has been processed by the Authority.
HBDI (Hesse)	No other data except your e-mail.
AZOP (Croatia)	By inspecting our databases, we did not find any referred to you data.
Tietosuojavaltuutetun (Finland)	Thank you for your request. We do not have any personal data about you.
APD/GBA (Belgium)	The Belgian Data Protection Authority does not possess any personal data.
LDA (Brandenburg)	We do not have any personal data about you, except for the data you have given us in request.
Commissioner (Cyprus)	We do not have any record with any of your personal data as described. We keep an electronic record of all complaints/ questions/ legal opinions etc
HDPA (Greece)	The audit carried out in the records we maintain (based on the identification details (e-mail address) that you have provided to us) reveals that the Hellenic DPA does not keep any personal data concerning you.
Datatilsynet (Denmark)	The Danish DPA does not process any other data than the personal data you have submitted.
ULD (Schleswig-Holstein)	We don not process any data about you and we don't store IP addresses at all.
Datenschutzstelle (Liechtenstein)	We don not process any data about you and we don't store IP addresses at all.
UOOU (Slovakia)	We confirm that our Authority does not process any personal data about you.
NAIH (Hungary)	The authority has not processed any data about you.
Datatilsynet (Norway)	We have stored no data that can be linked to the personal information that you have provided.
HBDI (Hesse)	No other data except your e-mail.
EDPB	Asked a national identity certificate to process the request and asked 2 more months time to process the request. Final reply- Thank you for your request. We do not have any personal data about you.
BlnBDI (Berlin)	We do not process any personal data concerning you.
Datainspektionen (Sweden)	conducted a search for your personal data in our systems. And the data regarding you is your request.
Datenschutzzentrum (Saarland)	From the data submitted by you via the aforementioned email, we have no further data. For more information you can inform us of reference in kind, especially in the form of a file number or similar.
DVI (Latvia)	Request was not made with enough documents and signature...
IDPC (Malta)	Checked records and the Office has never processed any personal data concerning your kind self. This Office is going to delete the here below email and this very answer too
LFDI (Rhineland-Palatinate)	Does not process any data that is personally identifiable or related to you.
EDPS	Thank you for your request. We do not have any personal data about you.
BfDI (Germany)	No personal data related to you could be identified in BfDI systems and files.
ICO (UK)	We have been unable to locate any data for you.

Table. 2.27: Response from DPAs.

Websites	Access using Tor network
Facebook	No access
Tripadvisor	No access
zoom	No access
pintrest	No access
fandom	No access
Google	Not a robot test
Ebay	Not a robot test
Yahoo	Not a robot test
Roblox	Not a robot test
Amazon	Verification code to mobile
Instagram	Verification code to email
LinkedIn	Verification code to email
Microsoft	Verification code to email
twitch	Verification code to email
bitly	Verification code to email
twitter	Verification code to email

Table. 2.28: Different types of verification performed while using Tor network as a registered user.

Category	Registered
No Reply	1
No, We have nothing about you	8
No account was found	0
No, it does not allow to identify you	4
No, Others	2
Provided data on IP	5

Table. 2.29: # of websites categorized based on the responses obtained for our IP-based SAR as a registered user.

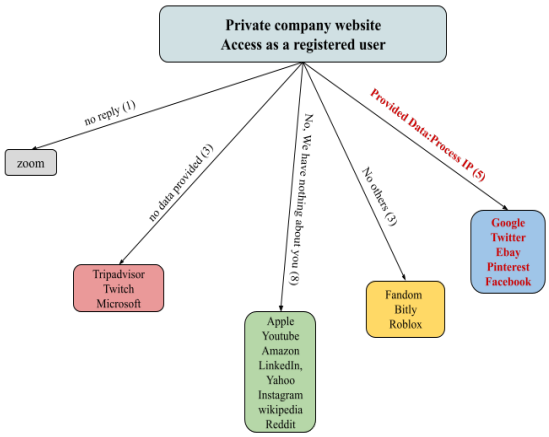

Figure. 2.6: Responses obtained by websites of private and public organizations for a IP-based SAR request.

No Reply: Zoom did not reply to our SAR request.

No, We have nothing about you: eight websites (Apple, Youtube, Amazon, LinkedIn, Yahoo, Instagram, wikipedia and Reddit) provided data either through dashboard or download link. This data did notinclude any IP address. For instance: *Apple data includes- Apple ID Number, Apple ID, Create Date, First Name, Last Name, Language etc.*

No, it does not allow to identify you: Four websites (Tripadvisor, Twitch and Microsoft) replied that they will not be providing the data for the given IP addresses as they are shared, dynamic etc. Response from Fandom states *We have no additional information to provide for IPs and we would not be able to prove that you specifically were the person who used them.*

No, others: 2 companies Bitly and Roblox replied to our SAR but did not provide any useful information, Response from Roblox states- *Sorry we can't identify your account so no data will be provided*.

Provided data on IP: Data downloaded from dashboards of 5 companies Google, Twitter, Ebay, Pinterest and Facebook include IP addresses (only IPv4 addresses) along with some other details like date/time of access, country, browser details, account creation IP address

Category	Data
Pinterest	Username, Email, Gender, Country, Birthday, Last login, Is active, Referrer Signup ip, Known login locations, Login history(Date, Entrypoint, IP), User sessions (IP Address, Platform, Accessed at, Created at) etc.
Ebay	Browsing and search history (Date, Device type, Search query, Item ID, Referrer, Session start date, Page name, IP) Date created- Site name, Channel, IP, signin using web browser
Google	Google Account ID, Alternate e-Mails, Created on, Terms of Service IP, Terms of Service Language, ACCOUNT RECOVERY (Contact e-Mail, Recovery e-Mail, Recovery SMS) IP ACTIVITY (Timestamp, IP Address, Activity Type, Raw User Agents)
Facebook	username, logins and logouts- IP address, data, time
Twitter	account (email, createdVia, username, accountId, createdAt), accountCreationIp, accountId, userCreationIp, set of loginIps, device token

Table. 2.30: As a registered user: Data downloaded from dashboards of companies include IP addresses and other information.

etc. Table 5 shows the data stored by these companies. Foe instance, **Pinterest** data includes- IP addresses along with some other details like date/time of access, country, browser details etc.

Responses on IP based SAR: Table 2.31 provides all the responses obtained for the IP based SAR with a registered user account.

Consistency with privacy policies and response to IP-based SAR

At least 18 websites out of 20 state the use of IP addresses for at least one of the purposes like identify location, personalize and improve products, customize services, troubleshooting and admin reporting the device a user is using to access the Internet based on IP address. In fact, these purposes require some degree of user personalization. As the 29WP refers, *'to argue that individuals are not identifiable, where the purpose of processing is precisely to identify them, would be a sheer contradiction in terms'*. As such, we reason that all these purposes, along with the registered user account information, potentially enable the collection of data that conducts to the identification of a user.

Table 2.32 links the companies that declare the use of IP addresses in their privacy policies and their answers to our SAR. From these findings, we observe that at least 10 companies did not provide any data, even if they have access to registered account information and the e-mail used for the SAR request. Tripadvisor and Ebay did not mention any collection or use of IP addresses in their privacy policies. However, in their SAR replies, we verified that both of these companies process IP addresses. Tripadvisor in the SAR reply, mentioned that

Websites	Response
zoom	No reply
Apple Youtube Amazon LinkedIn Instagram Reddit	Data downloaded from the dash board (no IP found)
twitter	Dashboard (IP address at account creation. Other IP addresses associated with each login, account ID, login time)
Google	Dash board (Timestamp, IP Address, Activity Type (login, logout) and Raw User Agents (Linux, Mozilla))
Facebook	Dash board (IP with login date and time for each access)
Ebay	Dash board (IP, login time and date for each access)
indeed	Link to authenticate the SAR did not work
Yahoo	National ID asked and then refusal to process request.
Roblox	Sorry we can't identify your account so no data will be provided.
Fandom	We have no additional information to provide for IPs and we would not be able to prove that you specifically were the person who used them.
Bitly	We'd be happy to help walk you through how to delete your account
wikipedia	responded with a link and no IP address stored or any other personal information except the username.
pinterest	Download file includes search engine used, browser details, Platform, Signup IP, all the known login locations, date-time with IP
office	Please note that because IP addresses can be dynamic or for shared devices, we do not provide data based on just an IP address.
Trip advisor	Please note that because IP addresses can be dynamic or for shared devices, we do not provide data based on just an IP address.
twitch	We received your request under Article 15 of GDPR for a copy of your personal data. Our general practice is to obtain your Twitch Username associated with your account, along with information that verifies that you are the owner of that account (such as your email address, IP address and recent purchase history). Your request, however, only includes what appear to be 2 IP Addresses that you claim you used to access Twitch.tv. To verify that you are the individual associated with these IP Addresses, we require additional information that can be associated with these IP Addresses (we note that you provided your DOB, however, that does not help us associate you with the IP Addresses). Twitch takes measures to verify that the individual making a request under Article 15 is the appropriate individual in order to ensure that personal information belonging to one individual is not inadvertently disclosed to others. This is especially true with IP Addresses that can be used by multiple individuals.

Table. 2.31: Response from 20 popular websites as a registered user.

Websites	Process IP address	SAR response - IP address Registered user
Reddit	yes	didn't provide data
Fandom	yes	didn't provide data
Apple	yes	didn't provide data
Wikipedia	yes	didn't provide data
Microsoft	yes	didn't provide data
Yahoo	yes	didn't provide data
Amazon	yes	didn't provide data
Youtube	yes	didn't provide data
Instagram	yes	didn't provide data
LinkedIn	yes	didn't provide data
Tripadvisor	No	didn't provide data (but agreed to processing of IP)
Twitter	yes	yes
Facebook	yes	yes
pinterest	yes	yes
Google	yes	yes
Ebay	no	yes

Table. 2.32: Consistency with privacy policy and response to IP-based SAR request.

"*IP address are dynamic and shared and hence they do not provide data*". The download from the dashboard of Ebay, includes IP addresses. From the responses and data downloaded, *we deduce that they indeed process IP addresses*. Websites like Twitter, Facebook, Pinterest and Google mention processing IP addresses in their privacy policies and the data we downloaded from the dashboards include IP addresses. This means that for these four companies, account information in combination with IP address is enough to identify the user.

Summary: We have seen that even as a registered user, private companies deny our IP-based request. In this case, companies have the details of a registered user and all the necessary information to authenticate the data subject and they still choose to deny.

2.5 Conclusion

In this chapter we presented our analysis on what is personal data according to GDPR. Mainly, we demonstrated two identifiers `Name` and `IP address` with various experiments to show when they are considered as personal data.

The analysis on the naming system showed that the difficulties in the application of GDPR varies across countries. Our tool on the `uniqueness of names in France` is a step towards educating online users on how sensitive there name is (currently our tool works only on French names). Analysis on usernames showed that online users tend to share a lot of personal information in their usernames. Websites that provide options to create a personalized account, do not warn or educate users to not share their personal information in their usernames. It is easy for an attacker to gain information on online users using just their usernames. Our experiments have shown that there is a strong need to spread awareness and warn users about information leakage and possible privacy risks associated to it.

Our work on IP address shows that the legal ambiguity surrounding the nature of an IP address may be misused by companies. This hypothesis is tested throughout a case study wherein subject access requests containing IP addresses were sent to companies and Data Protection Authorities. We found out that many of these organizations do not respond properly to data access requests with IP address even if they specify the use of IP address as a personal data in their privacy policies. Our survey of 109 websites show that websites consider IP addresses as personal data in their privacy policies. However, it is not possible to access any data related to the IP addresses used while visiting their websites. *IP addresses as personal data* is only a theoretical statement for Internet users because in practice there is no practical means for them to exercise their rights. Internet users cannot prove that they have used an IP address. Therefore, it is easy for websites to deny IP-based subject access requests.

Denial of IP-based subject access requests is not going to change unless drastic modifications are made to change the way IP addresses are allocated to users. The current scheme is device-oriented and it simplifies the task for ISPs. GDPR-friendly IP address allocation needs to be user-oriented. Changing the situation is complicated because it requires to change a core Internet mechanism: IP addresses allocation. However, it is interesting to see that an IPv6 address allocation scheme [78, 77] was GDPR-friendly and considered for standardisation [12]. This modification might be possible.

2.5.1 Future work and open questions

The proposals made in this chapter on personal data have scope for improvements and some extensions. Here we mention a few of them that we plan to do in the near future.

Usernames

On the subject of username, we developed a questionnaire as shown in Listing 2.2, that will be published on INRIA Mooc in 2021. This questionnaire has two objectives, firstly it is going to determine if the online users are aware of the information leaked in their e-mail and usernames. Secondly, it will also spread awareness about different information users are not supposed to share while creating online identities.

```
Q1: Have you used any of your personal information in your username or E
    -mail address?
     (Yes/No)

Q2: Have you used any of the following details in your username or Email
     Address?
     First Name(Yes/No)
     Last Name(Yes/No)
     Initials of your name(Yes/No)
     Country/City name or Department code(Yes/No)
     Date of Birth(Yes/No)

Q3: Do you think using personal details in Email address or Username
    could lead to PRIVACY ISSUES or RISKS?
     (Yes/No/Maybe/Don't know)

Q4: Do you think using first name/last name in your username or Email
    Address could reveal the following details?
     Gender(Yes/No)
     Location details -(Country/City you belong to or living)(Yes/No)
     Nationality(Yes/No)

Q5: Would you be interested to know the PRIVACY ISSUES or RISKS that are
     associated with your username or E-mail Address?
     (Yes/No/Maybe/Don't know)
```

Listing 2.2: Questionnaire on the username.

Websites revealing the information on the registered customers: While creating an username to have a personalized account in some online services, often websites show the availability of that username. Websites display a message if a particular username is already taken by one of there customer. This is also true when an user uses his/her email addresses to register to an service. This kind of information can reveal private information on the users,

especially if the websites are related to the subjects such as finance, medical, websites that could reveal sexual orientation or preferences etc.. In the near future, we plan to examine different websites that have high privacy impacts when there customer information are revealed and other information associated to this leakage.

IP addresses

On the subject of IP addresses, we have identified five websites (Tripdvisor, Microsoft, lyst.co.uk, twitch and rubiconproject.com) which consider that IP addresses do not allow to identify the requester. We acknowledge the fact that our subject requests were not providing enough information to let websites provide us any data. The position of these four companies is legitimate and it aligns with the position of the UK DPA (ICO) in [90]: *Where a reliable link between the subject access applicant and the information held cannot be established and where, therefore, there is an obvious privacy risk to third parties, the ICO would not necessarily seek to enforce the right of subject access unless there is a genuine risk to an individual's privacy if he fails to do so.*

We believe that it was actually not possible for us to create a valid IP-based subject access request. It was not possible for us to prove that we have used a certain IP address at a given time. It is currently impossible to link an IP address to the identity of a data subject. It is necessary to change how an IP address is allocated to a data subject in order to be compliant with the GDPR.

Still, we believe that this situation can change: it is possible to create IP-based subject access request which are answered positively by websites if we change how IP addresses are allocated to devices. It requires to include in the request, a proof that a given IP address was indeed used for a certain period of time by the individual requesting for data. We believe that this proof could have convinced these four websites.

There are two strategies to obtain a *proof of usage* for an IP address. The user can ask his/her ISP to provide a certificate when an IP address is allocated to his/her device. This certificate needs to include the user's identity, timestamps and anybody including the websites must be able to verify it. The certificate would attest that a given individual has used a certain IP address for a given period. Such a certification scheme would need to rely on public key infrastructure to attest that it was created by an authorized ISP. We have attempted to ask an ISP to provide us such a certificate during our study. Unfortunately, our request was never answered. This solution would increase the complexity of IP address allocation protocol but it does not modify significantly Internet protocols. The second option is to let the users create

their own IP addresses. The computation of an IP address can be based on the user's public key for instance. Such a stateless scheme has been proposed in the past for IPV6 addresses [78, 77]. These schemes[78, 77] are security oriented but it appears that they are also GDPR compliant. They can be used to make IP-based subject access request possible. The IP address and all the elements used to create it constitutes the proof of usage. However, this would be a major modification of Internet Protocol.

Right to be forgotten

On the practices on how companies and organization fulfill a user request w.r.t their personal data, we intend to examine one of the data subject's right: **Right to be forgotten**. The right to be forgotten is also known as the right to erasure. This right allows individuals to ask for their personal data to be deleted from the organization. In near future we will access this right and determine if companies and organizations actually erase all the content on the requested data subject or retain some information. This is an interesting topic to explore to check if GDPR rights are properly implemented or not. We can also check different kinds of information retained by organizations and there purposes.

3 PDF: Portable Document Format

Abstract: Organizations publish and share more and more electronic documents like PDF files as they are independent of software, hardware or the Operating System used. Even though this format has been evolved over time and newer versions have been improved to provide better security, it is still not immune to flaws and attacks that could be misused. Unfortunately, most authors and organizations are unaware that these documents can compromise sensitive information. In this chapter, we have been able to measure the quality and quantity of information exposed in the PDF files published by different security agencies around the world and preprints of scientific community. We have also measured the adoption of PDF files sanitization by these organizations. We identified only 7 security agencies which sanitize few of their PDF files before publishing. Unfortunately, we were still able to find sensitive information within 65% of these sanitized PDF files. We also observed that sanitization was not popular in the scientific community, we found only one poorly sanitized PDF file published by a researcher. In the second part of our work, we have analyzed different PDF files created using several popular PDF producer tools. Observing the patterns used by these producer tools, we developed a methodology to detect the software that has been used to produce a PDF file based on its coding style.

Portable Document Format also known as **PDF**, is a format that is arguably the most popular way to exchange documents on Internet. Adobe invented the PDF file to exchange documents reliably and hence, these files can be viewed and printed the same way on any device. PDF is an open standard, maintained by the International Organization for Standardization (ISO), PDF files meet ISO 32000 standards for electronic document exchange. PDF format is rich and hence apart from text and images, it also provides a facility to include links, buttons, form fields, audio, video, business logic and it can also be signed electronically[1]. Since 2008, PDF is standardized as an open format ISO and the latest version of the standard is PDF 2.0 [56].

PDF file formats not only contains all the information (authors identity, organization...) that its authors have decided to provide to the readers. But it also contains some hidden information which are not often provided by the authors or they are not aware of its presence. These hidden information can be easily exploited by attackers to footprint and attack a

[1] `https://acrobat.adobe.com/us/en/acrobat/about-adobe-pdf.html`

tarted author or an organization. In order to understand PDF files, we first look at its basic structure.

3.1 PDF in a nutshell

PDF has been created by Adobe Systems in 1993 to extend Postscript. PDF includes information such as fonts, hyperlinks, instructions for printing, images, keywords for search and indexing etc. The data in the PDF files is stored in streams of objects which are mostly encoded and/or compressed differently by their PDF producer tools. PDF files have structure and navigation capabilities, their organization and syntax is defined by `Adobe Systems`[2] in an object oriented format. The most comprehensive description of the basic features of a PDF file (version 1.4) is available in [87].

Figure 3.1 describes the basic structure of a PDF document [87]. The file itself is organized into four parts: *header*, *body*, *cross reference table* and *trailer*. In this section, we introduce the basic structure of a `PDF` file.

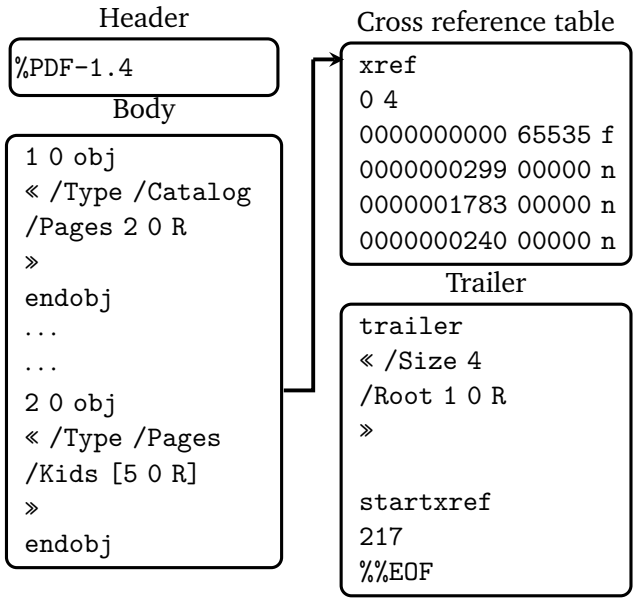

Figure. 3.1: `PDF` in a nutshell.

[2]https://www.adobe.com/devnet/pdf/pdf_reference.html

3.1.1 Header

All PDF files start with the same magic number %PDF (0x25 50 44 46) with % being the comment symbol. The version of specification used to encode the file is then appended: -m.n where m is the major number and n is the minor number (example - 1.4). The header can be composed of a second comment line if the file contained some binary data which is often the case. It consists of at least four binary characters [87]. These binary characters may vary across different PDF producer tools, for instance Microsoft Office Word tool uses 0xB5B5B5B5 and Ghostscript uses 0xc7ec8fa2. The header information is used by PDF viewers to determine if a file follows the PDF format.

3.1.2 Body

The body section of a PDF file is a collection of indirect objects. It is the actual content of the document as viewed by the user. These objects represent the fonts, pages, sampled images and object streams of the PDF file. There are eight objects: *boolean, numbers, strings, names, arrays, dictionaries, streams* and *the null object*. All the objects in the document are labeled so that they can be referred to by other objects. The label is unique for each object in the PDF file. Each object is identified by two positive integers, the first number is the object number followed by a second number that is used as a generation number. Initially all the generation numbers are set to zero and can be changed when the document is updated.

Listing 3.1 shows the syntax of an object consisting of object identifier, dictionary and stream objects. *X* is an object number (identifier) followed by generation number *0* and the keyword *obj*. Content of the object is enclosed between keyword *obj...endobj*. «*keystrings*» is a dictionary object, a dictionary object consists of sequence of key-value pairs enclosed in « ». Dictionary object is followed by the *stream* object and it consists of sequence of bytes enclosed between *stream...endstream*, this sequence of bytes is the actual content of the PDF file as viewed by the user.

```
1 X 0 obj
2 <<keystrings>>
3 stream
4 .....
5 endstream
6 endobj
```

Listing 3.1: A PDF object in the body section.

3.1.3 Cross reference table

Also called the xref table is the only component within the PDF file that follows the same organization across different tools. It consists of a collection of entries which gives the byte offsets for each indirect object in the document and is essentially used for quick and random access to objects. Cross reference table starts with the keyword xref followed by one or more cross reference subsections. For a PDF document that is not yet updated, the cross reference table contains just one subsection entry. When the PDF document is updated by adding the objects, the cross reference subsections are added to the cross reference table.

The cross reference table starts with the keyword `xref` followed by two numbers separated by a space, first number indicates the object number of the first object in the subsection and the second number indicates the total number of entries in the subsection. In the Listing 3.2, cross reference table section consists of single subsection entry with 4 objects from 0 to 4, where 0 is the first object and 4 is the number of entries in the subsection of the cross reference table. 0 4 also indicates that there are 4 consecutive objects from 0 to 4 in this subsection. Then there are cross reference entries one per line associated to exactly one object and it is 20 bytes long and has the format `"nnnnnnnnnn ggggg n/f eol"`, where the first 10 bytes are `nnnnnnnnnn`, indicating the byte offset of the referenced entry, followed by a space and then by 5 digits entry `ggggg`, which is a generation number of the object followed by a space and then a `n`, where `n` is a literal keyword to indicate that the object is in use while `f` is used to indicate that an object is free and this is followed by a space and last 2 bytes constituting the `end-of-line`.

```
1 xref
2 0 4
3 0000000000 65535 f
4 0000000299 00000 n
5 0000001783 00000 n
6 0000000240 00000 n
```

Listing 3.2: Xref/Cross reference table in a PDF File.

3.1.4 Trailer

The trailer part of the PDF document is used for quick access to find the cross reference table and certain special objects in the document. Figure 3.1 shows the syntax of the trailer section.

It is composed of two parts: the first part is a trailer dictionary while the second part defines the *startxref* information.

In the first part, the trailer object consists of some structural information about the document such as different key strings and some other information related to those keys used by the PDF applications. In the Listing 3.3 `/Size [integer]` specifies the number of entries in the cross reference table. `/Root [integer]` gives information of the reference object for the document catalog object that consists of various pointers to different special objects. There are several keys like `/Info, /Prev, /Encrypt, /ID, /XrefStm...` with specific references, particular meaning. All these keys are used differently by producer tools.

```
1 trailer
2 <<
3 /Size 4
4 /Root 1 0 R
5 >>
6
7 startxref
8 217
9 %%EOF
```

Listing 3.3: Trailer section in a PDF File.

Second part of the trailer is the `startxref` which points to the pointer to the start of the cross reference table. A PDF reader first goes to the end of the file where the last line of the document consists of `%%EOF`. Above the `%%EOF` is the `startxref` followed by the offset to refer the cross reference table. In Listing 3.3, `[integer]` below `startxref` is the offset where the `'xref'` table is present. `%%EOF` specifies EOF, every time a PDF file is opened, the PDF viewer goes to `EOF` and gets the offset of `xref` table present above `EOF` to read objects.

There are several different PDF producer tools to create a PDF file and these producer tools use these 4 sections as the basis for the creation of a PDF file. Along with these different sections, PDF files also include metadata information.

3.1.5 Metadata

Opposed to its content, a PDF file may include some general information about the title, author, creation date/time, modification date/time etc., this information related to the PDF file is

termed as metadata. Depending on the PDF producer tools used, metadata of a PDF file can either be stored in a `document information dictionary` associated with the document or in a `metadata stream` associated with the document or a component of the document [87].

Document Information Dictionary

The PDF file's trailer can include an optional Info entry that holds the document information dictionary consisting of metadata information associated to the PDF. A small set of key/value fields such as author, title, subject, creation and update dates is defined and can be extended with additional text values if required. This method was used to store metadata until PDF 1.4. Listing 3.4 shows an example of a metadata object stored using a typical document information dictionary.

```
481 0 obj
<<
/Creator(LaTeX with hyperref package)
/Producer(pdfTeX-1.40.18)
/CreationDate (D:20210409102510+02'00')
/ModDate (D:20210409102510+02'00')
/Trapped /False
/PTEX.Fullbanner (This is pdfTeX, Version 3.14159265-2.6-1.40.18
                  (TeX Live 2017/Debian) kpathsea version 6.2.3)
>>
endobj
```

Listing 3.4: Example of a metadata object stored as Document Information Dictionary.

Metadata Stream

Either for the complete PDF document or for components within the PDF file, metadata information can be stored in streams called metadata stream. In PDF 1.4, support was added for Metadata Streams, using the `Extensible Metadata Platform (XMP)` to add `Extensible Markup Language (XML)` to add XML standards-based extensible metadata as used in other file formats. Listing 3.5 shows an example of a metadata object stored as a metadata stream.

```
19 0 obj
<</Type/Metadata/Subtype/XML/Length 3009>>
stream
<?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc9d"?><x:xmpmeta xmlns:x="
    adobe:ns:meta\/" x:xmptk="3.1-701">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#">
<rdf:Description rdf:about="" xmlns:pdf="http://ns.adobe.com/pdf/1.3/">
<pdf:Producer>Microsoft Word for Office 365</pdf:Producer>
</rdf:Description>
<rdf:Description rdf:about="" xmlns:xmp="http://ns.adobe.com/xap/1.0/">
<xmp:CreatorTool>Microsoft Word for Office 365</xmp:CreatorTool>
<xmp:CreateDate>2019-04-16T10:34:39+02:00</xmp:CreateDate>
<xmp:ModifyDate>2019-04-16T10:34:39+02:00</xmp:ModifyDate>
</rdf:Description>
</rdf:RDF></x:xmpmeta><?xpacket end="w"?>
endstream
endobj
```

Listing 3.5: Example of a metadata object stored as Metadata stream.

3.2 Hidden Information in PDF files

Over time, to improve the content and the user experience, the PDF format has evolved to support more features like security, searchability, description by metadata etc.. Creators of PDF files are often unaware that these features can also expose a lot of information. Even though this format has been evolved over time and later versions have been improved to provide better security, it is still not immune to flaws and attacks that could be misused by attackers. Many works have been done in the past on the PDF file security [103, 100, 17, 115, 68] and privacy [99, 57, 42]. Some works also include results on PDF file sanitization [13, 42, 39]. Aura *et al.* [13] were the first to realize that PDF can contain identifying information which are not obviously visible when the file is created or viewed. Garfinkel [42] also discussed hidden information found in PDF document including images, cropping, comments and of course the metadata. During their work, they have provided files for demonstration.

The visible content of a PDF file can directly reveal many information on the authors: their names, their organization... A PDF file may sometime contain hidden data, metadata and embedded content that are not often provided. This information is mostly invisible to anyone who is not looking for it (anyone who merely opens, views, edits or prints the file). There exists several ways to retrieve this hidden information using appropriate software. By inspecting a

PDF file, it is possible to recover the Operating System and also different software (references, pictures etc.) used to create the file as well as many other information.

The hidden information in documents is considered as a security issue by National Security Agency (NSA) in [106]:

- **Metadata and Document Properties** - In addition to the visible content of a document, most office tools, such as MS Word, contain substantial hidden information about the document. This information is often as sensitive as the original document and its presence in downgraded or sanitized documents has historically led to compromise.

During our analysis of PDF files, we observed that a PDF file is basically collection of indirect objects. There are eight types of objects: boolean, numbers, strings, names, arrays, dictionaries, streams and the null object. Using these objects, data is stored within the file. It is observed that these objects within a PDF file may contain different hidden information like *metadata, document info, attachments, JavaScript actions, links, form fields, comments, unused resources, unreferenced data, hidden layers, overlapping objects, embedded search indexes, annotations and hidden text*.

The following section explains different types of hidden information that could be found in the PDF files. We have summarized them into three groups: **metadata, hidden data in images and other hidden information.**

3.2.1 Metadata

Electronic documents like word files, PDF files, image files and other file formats mostly contain metadata which is embedded inside the record itself. Typically, PDF metadata is populated automatically by the PDF producer tools. PDF file's metadata provides additional information about the file and includes information such as document's title, producer, creator, author, creation time and modification time etc.. The metadata are often used in cataloging to help searching for documents in external databases.

As discussed earlier, PDF Metadata are stored either in a document information dictionary or in a metadata stream. When the metadata are stored in a document information dictionary, the PDF file's trailer section includes an optional Info entry that holds the document information dictionary consisting of metadata information associated to the PDF file. Metadata information can also be stored in streams called metadata stream. It is either for the complete PDF file or

for some components within the PDF file. Metadata stream is represented using Extensible Markup Language (XML).

There are several ways and tools that could be used to view and extract metadata information from PDF files: `exiftool`[3], `Metagoofil`[4] and `pdfxplr`[5] to name a few. Table 3.1 shows the metadata of a PDF file. During our analysis, we noticed that number of metadata fields and their names depend on the PDF producer tool used. Some software also provide an option to the user to either add or remove metadata fields while creating a PDF file.

ExifTool Version Number	: 11.49
File Name	: 127-Zadost-2-10-2018.pdf
Directory	: ./www.vzcr.cz
File Size	: 259 kB
File Modification Date/Time	: 2019:04:04 13:21:11+02:00
File Access Date/Time	: 2020:08:21 13:22:52+02:00
File Inode Change Date/Time	: 2020:07:31 11:23:04+02:00
File Permissions	: rw-r–r–
File Type	: PDF
File Type Extension	: pdf
MIME Type	: application/pdf
PDF Version	: 1.5
Linearized	: No
Page Count	: 1
Language	: cs-CZ
Tagged PDF	: Yes
Title	: MINISTERSTVO OBRANY
Author	: chocholaty
Creator	: Microsoft Word 2010
Create Date	: 2019:04:04 13:16:51+02:00
Modify Date	: 2019:04:04 13:16:51+02:00
Producer	: Microsoft Word 2010

Table. 3.1: Metadata information of a PDF file (extracted using `exiftool`).

In the past, many studies have been made on the privacy impact of PDF metadata [8, 100, 42, 74]. Smutz *et al.* [100] have built a classifier to detect malicious PDF based on the metadata

[3]https://exiftool.org/
[4]https://tools.kali.org/information-gathering/metagoofil
[5]https://github.com/sowdust/pdfxplr

and structural features of the file, unfortunately they have not provided more description on the structural features they have exploited. Mendelman has demonstrated in his master thesis [74] that PDF metadata can be used to fingerprint an organization. He has gathered 1580 PDF files from 3 organizations and was able to collect printer names, internal domain names, Operating Systems, personal information and producer tools. Aura *et al.* [13] have tested 43 anonymous PDF submissions of a conference to detect if any leakage was present. They found 3 submissions which were not properly anonymized. One submission contained a metadata field with the authors names. They propose a tool based on regular expression to detect usernames, device or organization names, identifiers and other information in a PDF file.

3.2.2 Hidden data in images

Images embedded in a PDF file can provide many information on the authors. If an author cleans the metadata of the PDF document and forgets to clean the metadata of the embedded image files, one can still extract information on the author. Image metadata can include author name, username, path from where the file was inserted... For a digital photograph, the metadata might include the information about the type of camera or other device/software used to create the image, the location where the picture was taken and the date/time. It can also include some *artifacts* that can leave traces of how an image has been edited or modified, information on different software used and the habits of the author. Geo-location features stores the geo-data revealing the location where the image was taken this kind of information may impacts the security and privacy of an individual.

All the information hidden in images can be retrieved using an appropriate software. Several tools like `ImageMagick`[6], `exiftool` and `exiv2`[7] can be used to view metadata. We have used `pdfxplr`[8] in our work to extract the `Username` and `PATH` information from the image files embedded within the PDF documents. Listing 3.6 shows the `PATH` extracted from one of the PDF file containing a image file. We can see that the author is providing the location where the image file was stored along with the username.

[6] `https://imagemagick.org/script/download.php`
[7] `http://manpages.ubuntu.com/manpages/hirsute/en/man1/exiv2.1.html`
[8] https://github.com/sowdust/pdfxplr

```
19693 0 obj
<<
/K 29/P 19690 0 R/S/InlineShape/Alt(Description: C:\\Users\\Mazhar\\
    Desktop\\scml.JPG)/Pg 19761 0 R
>>
endobj
```

Listing 3.6: A PDF object displaying the PATH information of an embedded image file.

Image extraction and analysis was proven to be successful in [13, 39] to recover information of the authors of PDF file. Aura *et al.* [13] have tested 43 anonymous PDF submissions of a conference to detect if any leakage was present. They found two PDF files that had a image with identifying metadata. Feng *et al.* [39] have also proposed a tool to detect privacy leakages in PDF files with a focus on extracted images. The main difference of their tool with the tool of Aura *et al.* [13] is the use of text mining, information retrieval and natural language processing to detect identifying information.

3.2.3 Other Hidden Information

A PDF file may include a variety of different embedded content and newer versions of PDF also support multimedia such as Flash, Windows Media Video and QuickTime content, scripts, form fields, stored user data and form processing information. Each of these content types may contain hidden data or metadata. Then there are unreferenced objects, comments, annotations that may be inserted into the binary data that are not displayed but can be accessed. PDF files can also include obscured text like white text on a white background, text that is inadvertently hidden behind images. Such information is easily extractable if all the contents are copied and pasted into a text editor.

Comments can be inserted anywhere in the PDF file by starting a line with % symbol, PDF viewers just ignore the comments and display the content stored in PDF objects. Comments are often a good source of information. Listing 3.7 shows an example of how comments can be included within the PDF files. We found a tool `PDF HIDE`[9] which is a steganographic tool implemented in Python for hiding data in PDF files. Authors in [37], also exploited these features and have proposed to hide secrets inside multiple PDFs by scattering the information. These kind of changes are undetectable if the PDF is viewed in the normal mode. They can only be detected by looking inside the PDF document's binary file.

[9] https://github.com/ncanceill/pdf_hide

```
4 0 obj
<</Length 6922>>
stream
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%       Comment in one line at a time                 %
% an AT&T facility containing networking equipment ....%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
endobj
```

Listing 3.7: Example to show how comments can be included in PDF files.

Annotations can stay invisible when a PDF file is viewed. However, they can be retrieved using simple string extraction commands like `strings`. Often authors forget to remove the annotations from their files and it can expose their identity. Listing 3.8 shows an object extracted from a PDF file that reveals the annotated message: `changes made by john doe` and also the Username: `sab`.

```
152 1 obj
<<
/Type /Annot /Rect [165.897704918 615.5800226116 189.897704918
    639.5800226116 ] /Subtype /Text /M (D:20210225232546) /C [1 1 0 ] /
    Popup 153 1 R /T (\FE\FF\00s\00a\00b) /P 3 0 R /Contents (\FE\FF\00c
    \00h\00a\00n\00g\00e\00s\00 \00m\00a\00d\00e\00 \00b\00y\00 \00j\00o
    \00h\00n\00 \00d\00o\00e)
>>
endobj
```

Listing 3.8: Annotation object of a PDF file containing the message and username.

Specific objects within a PDF file can also include some sensitive information. We found that registry objects, font objects... can include metadata information. Listing 3.9 shows two objects extracted from PDF files that reveal information about the OS and other software used.

```
15 0 obj
<<
/DW 1000/CIDSystemInfo
<<
/Supplement 0/Registry(Adobe)/Ordering(Identity)
>>
/Subtype/CIDFontType2/BaseFont/CAFBBG+TimesNewRomanPSMT/Type/Font/
   FontDescriptor 23 0 R/W[267[610 443] 284[333]]
>>
endobj


1459 0 obj
<<
/Platform(Macintosh)/Creator(FileMaker Pro Advanced 14.0.1)/
   DLI_Copyright(Datalogics Interface \(DLI\) Copyright \(C\) 1998-2012
    Datalogics, Inc. -- www.datalogics.com)/Producer(Adobe PDF Library
   10.1; modified using iText 2.1.7 by 1T3XT)/Title()/Keywords()/
   ModDate(D:20180223153614+01'00')/Subject()/DLI(10.1.0.50)/Author()/
   CreationDate(D:20180220152519+01'00')
>>
endobj
```

Listing 3.9: Specific objects revealing information on the authoring process of a PDF file.

Castiglione *et al.* [20] demonstrated that a PDF can be tracked when it is read. A malicious author can use this technique to obtain information on his/her reviewer during the review process of a conference. It is based on the possibility to embed scripts that are executed by the PDF viewer when the file is displayed. The script needs to download external resources controlled by the author. It exposes the IP address, user-agent and the time of request of the reviewer. It was even discovered in 2019 that embedding a script was not needed to track a PDF file. CVE-2019-8097[10] shows that it is possible to insert an URI in a PDF file which is downloaded automatically by certain viewers. In emails and online tracking, similar techniques are used [99, 57, 111]. The dangers of scripts and dynamic resources in PDF files are well-known and many solutions exist (see [63, 103, 102, 69, 70]) and the method proposed in [20] to track a PDF when it is read is unlikely to work anymore.

No doubt PDF files offer several benefits over other file formats, at the same time, it is also true that they have many possibilities to include hidden information and also malicious content that could compromise author's security and privacy.

[10] https://nvd.nist.gov/vuln/detail/CVE-2019-8097

3.3 Impact of hidden information in PDF files

In this section we describe different information that we were able to extract on the author and on the organization from PDF files. We have conducted a large scale study of the PDF files published by two entities: security agencies around the world and preprints of scientific community (595529 PDF files in total).

Motivation to choose these organizations

- **Security agencies**:The National Security Agency (NSA) is one of the main actor in cyber security and it has provided many guidelines to sanitize PDF files before sharing them [106, 80, 81]. This was a motivation in our work to check if security agencies around the world clean their PDF files before publication. They are supposed to have the strongest and best security practices and hence we decided to work on PDF files published by them. We also wanted to know if any security agency care to follow NSA guidelines on PDF sanitization. We have crawled the websites of 75 security agencies mentioned by Wikipedia[11] belonging to 47 countries. We downloaded 39664 PDF files in total. The distribution of PDF files over the agency is not even. We found between 5 to around 6000 PDF files for each agency. Figure 3.2 shows the discrepancy in the number of PDF files for each country in our dataset. We have used `wget` command to crawl websites and download PDF files. From now onwards, we call this dataset of PDF files as the **security dataset** for the convenience.

- **Preprints of scientific community**: We observed that, over 1.6 million PDF files have been published on arXiv[12] and 0.8 million have been uploaded on the open archive HAL[13]. Researchers manipulate PDF files on a daily basis to learn new results, write proposals and also to review articles or conference papers and to submit their work to different venues. Since researchers publish many PDF files, they are particularly exposed to an adversary. We wanted to discover if the PDF sanitization is taken seriously by researchers in the scientific community or not. It is a common practice for researchers to submit their work to conference proceedings and preprint servers, hence they are clearly the best options for an adversary to obtain interesting PDF files on authors. Our first preprint dataset was downloaded from the Cryptology ePrint Archive[14], it includes 11405 PDF documents from 2004 to 2018. We observed that all the PDF documents were originally compiled by the authors using a PDF producer tool of their choice. Open

[11] `https://en.wikipedia.org/wiki/Security_agency`
[12] `https://arxiv.org/`
[13] `https://hal.archives-ouvertes.fr/`
[14] https://eprint.iacr.org/

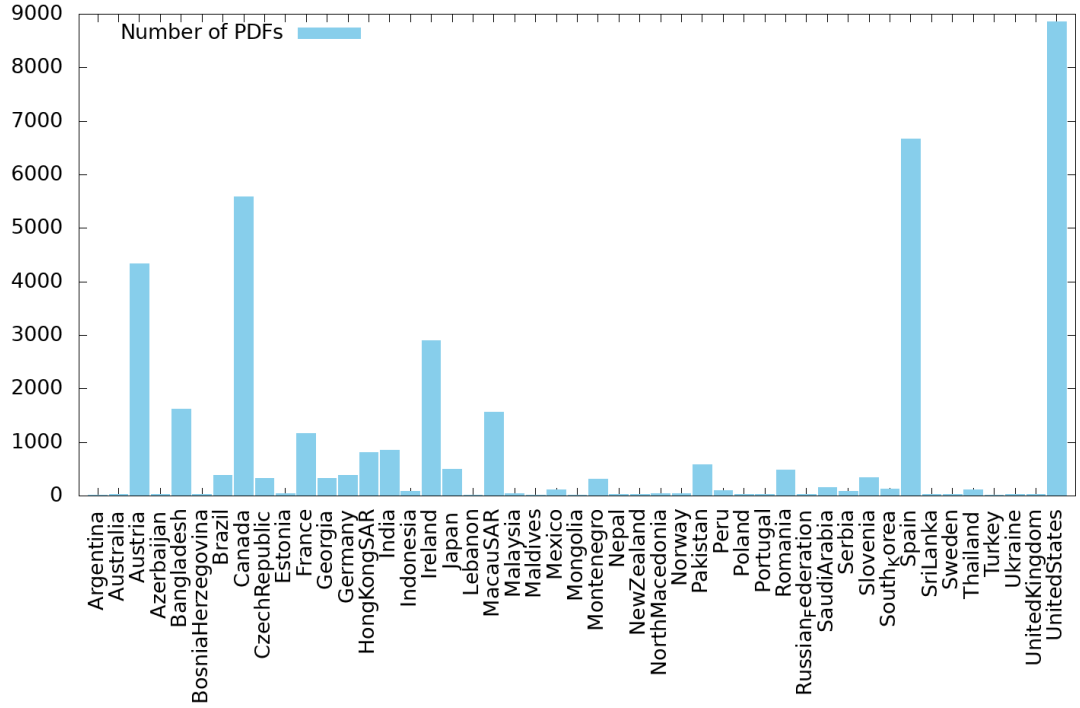

Figure. 3.2: # PDF files for each country in Security dataset.

archive HAL[15] administrators provided us with an access to a second dataset of PDF files and it consists of 544460 PDF files from 1996 to 2019. They are respectively called IACR dataset and HAL dataset for convenience.

In this section we explain the different information that we could exploit from the PDF files of security agencies and preprints of scientific community. We have analyzed the metadata of the PDF files using `exiftool` (Table 3.1) and some other hidden data like path, e-mail addresses etc.. using `pdfxplr` tool.

Please note that, we assume that the metadata can be 100% trusted. There exists a possibility that metadata fields were altered by the authors and some values can be decoys or random noise but we cannot detect such situations. We consider metadata information to be reliable information. It is also important to notice that we did not know if the authors of the PDF files were some employees of the security agencies or if they were working for third party companies that work for the security agencies. We do not know the ground truth, we assume that the authors of the PDF files were working for a security agency. Similarly for scientific

[15]https://hal.archives-ouvertes.fr/

community, PDF files might have been created by someone in the university or organization they work for and not directly by the author himself.

We found different results for both security agencies and preprints dataset and based on the targeted individuals, an adversary may look for different kinds of information. Hence, our results on the analysis of PDF files are presented separately for both security agencies and preprints.

3.3.1 Security Agencies

We have analyzed the content (as viewed by the viewer) of all the 39664 PDF files to find if the names of the author appear directly within the document. Using string extraction commands, we found the author names in only 1783 (4%) PDF files and the rest of the 37881 (96%) PDF files were anonymous.

Information leaked on the authors

Name of the author: During our analysis of PDF files, we observed that three metadata fields `author`, `creator` and `Tag Author Email Display Name` can reveal the name of the author producing the PDF file. In our dataset, 13166 (33%) PDF files reveal the identity of the individual who have created the file.

PDF producer tool: This information can be exposed in the metadata fields like `producer`, `creator` and `creator tool`. We found that, in our dataset 30155 (76%) PDF files include the metadata information on the PDF producer tool used. `Acrobat Distiller`, `Microsoft Office Word` and `Adobe PDF Library` are the most popular tools in our dataset (see Table 3.2).

Operating System: Interestingly, producer tool name in the metadata can sometimes reveals the Operating System (OS) used by the author. This is due to the fact that some tools are OS specific and many others have a label which includes OS information: *Mac OS X 10.6.6 Quartz PDFContext, Acrobat Distiller 20.0 (Macintosh), Acrobat Distiller 8.3.1 (Windows) or Antenna House PDF Output Library 6.2.553 (Linux64)*. In our dataset at least 16805 (42%) of the PDF files reveal OS information. Table 3.3 shows the distribution of PDF files between the three main Operating Systems: Microsoft Windows, Mac OS and Linux. We can see that Microsoft Windows is a popular choice among the employees of many agencies.

Producer tool	# PDF	# Agencies
Acrobat Distiller	9054 (23%)	46
Adobe PDF Library	6171 (16%)	50
Microsoft Office Word	4850 (12%)	66
LibreOffice	2171 (5%)	07
Ghostscript	1133 (3%)	36
Mac OS X Quartz	94 (0.2%)	20
SKia/PDF	106 (0.2%)	08
Other tools	16085 (40.5%)	75

Table. 3.2: # of PDF files associated to popular PDF producer tools (76% PDF files).

OS used	# agencies	# PDFs
Microsoft Windows	71	11,174 (28%)
Mac OS	29	3,444 (8%)
Linux	7	2,187 (6%)

Table. 3.3: # of agencies and the choice of OS used.

Brand of the device, e-mail and Path information: We observed that sometimes authors reveal the brand of their hardware device in the metadata fields `author, creator tool` and `creator` instead of their name or along with their name. We found four brands like `Toshiba, HP, DELL and Lenovo`. Authors of at least 24 security agencies have such practices (Table 3.4). Using the `pdfxplr tool`, we could also extract different information like the e-mail address of authors, the PATH or location of the folder from where images/files were inserted within the PDF files. We found 52 personal or official e-mail addresses in the metadata field `Tag Author Email, author` and `Current User Email` etc.. 47 of these e-mail addresses are official e-mail address of the employees associated to the security agencies they work for. Four are `gmail addresses`, one of them is `outlook address` (Table 3.4).

OS	# PDF	# Agencies
E-mail	52	13
Hardware brand	581	24
Paths	1814	47

Table. 3.4: # of PDF files revealing e-mail, hardware and PATH information.

As mentioned previously, images/files included within the PDF file can also contain the metadata and the path from where the images are included. We did not exploit this possibility in

Agency	Author Name	Author habits	Year	PDF producer tool	# PDF published
fia.gov.pk	**Author-X**	**Using same tool**	**2014-19**	**Microsoft Office Word 2007**	**29**
defensa.gob.es	Author-Y	Updating regularly	2010	Acrobat Distiller 7.0.5 (Windows)	4
			2011	Acrobat Distiller 8.0.0 (Windows)	1
			2011-14	Acrobat Distiller 8.2.5 (Windows)	9
			2014-15	Acrobat Distiller 10.1.0 (Windows)	26
			2017-18	Acrobat Distiller 11.0 (Windows)	3
customs.gov.hk	Author-Z	Changing tools	2017	Adobe Acrobat 11.0.20	1
			2018	Adobe Acrobat 11.0.0	1
			2019	PDFCreator 2.1.2.0	3
			2019	PDFCreator 3.2.2.13517	2
			2019-20	Adobe Acrobat Standard 2017 17.11.30150	2

Table. 3.5: Interesting author behaviors observed using PDF producer tool information.

our work. We found complete image paths for 1814 PDF files using `pdfxplr` tool (Table 3.32).

Combining information: *It is possible to combine author, producer and time information provided in PDF metadata fields to understand how employees in security agencies update or change their PDF producer tools*. We have provided three examples in Table 3.5. Author-X is working at Federal Investigation Agency (Pakistan) and he/she has never changed/updated his/her PDF producer software from 2014 to 2019. This author is using `Microsoft Office Word 2007` software which is a older version of the software and may contain some vulnerabilities that could be exploited. Author-Y works for the Spanish Ministry of Defense and updates his/her `Acrobat Distiller software` on a regular basis. Author-Z is working at the Customs and Excise Department (Hong Kong) and produced PDF files using Adobe Acrobat. Author-Z sometimes used `PDFCreator` tool to convert documents into PDF files on Microsoft Windows Operating System.

Information leaked on the organization

Now we show that it is possible to aggregate the information on the authors to obtain results on an organization. The main goal is to observe the trends in the usage of software and OS in the organizations.

By aggregating the PDF files published for several years, we have considered three different profiles for employees on the organization level (behaviors previously shown in Table 3.5). **Profile-1** employees update their software on a regular basis. **Profile-2** employees change their software. Finally, **Profile-3** employees do not change their software during a period of

at least 2 years. Figure 3.3 shows the number of employees following **Profile-1**, **Profile-2** or **Profile-3** for each security agency. In our dataset, we found at least 19 agencies that have 154 employees following **Profile-3** and who do not change/update their tools over a period of two years or more.

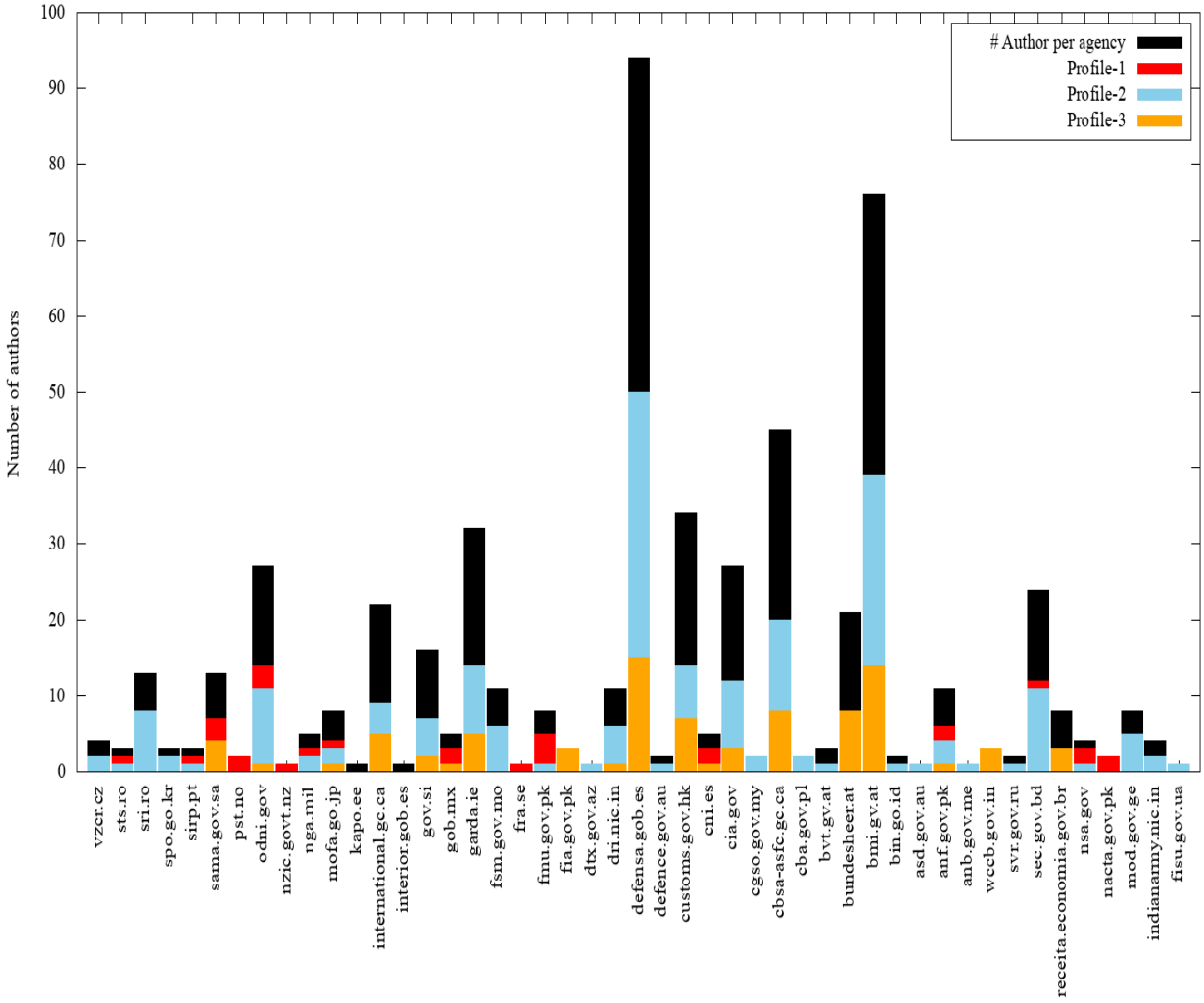

Figure. 3.3: Profile of the employees working in security agencies.

OS used by agencies: In our dataset, OS details are revealed in 16805 (42%) PDF files. Table 3.15 shows the distribution of PDF files between the 3 main Operating Systems: Microsoft Windows, Mac OS and Linux. It is also possible to spot how many Operating Systems are used in a security agency. We give the example of Austria's Interior Ministry (`bmi.gv.at`). Figure 3.4 shows the different OSs we have been able to spot during the last 24 years. In the last five

year	# agencies using same OS	# agencies using multiple OS (mix)
2000-2005	15	9
2006-2010	18	10
2011-2015	33	17
2016-2020	37	26

Table. 3.6: # of agencies and OS used.

years, the authors of this agency mainly used Microsoft Windows OS to produce their PDF files and fewer PDF files have also been created using MAC OS.

OS Trends: An organization can either allow its employees to use any OS or every employee uses the same OS. Table 3.6 shows the use of OSs by agencies for a period of 20 years that we have observed in the metadata. Even though the number of agencies using same OS is more, we can observe that in the last five years, number of agencies leaning towards using multiple OSs are increasing.

3.3.2 Scientific community

We observed that the PDF files in IACR and HAL dataset are not anonymous, *i.e.* they include the name of the author(s). Also, it is possible to find a description (metadata) of author names, title, year of publication of each PDF file on there respective websites. Both metadata field names and values can expose information on the authoring process. Therefore, we have analyzed metadata field name and values separately and describe our findings as follows.

Metadata Field Names: we found 153 unique fields in IACR dataset and 2190 unique fields in HAL dataset. We also observed that some fields are very common like `Producer` or `author` etc. and others like `PTEX Fullbanner`, `Google Documents Tracking` are very rare. We define their probability of occurrence below 0.0001. Rare values are the most interesting metadata fields since fewer PDF files include them and they could be exploited to identify author(s). Table 3.7 shows the number of rare metadata fields in IACR and HAL dataset and Table 3.8 shows the probability of each metadata field. HAL dataset includes many different rare fields, this is due to the size of this dataset and because many different organizations contribute to HAL. For IACR dataset, we observed that nearly 91% of the PDF files were produced using three PDF producer tool (PdfTeX, xdviPDFmx and Ghostscript tool-see Table 3.31), this explains the reason for rare metadata fields to be not frequent in IACR dataset.

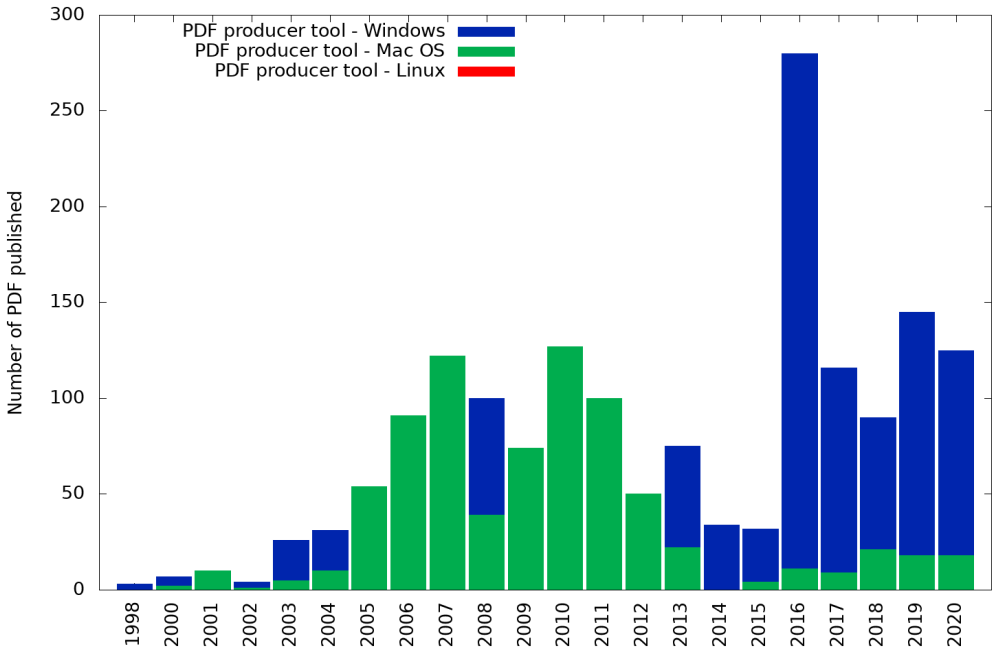

Figure. 3.4: Use of different OS over years at `bmi.gv.at`.

Description	IACR	HAL
# other fields	46	1079
# rare fields	110	1111
# fields	153	2190

Table. 3.7: Number of rare metadata fields in IACR and HAL dataset.

Software, language and organization details: We were surprised to realize that just the name of the rare fields is already enough to learn information about the authoring process of a PDF file. We found the occurrences of `Google`, `Mendeley` field names.We have classified the information provided by the field names into three categories: software, language and organization. Table 3.9 shows the number of metadata fields and number of PDF files that reveal such information in our dataset.

Our analysis of metadata field names is rather not conclusive for IACR dataset. Only 8% of the PDF files have a specific field names. This number is 43% for HAL dataset. There is no need to perform complex forensics to obtain information. Table 3.10 shows the different software used by authors.Software like Bibtex and Mendeley provide the habits of authors for managing their references. Full banner (nearly 13% of PDF files) provides the information on the LaTeXPDF producer tool, the distribution, its version number and OS used. All such

Field	Probability	Field	Probability	Field	Probability
About	0.0509	Aggregation Type	0.0003	Apple Keywords	0.0009
Application	0.0004	Appligent	0.0004	Attribution Name	0.0004
Attribution URL	0.0004	Author	0.4119	Author Info Name	0.0001
Author Info Orcid	0.0001	Authors Position	0.0001	Conformance	0.0018
Comments	0.0014	Company	0.0085	Caption Writer	0.0002
Copyright	0.0003	Create Date	0.983	Creation Date	0.0043
Creation Date–Text	0.0002	Creator	0.9695	Creator Address	0.0001
Creator City	0.0001	Creator Country	0.0001	Creator Postal Code	0.0001
Creator Tool	0.0972	Creator Work Email	0.0001	Format	0.0987
Cross Mark Domains	0.0003	Description	0.0237	Description (pdflang)	0.0002
Cross Mark Domains 1	0.0003	Crossmark Major Version Date	0.0003	Date	0.0103
Cross Mark Domains 2	0.0003	Derived From Document ID	0.0002	Derived From Instance ID	0.0002
Description (x-repair)	0.0004	Digital Object Identifier	0.0003	Document ID	0.0989
Doi	0.0003	Encryption	0.0016	Crossmark Domain Exclusive	0.0003
GIT Rev	0.0001	GTS PDFA1 Version	0.0001	GTS PDFX Conformance	0.0002
GTS PDFX Version	0.0002	Has XFA	0.0027	Headline	0.0076
HG Rev	0.0004	History Action	0.0004	History Instance ID	0.0004
History Parameters	0.0004	History Software Agent	0.0004	History When	0.0004
ICN App Name	0.0001	ICN App Platform	0.0001	ICN App Version	0.0001
ID	0.0001	Identifier	0.0017	Instance ID	0.0482
ISSN	0.0003	Journal article version	0.0002	Keywords	0.0417
Language	0.0163	License	0.0004	Linearized	1.0
Major Version Date	0.0003	Manifest Link Form	0.0001	Modify Date	0.78
Manifest Placed X Resolution	0.0001	Manifest Placed Y Resolution	0.0001	Manifest Reference Document ID	0.0001
Manifest Reference Instance ID	0.0001	Marked	0.0029	Mod Date	0.0041
Metadata Date	0.044	MIME Type	1.0	Manifest Placed Resolution Unit	0.0001
More Permissions	0.0004	MT Equation Number 2	0.0004	MT Equation Section	0.0003
MT Preferences	0.0004	MT Preferences 00201	0.0004	MT Preferences 00202	0.0004
MT Preferences 00203	0.0004	MT Preferences 1	0.0004	MT Preferences 2	0.0004
MT Preferences 3	0.0004	MT Preference Source	0.0004	MT Use MT Prefs	0.0001
MT Win Eqns	0.0022	Page Count	0.9996	Page Layout	0.0126
Part	0.0018	PDF Version	1.0	Producer	0.9986
PTEX Fullbanner	0.6801	PTEX Full Banner	0.0007	Publication Name	0.0003
Publisher	0.0004	PXC Viewer Info	0.0008	Rendition Class	0.0004
Rights	0.0025	Rights (en-US)	0.0001	Rights (pdflang)	0.0002
Robots	0.0003	Schemas Namespace URI	0.0006	Schemas Value Type Prefix	0.0002
Schemas Property Category	0.0006	Schemas Property Description	0.0006	Schemas Property Name	0.0006
Schemas Property Value Type	0.0006	Schemas Schema	0.0006	Schemas Value Type Description	0.0002
Schemas Value Type Field Description	0.0002	Schemas Value Type Field Name	0.0002	Source	0.0082
Version ID	0.0078	Source Modified	0.0094	SPDF	0.0002
Schemas Value Type Namespace URI	0.0002	Subject	0.3747	SVN Rev	0.0001
Schemas Prefix	0.0006	Schemas Value Type Type	0.0002	Tag Doc Home	0.0001
Tagged PDF	0.0231	Title	0.4781	Title (en-US)	0.0001
Title (pdflang)	0.0003	Trapped	0.6324	Type	0.0053
URL	0.0003	Usage Terms	0.0004	Usage Terms (en)	0.0004
Usage Terms (fr)	0.0004	User Access	0.0016	Version	0.0002
Schemas Value Type Field Value Type	0.0002	Warning	0.0026	Web Statement	0.0024
XMP Toolkit	0.1001	Page Mode	0.3723		

Table. 3.8: Probability of all the metadata fields in IACR and HAL dataset.

Description	IACR		HAL	
	# Field	# PDF	# Field	# PDF
Software	6	1162 (7%)	805	126996 (23%)
Language	4	10 (0.08%)	57	110452 (20%)
Organization	6	4 (0.03%)	69	5587 (1%)

Table. 3.9: Information leaked by metadata field names.

information is sensitive and the results in Table 3.10 shows that authors do not pay any attention to the information leaked in the PDF metadata.

Software name	# PDF	Organization	# PDF
Full banner	72586 (13%)	Elsevier	5320 (1%)
Mendeley	7526 (1.3%)	IEEE	154
Apple	6702 (1.2%)	OECD	60
ZOTERO	3408 (0.6%)	Yann Desjeux	10
Bibtex	537	Medarb	6
Prism	162		
Google	65		
Microsoft	10		

Table. 3.10: Software names revealed in PDF files of HAL dataset.

Metadata Field Values: After analyzing the metadata field names, we analyzed the metadata field values to extract sensitive information. Metadata field values in our dataset reveals information on the OS (Microsoft Windows, Mac OS and Linux), country name, organization of the author as shown in the Table 3.11.

Author details	IACR	HAL
OS (Linux, Microsoft Windows, Mac OS)	1703 (15%)	121427 (22%)
Organization Details	107 (1%)	11328 (2%)
Country	32 (0.2%)	156432 (29%)

Table. 3.11: Sensitive information revealed in our datasets.

Table 3.12 shows three popularly used OS and few organization names present in the field value of HAL dataset. PDF metadata includes the information of the organization they work for and also the OS used to produce their PDF document.

We saw that, just the metadata field names and field values can be used to exploit author's habits, software and the organization. Results in Table 3.9, 3.10, 3.11 and 3.12 show the possibilities to find information about the author using metadata.

OS	# PDF	Organization	# PDF
Microsoft Windows	72372 (13%)	CNRS	712 (0.1%)
Mac OS	6036 (1.1%)	Hewlett-Packard	468 (0.08%)
Linux	16815 (3%)	Microsoft	391 (0.07%)
		INRIA	23
		Univ. of Manchester	7

Table. 3.12: Details on the information.

Producer tool information: Previous observations have been made without taking producer field into consideration. We now focus on the `producer` field and the information revealed by this single metadata field. We examined all the PDF files in our dataset and found that 99.85% PDF files in IACR and 99.30% PDF files in HAL dataset contained the producer field in the metadata.

We discovered that in IACR dataset there were 449 unique producer tools while in HAL dataset there were 3699 unique producer tools. In Table 3.31 we list some of the producer tools we found in our dataset. Additionally, we noticed that few PDF producer tools were used infrequently and hence, we term them as rare producer tools. Subsequently, we found that in IACR dataset there were 46 rare producer tools while in HAL dataset there were 1031 rare producer tools. (Rare producer tools are computed considering the PDF producer tools with PDF files count >1 and ≤ 25 in our dataset). Table 3.31 shows the diversity in the producer tools and the number of PDF document produced by them. Distribution of PDF files among HAL dataset is interesting to analyze due to the variation in number of PDF files associated to PDF producer tools. This variation also leads to higher number of rare metadata fields in HAL dataset.

`PdfTeX` in Table 3.31 is the label of LaTeX software whereas `PDFLaTeX` is a standard label used by HAL for there PDF producer tool. We can observe that only 1 PDF file was produced using `PDFLaTeX` tool in IACR dataset while 46% of the PDF files in HAL are compiled by `PDFLaTeX`. It is interesting to note that one of the PDF file produced using `PDFLaTeX` in IACR dataset was initially submitted to HAL and was produced using HAL PDF producer tool and then re-submitted to IACR preprint.

Linking producer tool to authors: Authors are more likely to use the same PDF producer tools for producing different PDF documents. Therefore, we attempt to link producer tool to the author. Considering rare PDF producer tools, we managed to easily link the authors by merely utilizing the the name of the producer tool. In Table 3.14, we show the results obtained for the association of PDF producer tool to an author. We grouped the PDF files

Producer tool	# PDF IACR	# PDF HAL
PDFLaTeX	1	249160 (46%)
Ghostscript	1200 (11%)	115948 (21%)
Acrobat Distiller	369 (3%)	96157 (18%)
Rare tools	513 (4%)	46516 (9%)
PdfTeX	7745 (68%)	10375 (2%)
Microsoft Office Word	147(1.3%)	15995 (2%)
Mac OS X Quartz	62 (0.5%)	7767 (1.4%)
Cairo	0	1100 (0.2%)
xdviPDFmx	1347 (12%)	836 (0.2%)
LibreOffice	4	451 (0.08%)
SKia/PDF	0	88 (0.01%)
LuaTeX	17	67 (0.01%)

Table. 3.13: PDF producer tools and number of PDF files associated to each tool in IACR and HAL dataset.

generated by a rare producer tool and verified the authors of all the PDF files associated to this PDF producer tool. If more than 2 PDF documents of a rare tool match then we associate the PDF producer tool to the author. Table 3.14 shows the results for author association to PDF producer tool where results include, one PDF producer tool used exactly by one author(s), one tool used by two authors and so on. In most cases, all these PDF files submitted by these authors lead to re-identification while others narrow down the search among 2 to 3 authors.

Description	IACR (46)	HAL (1031)
1 tool - 1 author match	29	737
1 tool - 2 authors match	4	150
1 tool - 3 authors match	3	29

Table. 3.14: Author association using producer field of the PDF.

OS information using producer tool: Metadata information provided for the `producer field` can also be used to reveal the Operating System used to generate the PDF document. In our dataset, by merely using the keywords pertaining to popular OS like `Microsoft Windows`, `Mac OS and Linux`, we were able to know the OS. The Table 3.15 shows the number of PDF files that revealed the OS in the metadata field producer.

The information revealed by producer field are very similar to the one exploited for browser-fingerprinting. We also observe from the results in Table 3.15 that OS information could be revealed for few PDF producer tools using just the value of producer field.

OS used by author	IACR	HAL
Microsoft Windows	373 (3.2%)	94466 (17%)
Mac OS	68 (0.6%)	12557 (2%)
Linux	2	969 (0.2%)

Table. 3.15: OS information obtained using producer field.

The detailed analysis and the corresponding results obtained for three dataset (Security, IACR and HAL) in this section show that PDF metadata includes some sensitive information about the authors. If the authors fail to or neglect to clean the metadata in their PDF documents, it can undoubtedly reveal sensitive information on them and the organization they work for.

3.4 PDF file sanitization

Today's world is more interconnected than ever before, increase in the online activities has also increased the risk of theft, fraud and abuse for individuals and organizations. Document sanitization is a critical step for any organizations who wants to publish electronic documents internally or online. Sanitization is required to avoid exposing confidential or sensitive data.Sanitization is the process to ensure that only the intended information is present in a file, it includes removal of all the hidden information that could pose a privacy or security risk on the author.

PDF files needs to be carefully sanitized, there are several solutions available to sanitize a PDF file. NSA provides a list of eleven main types of hidden data, metadata and embedded content that may be found in PDF files [80, 109, 81]. NSA recommends that after all the following eleven types of content are removed then a PDF file is properly sanitized for safer distribution.

1. Metadata
2. Embedded Content and Attached Files
3. Scripts
4. Hidden Layers
5. Embedded Search Index
6. Stored Interactive Form Data
7. Reviewing and Commenting
8. Hidden Page, Image and Update Data
9. Obscured Text and Images

10. PDF (Non-Displayed) Comments
11. Unreferenced Data

There is more than one way to sanitize a PDF file, several softwares are available that could be used locally and many websites also provide an option for PDF sanitization. During our experiment, we have tested some tools and list our observations below.

Adobe Acrobat: is often mentioned in NSA guidelines [106, 80, 81] as a reliable sanitization tool. It cleans the metadata and all the hidden content of the PDF file. This is the most complete sanitization tool we have used in our work. For a safe distribution of a PDF file, NSA [80] recommends to remove all types of hidden data, metadata and embedded content. NSA provides direction to properly sanitize a PDF file using Adobe Acrobat in five steps.

- Step 1: Before converting a file format to PDF, **if possible**, it is recommended to minimizing the presence of hidden data, metadata and embedded content in the source file and then convert to PDF.
- Step 2: is to configure the security settings of Acrobat tool to minimize any risks associated with opening the file for sanitization.
- Step 3: is to run the **Preflight** utility to ensures that the file can be successfully converted to the PDF format.
- Step 4: is to run the PDF Optimizer utility. This step regenerates the PDF content and strips out all the hidden data, metadata and embedded content as well as file attachments.
- Step 5 : is to run the **Examine Document** utility to identify and remove any residual hidden data, especially hidden text.

GhostScript: Converting PDF file to postscript is mentioned online as a sanitization method . This conversion can be done using `GhostScript` tool (Listing 3.10) and it clearly removes a lot of information (multiple XMP entries of metadata) on the resulting PDF file. However, it is difficult to determine what is removed or retained by this conversion.

```
pdf2ps filename.pdf
ps2pdf filename.ps
```

Listing 3.10: PDF metadata sanitization using Ghostscript.

Exiftool: Several threads on sanitization in online forums mention the possibility to use `exiftool` (`https://exiftool.org/`). During our experiments, we observed that this tool only cleans the metadata of a PDF file (Listing 3.11).

```
exiftool -all:all= file.pdf
```

Listing 3.11: PDF metadata sanitization using exiftool.

Text processing software: PDF files can be produced without any metadata using text processing software. This is the case for Microsoft Office Word or LibreOffice.

LaTeXsoftware: It is also possible to include options in LaTeXsources to remove the metadata (see Listing 3.12) or one can just use the `pdfprivacy` package[16].

```
\usepackage{hyperref}
\hypersetup{
pdftitle={},
pdfauthor={},
pdfproducer={},
pdfcreator={},
pdfkeywords={},
}
```

Listing 3.12: Producing PDF file without metadata with `hyperref in LaTeX sources`.

Online methods: Some online PDF tools like `scanwritr` provide options to sanitize PDF files. The users needs to first upload his/her PDF file on their website and then once PDF is sanitized, user is given an option to download the sanitized PDF file. This process has privacy risks as the websites can collect the hidden data.

Other methods: There are few other methods that is mentioned online to clean PDF metadata, surely these methods remove metadata from the PDF files. Method 1 based on `qpdf`[17] and regular expressions can be used (Listing 3.13). The first few commands are used to clean the content of the PDF file. And then the last command copies the PDF file in an empty document without any metadata.

```
qpdf --qdf --object-streams=disable $file tmp
perl -pe 's/(?<=\/T \()(.*?)(?=\))/ "x" x\
length($file) /e' tmp > tmp.perl
qpdf --compress-streams=y tmp.perl tmp
qpdf --empty --pages tmp 1-z -- anonymous.pdf
```

[16] `https://ctan.org/pkg/pdfprivacy`
[17] `https://gist.github.com/peci1/67bc29310fd4208312222c2de97ba0eb`

```
rm tmp tmp.perl
```

Listing 3.13: PDF sanitization method 1- using `perl` and `qpdf`.

Method 2[18] consists to combine `pdftk` with regular expressions and `exiftool` (Listing 3.14). The first commands are used to purge the metadata included in the document information dictionary. Then, `exiftool` is used to place the metadata into an unused object of the file. The option `-linearize` of `qpdf` has the side effect to remove any unused object and hence removes the metadata object.

```
pdftk $file dump_data | \
sed -e 's/\(InfoValue:\)\s.*/\1\ /g' | \
pdftk $file update_info - output clean-$file

exiftool -all:all= clean-$file
exiftool -all:all clean-$file
exiftool -extractEmbedded -all:all clean-$file
qpdf --linearize clean-$file clean2-$file

pdftk clean2-$file dump_data
exiftool clean2-$file >>tmp1.txt
pdfinfo -meta clean2-$file
```

Listing 3.14: PDF sanitization method 2- removing `PDF` metadata using `pdftk`, `exiftool` and `qpdf`.

Recommendations provided by NSA and many tools presented in this section propose to erase the metadata of a PDF file and it is also possible to find scripts or command lines to sanitize. From a technical and user point of view, the problem of PDF file sanitization seems to be solved: technical issues have been identified and they have been implemented in widely used software. There are no obstacles left to prevent users from sanitizing their PDF files. To check the adoption of PDF sanitization and different kinds of information leaked in PDF files, we performed several experiments as detailed below.

We have seen that different kinds of information is revealed in the PDF files and especially in PDF metadata, now let us see if any authors in our dataset care to sanitize their PDF files. During our analysis of PDF files, we observed that different authors have used different levels of sanitizations. We observed that many authors tend to sanitize their PDF files either removing complete metadata or by removing partial metadata or by removing all the hidden data. Hence, we distinguish four different levels of PDF file sanitization

[18] `https://gist.github.com/hubgit/6078384`

- **Level-0** consists of PDF files that include metadata information. There is no sanitization.
- **Level-1** consists of PDF files with partial metadata. Some metadata fields have been removed.
- **Level-2** consists of PDF files without any metadata. They have been sanitized using `exiftool` or by producing PDF files without any metadata.
- **Level-3** consists of PDF files with no information leakage and properly cleaned. All the objects within the PDF file holding sensitive information have been removed (This level can be obtained using Adobe Acrobat).

When Level-2 and Level-3 sanitization are observed, we know that the authors of the PDF file have a clear will to sanitize their PDF files. Level-0 and Level-1 are observed when the authors have not applied any sanitization method on their PDF files. Solution based on `exiftool` and by producing PDF files without metadata (Level-2) are not as strong as Adobe Acrobat (Level-3) with respect to sanitization. In fact we observed that, if `exiftool` is used to sanitize a PDF file, it is still possible to recover all the metadata. Metadata are stored in a separate object within a PDF file and exiftool only removes the reference to this metadata object in the file. Hence, it is still possible to access this object. Accessing each field of the metadata requires only the use of the grep command. Therefore, Adobe Acrobat should be considered as the only reliable solution for Level-3 PDF file sanitization.

3.4.1 Sanitization followed by security agencies

Using the metadata information we have evaluated each PDF file to check their level of sanitization. Table 3.16 provides the number of PDF files for each level of sanitization in our Security dataset. We found that a total of 9509 PDF files have been sanitized before being published online. Clearly, PDF sanitization is a concern for several security agencies. However, we found that only 3313 PDF files were sanitized with Level-3.

Level of sanitization	# PDF
Level-0	16199 (41%)
Level-1	13956 (35%)
Level-2	6196 (16%)
Level-3	3313 (8%)

Table. 3.16: Different levels of sanitization used on PDF files by security agencies.

We have computed a score for each agency based on the level of sanitization of the PDF files published. This score is the weighted sum of the number of PDF files sanitized with a certain level times the corresponding level. The value n_i is the number of PDF files sanitized with Level-i for $0 \leq i \leq 3$ (see Equation (3.1)). The highest possible score is 3 and it can only be achieved if the agency only publishes PDF files with Level-3 sanitization.

$$\text{Score} = \frac{0 \times n_0 + 1 \times n_1 + 2 \times n_2 + 3 \times n_3}{n}, \tag{3.1}$$

For instance, we have downloaded $n = 82$ PDF files on `nsa.gov`. We found that $n_0 = 45$ (Level-0), $n_1 = 24$ (Level-1), $n_2 = 13$ (Level-2) and $n_3 = 0$ (Level-3). Therefore, NSA has a score of 0.60. Table 3.17 shows the score distribution of the security agencies. One security agency (nabis.police.uk) does not care to sanitize any PDF files before publishing. At the other side of the scale, we found no agency with the perfect score. We found 7 agencies with a score greater or equal to 2: most of their files are sanitized. Four of these agencies `ssi.gouv.fr`, `bmi.bund.de`, `interior.gob.es` and `secp.gov.pk` have performed Level-2 sanitization on most of their PDF files. Three agencies `sie.ro`, `garda.ie` and `bvt.gv.at` have taken care to sanitize most of their PDFs with Level-3 sanitization. Clearly, some security agencies are more concerned by PDF sanitization. Even if they do not sanitize all the PDF files published, they take care to sanitize fewer of them.

Score	0	0 >1	1	1 >2	2	2 >3	3
# agencies	1	50	6	11	4	3	0

Table. 3.17: Sanitization score of security agencies.

Our study shows that the PDF files published by different security agencies are not sanitized to the level expected by such organizations. Many PDF files published by these agencies contained hidden information which can be used to target their employees. Footprinting an organization using its published PDF files is quite effective.

3.4.2 Sanitization followed by Scientific Community

We were surprised with the results of sanitization of PDF files found in IACR and HAL dataset, it includes just one PDF file of Level-2 sanitization in IACR dataset (Table 3.18). This file has been sanitized using `exiftool`. Clearly, researchers do not pay any attention to sanitize their PDF files.

Level of sanitization	# PDF- IACR	# PDF- HAL
Level-0	11389 (99.85%)	540690 (99.3%)
Level-1	16 (0.1%)	3770 (0.6%)
Level-2	1	0
Level-3	0	0

Table. 3.18: Different levels of sanitization used on PDF files in IACR and HAL dataset.

It seems that the authors have the feeling that sanitization is not necessary for them because their identity is already given in the visible content of the PDF file. Authors may think that PDF metadata creates a limited risk, however, they might be unaware that there are still a lot of other information concerning the authoring process that could be exposed (OS, producer tool, software versions...) and pose a risk to the author.

Since sanitization is not popularly practiced by any researchers, we wanted to analyze the way distribution of PDF files takes place in this community. We have performed some analysis on different practices and presented our observation in the following section.

3.5 Fairness of submission process in scientific conference

The results obtained for the dataset of scientific community (IACR and HAL) clearly shows that sanitization is not practiced by researchers. To put the results into perspective, we have examined different procedures followed by scientific community for the distribution of their work using PDF files. We have conducted a study of how the *submission and review system* and *preprint* servers operate.

Recent changes in some submission and review system have shown that, PDF sanitization is considered as a vital process to provide fairness in the reviewing process. We discovered that certain conferences are well aware that the submitted PDF files must be sanitized. We studied 47 computer science conferences that follow the double-blind review policy which are listed on `http://double-blind.org/`. We have examined the call for papers and the submission guidelines to the authors for the anonymous submission for the 2020 edition of these conferences. We searched if they provide explicit instructions concerning PDF metadata. Only seven conferences (*ACM - ASPLOS 2020*, *HPCA 2020*, *ICSE 2020*, *ISCA 2020*, *MICRO 2019*, *SIGMOD 2020* and *CHI 2020*) mention explicitly that the authors must remove all identifiable information from metadata of the PDF file. ICSE conference also provides instructions to check

the metadata using `pdfinfo` or `Adobe Acrobat`. But none of these seven conferences provide any instructions to clean metadata and sanitize the PDF files. Even though the purpose for PDF sanitization is different here, we can see that it is getting enforced at least during the reviewing process.

Authoring process in scientific community
During the analysis of PDF files, we found that three different ways of authoring process (direct, indirect and modified) are usually used to create a PDF file in the scientific community. We say that the authoring process is *direct* when the PDFs have been created by the researcher's computer. Otherwise, the authoring process involves a computer from a third party. In this case, their are two different authoring processes: *indirect* and *modified*. It is *indirect* when a third party uses it's computer to create the PDF file from the source files (doc/docx or LaTeX files for instance) of the researchers. The authoring process is *modified* when the researchers have created the PDF file but it was later modified on a third party's computer. The hidden information in the PDF files for each of these authoring process are different and the level of risk posed are different too. When direct and modified authoring process is used, it is possible to get information on the author. Indirect creation is safe since no author information is leaked in the hidden content of the PDF file.

3.5.1 Role of Submission & Review Systems

In order to be fair, many conferences promote double blind review concept. Here the researchers and reviewers are anonymous. But the metadata and other hidden information present in the PDF file could lead to compromising the fairness of the review process. In this context a reviewer can obtain information on the authors of a submission, this situation was previously discussed in the work done in [13]. We have analyzed several systems in Table 3.19 to determine what type of PDF file is available to the reviewers.
Elsevier Editorial System, Editorial Manager and evise are operated by Elsevier and they all modify the PDF files submitted by the authors to add a header page. ScholarOne Manuscripts is used by IEEE and ACM and the PDF submitted by the authors are modified. Whereas, EasyChair (`https://easychair.org/`) and HotCRP (`https://hotcrp.com/`) directly provides the PDF of the authors to the reviewers. A submission on EasyChair or HotCRP exposes the information of it's authors to the reviewers. This can affect the peer-review process if it is supposed to be double-blinded.

Submission & review systems	PDF type
evise	Modified
Editorial Manager	Modified
Elsevier Editorial System	Modified
ScholarOne Manuscripts	Modified
EasyChair	Direct
HotCRP	Direct

Table. 3.19: PDF types of submission & review systems.

3.5.2 Online Publication of technical documents

All the accepted work during the review process are published by respective conferences and journals. We have analyzed how publishers and preprint servers publish the PDF files. We have sampled PDF files over the last five years from different academic publishers like IEEE, ACM, Elsevier, Springer, open-access journals and online preprints. During this research, we have always respected the agreements between our organization and the various academic publishers. When we have downloaded large number of files, it was always with the authorization to the publishers. We have analyzed their authoring processes to determine the type of the PDF files they publish. The results are given in Table 3.20.

Publication	PDF type
IEEE Conferences	Direct
IEEE Journals	Indirect
ACM Conferences	Direct
ACM Journals	Indirect
Elsevier	Indirect
Springer	Indirect
Open Access Journals	Direct Indirect
Cryptology ePrint Archive	Direct
Open Archive HAL	Direct/Modified/Indirect
arXiv	Word - Direct LaTeX - Indirect

Table. 3.20: PDF publication policy for publishers and preprint.

All the publishers listed in Table 3.20 accept PDF files created by either LaTeX or any Microsoft Word compatible software. And some of them request to access the source files used by the authors to create the original submission file. It is interesting to observe that IEEE and ACM

publish different types of PDF, the proceedings of conferences are direct, while journals are indirect. Elsevier and Springer always request the source files from the authors and then produce the PDF files.

Open Access journals do not share a common PDF type. We have used the Directory of Open Access Journals[19] to sample PDF files and to evaluate each journal. We have evaluated a total of 486 journals. We found that 259 (53 %) used the direct authoring process. Other journals have an indirect PDF publication policy. We found just one open access journal[20] that has sanitized all the PDF files by removing all the metadata information.

Preprint servers are very interesting. The Cryptology ePrint Archive[21] publishes directly the file submitted by the researchers. arXiv[22] has a very original PDF publication policy: researchers using LaTeX must provide their sources to arXiv (indirect policy). Otherwise (Microsoft Word users for instance), the PDFs are directly submitted. This policy is enforced by a detector which checks if a PDF file has been created using LaTeX or other producer tool. The Open Archive HAL[23] has a more complicated submission type. The authors are free to choose to submit their PDF files or their sources (LaTeX or Microsoft Word compatible software). If they choose to submit a PDF file, HAL server is editing the file to add a front page. The types of PDFs found on HAL are modified or indirect. We contacted HAL administrators. They have published a total of 808413 PDF files. Only 65421 PDF files (8%) have been created by HAL (indirect). When the researchers have the choice, they prefer to publish directly their PDF files to avoid losing time needed to comply with the rules enforced by HAL on the submission of source files.

So far we have seen how different conferences, Journals and preprint publish PDF files. Every time direct and modified authoring process is used, there exists a possibility that an adversary can exploit information on the author.

3.6 Targeting an author/organization

Organization footprinting [38] regroups all the techniques used by hackers to collect as much information as possible about their victims. The goal of this reconnaissance is not only to obtain details on the people but also on the infrastructures they use (network, hardware,

[19] https://doaj.org/
[20] https://imt.uoradea.ro/auo.fmte/
[21] https://eprint.iacr.org/
[22] https://arxiv.org/
[23] https://hal.archives-ouvertes.fr/

system...). Footprinting includes techniques like OS fingerprinting [98] or the exploitation of all the documents published by the organization [9, 7] using tools like FOCA[24]. This later technique is particularly attractive for hackers because it is rather inexpensive and effortless. Organizations publish on their websites many Microsoft Office (`doc`, `docx`...) or Portable Document Format (PDF) files. All these file formats are particularly interesting for a hacker because they include hidden data which describe the authoring process. The impact of hidden data was highlighted during two events that occurred during the Iraq War.

In February 2003, the British government of Tony Blair published on its website a dossier on Iraq's security and intelligence organizations. The dossier was a Microsoft Word file[25]. The file was analyzed by the IT researcher Richard M. Smith[26] who retrieved the revision logs. It was easy to identify the authors and their positions in government from these revision logs. The British government was greatly embarrassed by the information exposed by those hidden information.

In 2005, an incident occurred between American soldiers and Italian Secret Service officers near Baghdad International Airport causing the death of an Italian officer. Multi-National Force-Iraq issued a report on its investigation of the shooting. That report was posted as an unclassified PDF file with classified sensitive data obscured from public view. However, it was discovered that copying and pasting the classified sections revealed the blocked text.

Even researchers are not immune to attacks, Professor Jean-Jacques Quisquater, a Belgian cryptographer whose work is said to have informed card payment systems worldwide, has reportedly become the victim of a spear-phishing attack by the NSA and/or GCHQ[27]. Belgium's De Standaaard reports that Professor Quisquater clicked on a fake `LinkedIn invitation` that infected his computer. The malware is said to have allowed tracking of the Professor's work, including consultancy for various firms.

PDF files published/shared by security agencies and researchers contain different information and knowing the information on the target just makes the work of the attacker easier.

[24] `https://github.com/ElevenPaths/FOCA`
[25] Retrieved 02/24/2021 at `http://web.archive.org/web/20040329171413/http://www.computerbytesman.com/privacy/blair.doc`
[26] Retrieved 02/24/2021 on `http://web.archive.org/web/20040113074742/http://www.computerbytesman.com/privacy/blair.htm`
[27] `https://news.hitb.org/content/nsa-gchq-accused-hacking-belgian-smartcard-crypto-guru`

3.6.1 Solving the authorship problem

In order to exploit the metadata information, the adversary needs to first resolve the authorship problem. It is possible that the author of the PDF content is one person and the PDF producer is all together an another person.

We analyzed the content and metadata of PDF files of **security agencies** and found that only 4% PDF files have the author name in the content. These author names do not match the author names in the metadata of the document. Clearly, there is some link between the author and producer of the PDF file. An adversary can still exploit such information to link the author publishing the PDF with the one who created it. Information on the organization level is leaked in this scenarios. It could be used to build profiles and draft malware etc..

When we consider the **scientific community**, the problem of authorship is much more complicated. Many a times scientific papers are written by multiple authors, all the author names are mentioned within the PDF file. Sometimes these authors can be from different organizations and countries. Metadata present in the PDF file will be related to the one producing the PDF file. Here the adversary will need to solve the co-authorship problem and gather data about the targeted author.

We have considered the co-authorship problem and present scenarios where an adversary can target a researcher in the scientific community. Many tools have been created to find scientific publications like Google Scholar or DBLP[28]. An adversary can make a single query to DBLP to obtain almost all the links to the publications of a targeted researcher.

Dealing with co-authorship: Let us assume that the adversary has obtained all the PDF files associated to a targeted researcher named Alice. Unfortunately the adversary can not yet claim that the metadata of the collected PDF files are related to Alice. In several PDF files, Alice has several co-authors (Bob and Charlie). Let us assume that the adversary has obtained 3 PDF files co-authored by Alice (see Table 3.21).

Publication	**Co-author names**			Year
PDF 1	Alice	Bob		2020
PDF 2	Alice	Bob	Charlie	2018
PDF 3	Alice			2018

Table. 3.21: Alice's co-authors.

[28] `https://dblp.org/`

PDF 3 in our example (Table 3.21) is very important for the adversary as the adversary knows that the PDF file has been directly produced by Alice. Therefore, the metadata of this file contains information associated to Alice. For PDF 1 and PDF 2 in Table 3.21, there are some uncertainty on who has created the file: it can be Alice, Bob or Charlie. The adversary can match the metadata provided by PDF 3 with the metadata of PDF 1 to check if they are consistent. If they match, there are two possibilities:

- Alice has created PDF 1.
- Bob has created PDF 1 but Alice and Bob use the same tool to create PDF files. This is a *collision* for the adversary.

If they mismatch, the adversary has two possibilities:

- Bob has created PDF 1.
- Alice has created PDF 1 but Alice is using different tools to create PDF files. This is a *instability* for the adversary.

To remove these ambiguities, the adversary can look at the profiles of Bob and Charlie. The publication of Bob and Charlie can have a impact on the privacy of Alice.

# Author(s)	# occurence	# PDFs
1 author	1035 (11%)	1988 (17%)
2 authors	2608 (27.5%)	3471 (30%)
3 authors	2701 (28.5%)	3072 (27%)
authors ≥ 3	3119 (33%)	2874 (25%)

Table. 3.22: Author statistics on the Cryptology ePrint Archive.

Finding PDF files associated to a single author is quite infrequent in the IACR dataset (See Table 3.22). Only 17% of the PDF files have a single author and it represents 11% of all the authors found in this dataset.

The attribution problem is equivalent to solving a linear system of equations over $\mathbb{F}_2$. The number of equations is given by the number of PDF files and the binary variables are the co-author's names. The Cryptology ePrint Archive consists of 11405 PDFs for 9558 authors and co-authors. We have solved the system of equations created for each author individually and then we have propagated the results whenever it was possible to associate the PDF file with an author. The system of equations for the Cryptology ePrint Archive is underdetermined:

this source by itself is not enough to attribute a PDF per author. We were unable to find conclusive results for 33% of the authors included in this dataset. For the rest of the authors, it is possible to make a guess to determine who has really created a PDF file. But there is 26% of collision and we found no instable author.

Co-authors are a big problem for an adversary who wants to extract information on the real author, *i.e.* who has created the file? We have demonstrated that use of linear algebra can deal with the issue of co-authors.

3.6.2 Who Should Produce The PDF?

Ideally, PDF producer tools should directly sanitize PDF files during the creation. But that is not the case, only some PDF producer tools provide option to sanitize the PDF file and it has to be done by the author following some guidelines. We have encountered 3699 different PDF producer tools during our analysis of PDF files and it is very unlikely that all these tools will implement sanitization. It might not be possible to change all PDF producer tools therefore it is best to change practices of the author producing the PDF files.

3.6.3 possible sanitization methods for organizations

For both security agencies and the scientific community, we have several different scenarios that needs to be handled. Figure 3.5 describes different scenarios that are available. PDF sanitization can be enforced by the authors (**direct sanitization**), a third party (**modified sanitization**) or the authors can fully delegate the creation of the PDF to a third party (**indirect sanitization**).

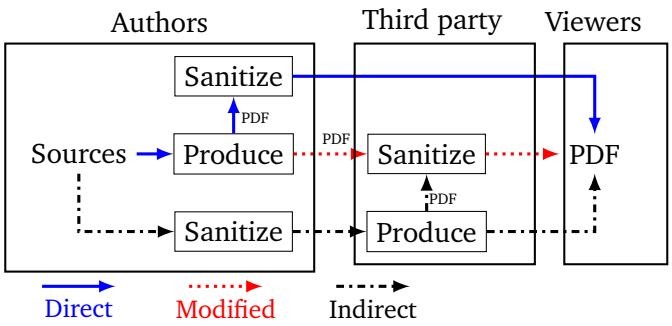

Figure. 3.5: Different possibilities to sanitize a PDF file.

Direct sanitization is based only on the actions of the authors creating a PDF file. Authors produce their PDF and then sanitize it before it is shared. This needs an incentive to ensure that the authors change their publication habits. If every security agencies enforce sanitization in their organization by providing guidelines, direct sanitization could work very well. And for scientific community, if the publication venue chooses a double-blinded peer review process with strong guidelines for sanitization, then we can put the authors in charge of sanitization of their PDF files. When the sanitization is enforced, authors will have stronger incentives than privacy to sanitize there files. However, this is very unlikely to happen.

In the **modified sanitization**, the authors produce their PDF which is then sanitized by a third party. In practice, for security agencies, this third party could be a dedicated set of employees that take care of sanitization or another organization that sanitizes and checks the PDF files before they are distributed or published. For scientific community, this third party could be the submission & review system, a preprint server or a camera-ready version system. Submission & review systems like ScholarOne Manuscripts already verify if a submitted PDF respects the format of the venue (A4 paper or letter etc.). In addition to this verification, they should sanitize by default PDF files. It will be beneficial for the reputation of the publishers who can advertise that they do their best to protect the privacy of researchers.

Indirect sanitization: This includes the source files to be submitted to a third party and the PDF files are produced by the third party and not the authors. The third party takes charge of producing clean PDF files and the author's sensitive information is not present in the PDF file. For the security agencies again it could be an organization that works for them. And for scientific community, the authors can provide their sources to a submission & review systems, a preprint server or a camera-ready version system. This is the case of arXiv for LaTeX users. The only information exposed in the PDF files metadata are related to the tools of the third party. This solution is attractive but it creates issues for both the authors and the system creating the PDF file. First, the authors need to provide their sources. Recently, it was demonstrated [73] that programmers have a tendency to put very sensitive information in their source codes. Authors can expose more information in their sources than in the metadata of their PDF files. Compiling LaTeX file can be risky as shown in [23, 24]. It can threaten the security of submission & review systems. Indirect sanitization seems to create more problems than it solves. However, each issue can be addressed. It is possible to sanitize sources, Mathieu Roy has written a tool `latexpand`[29] that simplifies distribution of LaTeX sources to satisfy the requirement of editors and archival sites (Springer, arXiv...). It produces a single LaTeX file by expanding `\include` and `\input` and also provides an option to remove comments from LaTeX sources. Authors can use `latexpand` to submit sanitized source files to third parties. Many

[29] `https://gitlab.com/latexpand/latexpand`

systems like HAL, arXiv or Overleaf[30] are also compiling LaTeX sources from unknown origin. They all rely on sandboxes to mitigate the security risks.

Indirect and modified sanitization try to emulate a solution in which all the authors create their PDF files with the same software. This is *privacy by uniformity*. Everybody is going to produce PDF files with the same values. Hence the information in metadata becomes useless as it has no entropy. However, these methods of sanitization expose the authors information to the third-parties. Indirect sanitization exposes the author's sources and modified sanitization is an unclean version of the PDF. Direct sanitization is better for the author's privacy with respect to data leakage to third parties. It does not require to expose any information on the authoring process (PDF file or source files) to any third parties before sanitization.

3.7 Preliminary Conclusion

We have seen how organizations distribute their PDF files without sanitizing them. Our experiments on metadata provides a lot of information on the authoring process used to create PDF files. We observed that it is possible to get information on the authors using the metadata present in the PDF files for both direct authoring (Security agencies and Cryptology ePrint Archive) and also for modified authoring (open archive HAL) process.

Our findings on security agency PDF files can be effectively used to find weak links in an organization: employees who are running outdated software. We have also measured the adoption of PDF files sanitization by security agencies. We identified only 7 security agencies which sanitize few of their PDF files before publishing. Unfortunately, we were still able to find sensitive information within 65% of these sanitized PDF files. Some agencies are using weak sanitization techniques. Security agencies need to change their sanitization methods. Sanitization can be done by imposing the sanitization on the employees during the creation of PDF files or before distribution of PDF files. Another method would be to use a well trusted third-party organization dedicated to carry out PDF creation or PDF sanitiztion.

PDF files in scientific community surely contain author names in the content but we have seen that metadata leaks a lot of other information that could be exploited. In **Scientific community**, authors and publishers need to collaborate. Firstly, the source files provided by the authors to a publisher must be sanitized and in exchange, the publisher should create the PDF files and accept to expose only the necessary information required for the authoring process. Changing the policies of all the scientific publishers may be difficult but not impossible.

[30] `https://www.overleaf.com/`

Each scientific community can attempt to change the policy of its main publishers. PDF files created like this are useless to an adversary.

Are all these sanitization methods really efficient to remove all the information from the PDF files? Or have we just shifted the problem somewhere else? We verify the effectiveness of sanitization in the following section.

3.8 Robust PDF Files Forensics Using Coding Style

Identifying how a file has been created is often interesting in security. It can be used by both attackers and defenders. Attackers can exploit this information to tune their attacks and defenders can understand how a malicious file has been created after an incident. In our work, we identify how a PDF file has been created. This problem is important because PDF files are extremely popular: many organizations publish PDF files online and malicious PDF files are commonly used by attackers.

Does it really matter to identify the PDF producer tool?
The answer is YES, a PDF producing tool detector has applications in offensive security and in incident response. In offensive security, it can be used to determine which software is used by an author or by an organization to create and view PDF files. The attackers can find vulnerabilities corresponding to the PDF viewer identified. Many vulnerabilities have been found in the past in PDF viewers, around **1090** vulnerabilities according to `https://cve.mitre.org` by December 2020. The attacker can craft and send malicious PDF files to the organization thanks to the knowledge obtained from PDF files. In incident response, a PDF producing tool detector is valuable to understand how a malicious PDF file has been created. It is a useful step toward an attack attribution. The most simple approach to design a PDF producing tool detector consists to look at the file metadata. By default, PDF producer tools put many information in the field `Creator` and `Producer` of the file's metadata. It is possible to find the name of the producer tool and its version as well as details on the Operating System. Unfortunately, metadata are not a reliable source of information: they can be easily modified using tools like `exiftool`[31] or removed using sanitization tools like Adobe Acrobat.

We have designed a robust PDF producing tool detector based on the coding style of the file. The PDF standard [56] defines the language that is supported by PDF viewers. Developers of PDF producer tool have their own interpretation of this PDF language. Therefore, it is likely

[31] `https://exiftool.org/`

that their coding style is reflected on the output of their PDF producer tool. The coding style elements [58] in PDF files can be used to identify the producer tool.

3.8.1 Ecosystem

Writing directly PDF commands being difficult for a human, there exist several options to create a PDF file using different file formats (Figure 3.6). There is a large ecosystem of PDF creation tools, converters and optimizers which are available locally as well as online. Many editors also propose in their software the option to convert certain file formats into PDF. Most people are accustomed to convert user-friendly (rich text) documents (doc/docx, ppt) into PDF files. These conversions can be straight-forward, for instance, the conversion in word processing applications such as Microsoft Office and LibreOffice which transform the doc/docx files into PDF file. Intermediary conversions maybe needed for some file formats, like for the LaTeXchain `tex` → `dvi` → `ps` → `PDF`. Popular Operating Systems (OS) and even browsers commonly provide support to print content (html pages, images...) into the PDF file. Then there are several online tools which convert different file formats into PDF files.

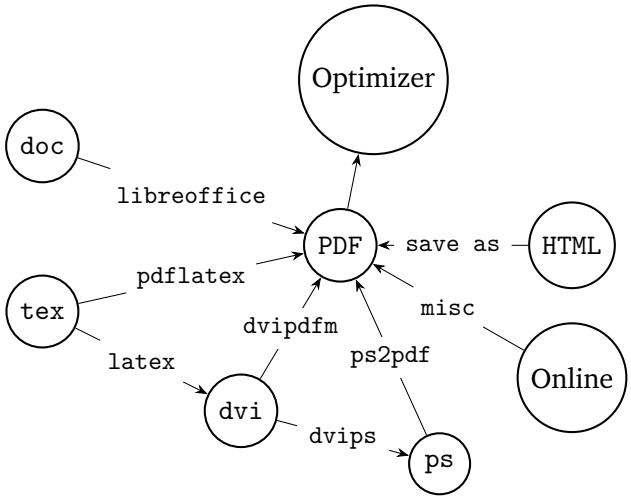

Figure. 3.6: Different options available to create a PDF file.

We broadly classify PDF file creation tools into five categories : OS based tools, word processors, LaTeXprocessors, browsers and optimizer/transformation tools.

OS based tools: In our study we have considered the three prominent Operating Systems and the default PDF creators used by them is shown in Table 3.23. These Operating Systems provide an option to print a file into a PDF using a *print to file* dialog box. While Linux uses

cairo as it's default PDF creator, Microsoft Windows and Mac OS provide two options: either choose any locally installed PDF creator tool or use the default option provided by the OS.

Operating System	Default PDF tools available
Mac OS	10.13.6 Quartz PDFContext
Linux	18.04.1 LTS cairo 1.15.10
Microsoft Windows	10 Pro Microsoft: Print To PDF

Table. 3.23: Operating Systems and their corresponding default PDF producer tools.

Browsers: Many widely used browsers offer the option to print or save any online documents like html files, images etc. into a PDF file. In our study, we have considered the five most popular browsers ranked by W3C[32]: Mozilla Firefox, Google Chrome, Microsoft Edge, Opera and Safari. These browsers either use the local tools present on the system or use their own solution to create a PDF file. Table 3.24 shows the PDF tools used by each of the browser under study. We observed that Google Chrome and Opera use their own software/library to create PDF files across all the three Operating Systems. Safari and Firefox browsers on Microsoft Windows and Mac OS mostly rely on the tool installed on the host Operating System. Firefox uses it's own tool on Linux, it uses cairo 1.9.5 while the system provides cairo 1.15.10.

OS	**Browser**	**PDF Producer**
Microsoft Windows	Mozilla Firefox	Microsoft: Print To PDF*
	Google Chrome	Skia/PDF m73
	Microsoft Edge	Microsoft: Print To PDF*
	Opera	Skia/PDF m71
	Safari	Microsoft: Print To PDF*
Linux	Mozilla Firefox	cairo 1.9.5
	Google Chrome	Skia/PDF m69
	Opera	Skia/PDF m73
Mac OS	Mozilla Firefox	Quartz PDFContext*
	Google Chrome	Skia/PDF m73
	Opera	Skia/PDF m71
	Safari	Quartz PDFContext

Table. 3.24: PDF creation tools for each browsers with respect to the OS (* or local tools available on the OS).

Word Processors: All the applications/tools in the office suites can convert any respective office documents (e.g., doc and docx) into a PDF file. Windows provides the Microsoft Office suite and it could also be used on the Mac OS. Adobe Distiller and LibreOffice could be used

[32]https://www.w3schools.com/browsers/default.asp

on all three Operating Systems. AppleWorks and Pages could be used as word processor tools on Mac OS.

LaTeXprocessors: The LaTeXworld is dominated by two distributions: MiKTeX (`https://miktex.org`) and TeXLive (`https://tug.org/texlive/`). LaTeXis popularly used in scientific community. LaTeXdistributions propose several commands to create a PDF file directly using `pdftex, Luatex and XeTeX`. There is also a possibility to create a dvi file using LaTeXand then convert it to a PDF file (`dvi` $\rightarrow$ `pdf` or `dvi` $\rightarrow$ `ps` $\rightarrow$ `pdf`).

Optmizers/tranformation tools: Two transformations are often applied to PDF files to improve the viewers experience: linearization and compression. Linearization can be applied to improve the access to the document. It is often applied to PDF designed to view online. Compression tools attempt to reduce the size of the `PDF` file by improving embedded fonts, removing unused data or compressing streams. Users can either use tool which could be locally installed or there are many websites which propose to create/optimize PDF files online.

The specifications of the PDF format [56] defines a language supported by PDF viewers. Different syntaxes are supported to describe the same element in a PDF file. We expected that each PDF producer tool has its own way to create the PDF code. To verify our guess, we chose 11 PDF producer tools (Table 3.25) which represents different categories in Figure 3.6. We have used these 11 PDF producer tools on three different Operating Systems: Microsoft Windows (Windows 10) , MAC OS (10.15.7) and Linux (Ubuntu 18.04.4 LTS) Operating Systems whenever it was possible.

We created a dataset that includes 25 source documents for Microsoft Word compatible software and 30 source documents for LaTeX compilation chains. It is important to notice that PdfTeX tool name used in our work is the label of LaTeX software whereas PDFLaTeX is a standard label used by HAL for it's PDF producer tool. We could not create a dataset of PDF files for PDFLaTeX tool. Since we did not want to pollute HAL with our test files, for the analysis of coding style of PDFLaTeX tool we used some random PDF files created by HAL using this tool. These 900 PDF documents include many different elements like text, images, tables, equations etc. to be representative of the usual content found in a PDF file. We then analyzed these PDF files to identify patterns and combination of patterns which create **unique producer signature or fingerprint**.

Producer tools
Acrobat Distiller
Microsoft Office Word
LibreOffice
Ghostscript
Mac OS X Quartz
PdfTeX
SKia/PDF
Cairo
xdviPDFmx
LuaTeX
PDFLaTeX

Table. 3.25: 11 PDF producer tools.

3.8.2 Patterns observed for different PDF producer tools

To identify the software used to create a PDF file, we explore the different structures of the PDF file. Indeed, a PDF file is organized into four parts [87] (see Figure 3.1): header, body, cross reference table and trailer. We described the patterns found in each of these parts separately. These patterns can be used to find the PDF producer tool used to create the PDF file.

Header

The header section of PDF files has always the same organization across all the producer tools. It consists of two comments line: first comment line always starts with the same magic number `%PDF` (Figure 3.1) and is often followed by another comment if the file contained some binary data. This comment in the second line is left undefined by the specification. During our analysis, we observed that it is often there and producer tools leave different values in the file. Listing 3.15 shows an header section of the `Microsoft Office Word` tool. All the PDF files created by this tool include the same binary data as the second comment.

```
1 %PDF-1.7
2 %\B5\B5\B5\B5
```

Listing 3.15: Binary data - Associated to Microsoft Office Word tool.

Table 3.26 shows the different binary data associated to the 11 producer tools. Some values are shared by different tools and some are specific to a distribution and Operating System. For instance, `0xE2E3CFD3` is the binary value associated to the tool `Acrobat Distiller` and it is unique. Our analysis also showed that it is possible that one tool uses several values

across different Operating Systems. In our analysis, LuaTeX uses 2 different values. It depends on the LaTeX distribution and the OS used. `0xD0D4C5D8` value is shared by two PDF producer tools `pdfTeX, LuaTeX`. It is also interesting that this value is specifically associated to Texlive distribution on Linux system only. This value could be used in the detection of the OS used to create the PDF file.

Nine producer tools in our evaluation have unique values and hence this value can be used to directly reveal the PDF producer tool. This value can be considered as *producer magic number*. Removing the *producer magic number* or even altering it does not have any influence on the display of a PDF file. It is not necessarily a robust method to identify a PDF producer tool because it can be easily modified. Still, the *producer magic number* can significantly help in the detection of the PDF producer tool or at least narrow down the identification to a small set of tools. Table 3.26 shows the producer magic number and the associated Operating System, for LaTeXtools, it also shows the distribution.

Unique Binary Data (in hexadecimal)	PDF producer tools	OS- Microsoft Windows, Linux, Mac OS	Distribution- TeXLive/MikTeX
0xE2E3CFD3	Acrobat Distiller	3 OSs	-
0xB5B5B5B5	Microsoft Office Word	Microsoft Windows	-
0xD0D4C5D8	pdfTeX	3 OSs	TeXLive & MikTeX
0xD0D4C5D8	**LuaTeX**	**Linux**	**TexLive**
0xCCD5C1D4C5D8D0C4C6	**LuaTeX**	**Linux**	**MikTeX**
0xCCD5C1D4C5D8D0C4C6	**LuaTeX**	**Mac OS & Microsoft Windows**	**TeXLive & MikTeX**
0xE4F0EDF8	xdvipdfm	3 OSs	TeXLive & MikTeX
0xc7ec8fa2	GhostScript	3 OSs	TeXLive & MikTeX
0xc3a4c3bcc3b6c39f	LibreOffice	Linux	-
0xC4E5F2E5EBA7F3A0D0C4C6	Mac Os X	Mac OS	-
0xB5EDAEFB	cairo	3 OSs	-
0xD3EBE9E1	Skia	3 OSs	-
0xF6E4FCDF	PdfLaTeX	online	-

Table. 3.26: Header- producer magic number and the associated producer tools.

Body

The body part is a collection of indirect objects representing the fonts, pages, sampled images and object streams of the PDF (Figure 3.1). The PDF document contains eight basic types of objects: booleans, numbers, strings, names, arrays, dictionaries, streams and the null object [56]. Each of these objects are described below:

1. **Booleans**: There are two keywords *true* and *false* that are used to represent boolean values.

2. **Numbers**: Integer and real are the two types of number used in PDF document. Both the types of numbers can be preceded by plus/minus sign (Integer values: 123 43445 +17 -98 0 Real values: 34.5 -3.62 +123.6 4. -.002 0.0).

3. **Strings**: A string object shall consist of a series of zero or more bytes and the length of a string may be subject to implementation limits. String objects are written in one of the two ways as **Literal Strings** - as a sequence of literal characters enclosed in parentheses ( ) or **Hexadecimal Strings** - as hexadecimal data enclosed in angle brackets < >.

4. **Names**: The names are represented by a sequence of ASCII characters (range 0x21 - 0x7E). There are some exception with characters like %, (,), <, >, [,], {, }, # and these must be preceded by a slash (/). These characters can also be represented using their hexadecimal equivalent which is preceded by the character "#". The length of the name element may be only 127 bytes long. Table 3.27 show the way name literals are written. A slash symbol must be used to introduce a name, the slash is not part of the name but is a prefix indicating that what follows is a sequence of characters representing the name. For special characters, two digit hexadecimal notations needs to be used.

Syntax for literal name	Resulting name
/Name1	Name1
/A#42	AB
/@pattern	@pattern

Table. 3.27: Examples of name literals.

5. **Arrays**: An array object is a one-dimensional collection of objects arranged sequentially. PDF supoorts one dimensional array only, by nesting and using array within an array one can acheieve higher dimensions. PDF arrays can be heterogeneous and may contain numbers, strings, dictionaries, or any other objects, including other arrays. An array may have zero elements.An array shall be written as a sequence of objects enclosed in SQUARE BRACKETS (Example: [549 3.14 false (string) /Something]).

6. **Dictionaries**: A dictionary object is an associative table containing pairs of objects. The first element of each entry is the key and the second element is the value. The key must be the name object, whereas the value can be any object, including another dictionary. The number of entries in a dictionary shall be subject to an implementation limit. A dictionary can be presented with key-value pairs enclosed in double angle brackets (« ... »). Listing 3.16 shows an example of a dictionary object.

```
<<  /Type  /Example
    /Subtype  /DictionaryExample
    /Subdictionary  << /Item1  0 . 4
                       /Item2  true
                    >>
>>
```

Listing 3.16: Examples of Dictionary object.

7. **Streams**: A stream object is represented by a sequence of bytes and the length can be unlimited. Due to this feature, images and other big data blocks in PDF are are represented as streams. A stream object usually is represented by a dictionary object followed by the keywords *stream* followed by newline and finally *endstream* denoting the end of this object. Listing 3.17 shows how a stream object is represented.

```
dictionary
stream
... Zero or more bytes ...
endstream
```

Listing 3.17: Examples of stream object.

8. **Null**: *null* is used to represent null object. Only one object of type null can be present in a PDF, denoted by the keyword null.

Any object in a PDF file can be labeled as an **Indirect objects**. Every object has it's unique object identifier by which other objects can refer to it. Object numbers may be assigned in any arbitrary order. The object identifier shall consist of two parts, a positive integer *object number or object ID* and a non-negative integer *generation number*. The combination of an object number and a generation number is used to uniquely identify an indirect object. In a newly created PDF file all the generation numbers are 0, this number can be updated when PDF is altered. We observed the following differences between the different producer tools:

- Number and type of keys used to describe objects;
- Use of escape sequences (\n (newline), \r (carriage return) and \t (tab) etc..) in literal strings;
- Total number of objects created;
- Arrangement of the objects (random, increasing order, incremental etc..);
- The way metadata information is stored;
- Length of the encoded text and images;

- The way font encoding information is stored.

Listing 3.18 and 3.19 shows example of the same stream object created by `pdfTeX` and `LuaTeX` tools. The way they are encoded is different but PDF viewers will display the same output. Even if the object ID and all the keys (Length, Filter, FlateDecode...) used look alike for these two tools, it is possible to distinguish them using the order of arrangement of keys and the use of escape sequences (\n (newline) and spaces). Such differences across different objects encoded in the PDF files can be used in the detection of producer tools.

```
4 0 obj
<</Length 2413      /Filter/FlateDecode
   >>
stream
.....
endstream
endobj
```

Listing 3.18: Object encoding using pdfTeX tool

```
4 0 obj
<<
/Length 2006
/Filter /FlateDecode
>>
stream
.....
endstream
endobj
```

Listing 3.19: Object encoding using LuaTeX tool.

Cross reference table

Cross reference table or the xref table gives the offsets (in bytes) for each indirect object which is used for quick and random access to objects in the body section (Figure 3.1). This section of a PDF file is optional and many producer tools do not include it. Some producer tools use linearization or incremental saves and the information related to this table is encoded in the trailer object. Cross reference table is always coded in the same way across the different tools, when it is present.

Listing 3.20 gives an example of an xref table generated by `Microsoft Office Word` tool. The cross reference table starts with the keyword `xref` followed by two numbers separated by a space. The first number indicates the object number of the first object in the subsection. The second number indicates the total number of entries in the subsection. Listing 3.20 shows an example of the cross reference table extracted from a PDF file created using `Microsoft Office Word` tool. The table consists of single subsection entry with 5 objects from 0 to 5, where 0 is the first object and 5 is the number of entries in the subsection of the cross reference table. 0 5 also indicates that there are 5 consecutive objects.

Each cross reference entries (one per line) is associated to exactly one object and it is 20 bytes long and has the format `"nnnnnnnnnn ggggg n/f eol"`, where the first 10 bytes are `nnnnnnnnnn`, indicating the byte offset of the referenced entry, followed by a space and then by 5 digits entry `ggggg`, which is a generation number of the object followed by a space and then a `n`, where `n` is a literal keyword to indicate that the object is in use while `f` is used to indicate that an object is free and this is followed by a space and last 2 bytes constituting the `end-of-line`.

```
1 xref
2 0 5
3 0000000010 65535 f
4 0000000017 00000 n
5 0000000166 00000 n
6 0000000222 00000 n
7 0000000486 00000 n
```

Listing 3.20: Cross reference Table - Microsoft Office Word

Since this section of PDF file has same patterns across all the producer tools, it's presence or absence can narrow down the detection of PDF producer tools to a smaller set of candidate tools. We found only nine tools among 11 that include the cross reference table: Acrobat Distiller and xdviPDFmx do not include this table. We observed that `pdfTeX` tool includes the table only in `MikTeX` distribution for all three OSs whereas the table has been removed for `TexLive` distributions.

Trailer

The trailer part is used for quick access to find the cross reference table and certain special objects in the document. The trailer part is very interesting to detect the producer tool used. The last object present in this part includes `Root` information and some other keys-values. It is

possible to distinguish producer tools based on the keys used to describe the trailer object. We have chosen two PDF files with same content created using `LibreOffice` and `Microsoft Office Word` tools, Listing 3.21 and 3.22 shows example of the trailer objects.

Both these tools have completely different coding style for the same content of the file and hence can lead to the detection of producer tool. It is interesting to note that `Microsoft Office Word` tool includes 2 trailer objects which is unique style associated to only this tool.

```
1 trailer
2 <</Size 14/Root 12 0 R
3 /Info 13 0 R
4 /ID [ <438A4EF8B552AF586C55DFFE40065998><438A4EF8B552AF586C55DFFE40065998>
    ]
5 /DocChecksum /7C2B6DC7F4AF6CC658C0703D8002E3D4
6 >>
```

Listing 3.21: Trailer object- Libreoffice

```
1 trailer
2 <</Size 25/Root 1 0 R/Info 9 0 R/ID[<70265267FB5C68469F73B4AB7F5E4003
    ><70265267FB5C68469F73B4AB7F5E4003>] >>
3 startxref
4 46566
5 %%EOF
6 xref
7 0 0
8 trailer
9 <</Size 25/Root 1 0 R/Info 9 0 R/ID[<70265267FB5C68469F73B4AB7F5E4003
    ><70265267FB5C68469F73B4AB7F5E4003>] /Prev 46566/XRefStm 46274>>
```

Listing 3.22: Trailer object- Microsoft Office Word

We have listed all the different keys in the trailer object in Table 3.28. *Tools like LibreOffice, Acrobat Distiller and Microsoft Office Word have their own keys that are distinguishable from any other tools and hence are unique and tool specific.* `LuaTeX, pdfTeX, Ghostscript` and `Mac OS X Quartz` share the same set of keys, but the order of arrangement of these keys is different and hence potentially lead to the detection of a producer tool.

Producer Tool	Key strings in Trailer section
Acrobat Distiller	/DecodeParms /Columns /Predictor /Filter /FlateDecode /ID /Info /Length /Root /Size /Type /XRef /W
TexLive LuaTeX	/Type /XRef /Index /Size /W /Root /Info /ID /Length /Filter /FlateDecode
TexLive pdfTeX	/Type /XRef /Index /Size /W /Root /Info /ID /Length /Filter /FlateDecode
MikTeX LuaTeX	trailer /Size /Root /Info /ID
MikTeX pdfTeX	trailer /Size /Root /Info /ID
Ghostscript	trailer /Size /Root /Info /ID
xdvipdfm	/Type /XRef /Root /Info /ID /Size /W /Filter /FlateDecode /Length
Microsoft Office Word	trailer /Size /Root /Info /ID /Prev /XRefStm
LibreOffice	trailer /Size /Root /Info /ID /DocChecksum
Mac OS X Quartz	trailer /Size /Root /Info /ID
cairo	trailer /Size /Root /Info
Skia/PDF	trailer /Size /Root /Info
PDFLaTeX	trailer /Root /info /ID /Size

Table. 3.28: Trailer - different trailer keys used by PDF producer tools (all the keys are same across 3 operating systems (Microsoft Windows, Linux and Mac OS X)).

PDF section	**YARA rule for MicroSoft Office Word**
Body	*string:* *$rule= /4 0 obj\r\n<<\/Filter\/FlateDecode\/Length [0-9]*>>\r\nstream\r\n/* *condition:* *$rule1*

Table. 3.29: Example of one YARA rule for an object present in the body part of a PDF file created using Microsoft Office Word tool.

It is important to notice that we have excluded the elements containing the metadata in the creation of our rules. Metadata information is stored in a object within the body section of the file. Removing the metadata object or even altering the values present in it has no effect on the functioning of PDF files.

We used the different patterns observed for each section of the PDF file and exploited them to detect the PDF producer tools. We have used regular expressions and expressed them using YARA[33] rules. Table 3.29 shows an example of a YARA rule written to match one of the text pattern generated by the `Microsoft Office Word` tool. We observed that, this object is present in every PDF file created using `Microsoft Office Word` tool.

We have seen earlier in this section that, for the same content of PDF files, number of objects and the way objects are created differs. Since the coding style is different across tools, the

[33] `https://github.com/virustotal/yara`

Producer tool	# rules
Acrobat Distiller	13
Microsoft Office Word	16
LibreOffice	15
Ghostscript	15
Mac OS X Quartz	30
PdfTeX	31
SKia/PDF	12
Cairo	16
xdviPDFmx	13
LuaTeX	22
PDFLaTeX	9

Table. 3.30: Rules for PDF producer tools.

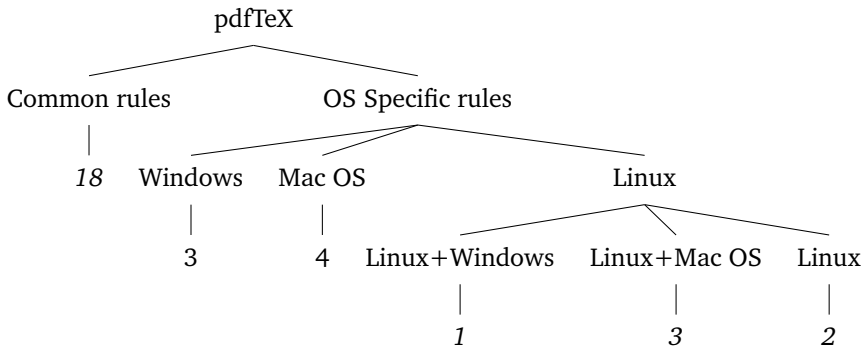

Figure. 3.7: Number of YARA rules across three Operating Systems for *pdfTeX* tool.

number of YARA rules are also different across tools. Table 3.30 shows the number of rules we have written for each producer tool.

In the header section, we have already provided an example that helps in the detection of OS and LaTeX distribution used (LuaTeX tool). During our analysis of 11 producer tools, for some LaTeX chain of tools, we observed that few of our rules can detect the OS. Figure 3.7 provides an example of the tool `pdfTeX` and number of rules associated across different Operating Systems. In the Figure 3.7, we can observe that `pdfTeX` has 18 common rules for three OSs and some rules are specific to one or two OSs. We used these OS specific rules in the detection of the OS used to create the PDF files.

Producer tool	Security	IACR	HAL
Ghostscript	1000(3%)	1200 (11%)	115948 (21%)
Acrobat Distiller	8859 (25%)	369 (3%)	96157 (18%)
Microsoft Office Word	4598 (13%)	147(1.3%)	15995 (2%)
Mac OS X Quartz	111 (0.3%)	62 (0.5%)	7767 (1.4%)
LibreOffice	2171 (6%)	4	451 (0.08%)
SKia/PDF	106 (0.3%)	0	88 (0.01%)
LuaTeX	0	17	67 (0.01%)
PDFLaTeX	0	1	249147 (46%)
xdviPDFmx	0	1347 (12%)	836 (0.2%)
Cairo	3	0	1100 (0.2%)
PdfTeX	2	7745 (68%)	10375 (2%)
Rare tools	18218 (52%)	513 (4%)	46529 (9%)

Table. 3.31: PDF producer tools and number of PDF files associated to each tool in our dataset.

3.8.3 Detection of PDF producer tools

We have tested our rules on the PDF files of three dataset: Security, IACR and HAL. We applied our rules to detect the PDF producer tool and then the results obtained using our tool are validated using the PDF producer tool name found in the metadata field producer. *We would like to clarify that, during the PDF producer tool detection, we have not considered the metadata object present in the PDF file. Our rules are used on the rest of the PDF file.*

cleaning the dataset: Many PDF files downloaded from security agency websites are sanitized. We can apply our tool to detect PDF producer tool on all the 39664 files but we do not know the **ground truth**. Therefore, we eliminate all the sanitized files. For 3 dataset, security, IACR and HAl we also eliminate all the files that were not produced using the 11 tools we have considered. Based on the values found in metadata field producer, Security dataset includes 16850 (48%) PDF files from the 11 tools we are inspecting and similarly 10892 (96%) PDF files from IACR dataset and 497944 (91%) from HAL dataset. We applied our rules on these sets of PDF files and findings are described below. Table 3.31 provides the total number of PDF files associated to each tool that we have examined. It also shows PDF files created using rare producer tools, these rare tools are out of scope of our work.

We first attempted to use the patterns found in each section separately and the results for each section of the PDF file are given in Table 3.32. It shows when the producer detected by our tool is correct, wrong or when it is unable to detect a producer tool. In Table 3.32 the detection of the producer tool based only on the header section of the PDF file has a higher accuracy of 94% for Security dataset. The results of section header and xref is not efficient for

IACR dataset but they are improved for Security and HAL dataset. Body and trailer sections offer better perspective for all the three dataset. However, these results are not conclusive.

Security (16850 PDF files)

Detection	Header	Body	Xref Table	Trailer
Correct	15909 (94.4%)	10387 (61.5%)	5038 (30%)	10372 (61.5%)
Wrong	210 (1%)	801 (5%)	1045 (6%)	318 (2%)
No result	731 (4%)	5662 (33.5%)	10767 (64%)	6160 (36.5%)

IACR (10892 PDF files)

Detection	Header	Body	Xref Table	Trailer
Correct	1676 (15%)	7302 (67%)	898 (8%)	8641 (79%)
Wrong	8325 (76%)	1839 (17%)	1277 (12%)	141 (1%)
No result	891 (8%)	1751 (16%)	8717 (80%)	2110 (19%)

HAL (497944 PDF files)

Detection	Header	Body	Xref Table	Trailer
Correct	235859 (47%)	197069 (40%)	232540 (47%)	229060 (46%)
Wrong	204214(41%)	241101 (48%)	232104 (47%)	259988 (52%)
No result	57871 (12%)	59774 (12%)	33300 (7%)	8896 (2%)

Table. 3.32: Detection of PDF producer tools for header, body, xref and trailer section.

For each individual section, our tool can sometimes detect two producer tools. This case is considered as a wrong prediction in Table 3.32. We have evaluated the frequency of this event in Table 3.33. Two cases are possible. The detection can be *confused*: for the same PDF file, our tool detects the correct producer and an another one as *incorrect*: we have an error when the two tools detected are incorrect.

The results in Table 3.33 show that the detection based on single section of the PDF file has too much uncertainty. For instance the header section for IACR dataset resulted in 66% of confused detection, since pdfTeX and LuaTeX use same producer magic number. Our tool often detects 2 tools for PDF files produced by either pdfTeX and LuaTeX. To improve our tool, we have combined the results of each section using a majority vote. In case of equality, *i.e.* two producers have received two votes, we have considered that our tool takes a wrong decision.

Table 3.34 shows results for the detection of producer tool for combination of all the sections using majority votes. Our tool finds the correct PDF producer tool 74% of the time for both Security and IACR dataset and 48% for HAL dataset. PDF files in HAL dataset are modified using PDFLaTeX tool. These modified PDF files includes coding style of both PDFLaTeX and the original tool initially used to create the PDF file. Currently our tool does not apply to the detection of modified/concatenated PDF files and hence the results obtained for HAL dataset are less impressive.

Security (16850 PDF files)

Detection	Header	Body	Xref Table	Trailer
Confused	116 (0.6%)	801 (5%)	1045 (6%)	250 (1%)
Error	94 (0.5%)	0	0	68 (0.4%)

IACR (10892 PDF files)

Detection	Header	Body	Xref Table	Trailer
Confused	7213 (66%)	433 (4%)	1228 (11%)	130 (1%)
Error	1112 (10%)	1406 (13%)	49 (0.4%)	11 (0.1%)

HAL (497944 PDF files)

Detection	Header	Body	Xref Table	Trailer
Confused	19673 (4%)	44751 (9%)	16028 (3%)	153 (0.03%)
Error	184541 (37%)	196350 (39%)	216076 (44%)	259835 (52%)

Table. 3.33: Frequency of detection of 2 PDF producer tools.

Detection	**Security** (16850)	**IACR** (10892)	**HAL** (497944)
producer tool	12411 (74%)	8018 (74%)	239967 (48%)
OS	5371 (32%)	3344 (29%)	105930 (19%)

Table. 3.34: Detection of PDF producer tool and OS using combination of different sections.

OS Detection: Along with detection of the PDF producer tool used, coding style can also reveal the OS for some PDF producer tools. Table 3.34 shows corresponding results obtained for the detection of OS. Header, trailer and xref section's coding style does not vary much across different Operating Systems and hence predicting the OS using these sections is harder and sometimes impossible. But using the body section's coding style it is possible to detect the OS for some producer tools. During the analysis of coding style we observed that body section coding style includes one or two objects that can be used to detect the Operating System. For instance, pdfTeX (MikTeX distribution) uses some keys across Microsoft Windows, Linux and Mac OS that are distinguishable. Even though we could identify OS for fewer number of PDF files (Table 3.34), these results are not conclusive for the majority of the files in our dataset.

Results for each producer tool

Our tool is more efficient for the detection of certain producers. Table 3.35 shows the differences between the different tools evaluated. The detection is still done using a majority vote over all the parts of the PDF file. PDF files produced by `Microsoft Office Word`, `Mac OS X Quartz`, `PDFLaTeX`, `Ghostscript`, `Skia/PDF` and `LibreOffice` are detected with the higher accuracy (more than 90%).

Producer tool	# PDF Security	# detection	# PDF IACR	# detection	# PDF HAL	# detection
Acrobat Distiller	8859	6245 (70%)	369	269 (73%)	96157	3497 (4%)
Microsoft Office Word	4598	2873 (62%)	147	141 (96%)	15995	4388 (28%)
LibreOffice	2171	2155 (99%)	4	4 (100%)	464	451 (95%)
Ghostscript	1000	993 (99%)	1200	993 (83%)	115948	1940 (2%)
Mac OS X Quartz	111	47 (42%)	62	61 (98%)	7767	2111 (27%)
SKia/PDF	106	96 (91%)	0	0	88	67 (76%)
Cairo	3	2 (67%)	0	0	1100	357 (32%)
PdfTeX	2	0	7745	6001 (77%)	10375	743 (7%)
xdviPDFmx	0	0	1347	536 (40%)	836	102 (12%)
LuaTeX	0	0	17	12 (71%)	67	4 (6%)
PDFLaTeX	0	0	1	1 (100%)	249147	226281 (91%)

Table. 3.35: Detection of individual PDF producer tool.

As previously explained the PDF files in HAL dataset are modified, some parts of the PDF files are replaced by the coding style of PDFLaTeX tool. Since we use the majority votes, the results for tools other than PDFLaTeX (91%) lead to wrong detection. If the PDF file is not modified, our rules detects the PDF producer tool with higher accuracy and results obtained for detection of producer tools in Security and IACR dataset supports it.

Challenges for detection of individual section of the PDF: For the *header* section we used the `producer magic number` to detect the PDF producer tool. We faced several challenges like the producer magic number shared between tools, altered/removed PDF files. Few PDF files with more than one producer magic number due to concatenation of PDF files produced using different producer tools. Challenges faced for detection of *body*, *xref* and *trailer* section are due to concatenation PDF files (produced using tool different producer tools).

The results obtained in this section show that it is possible to detect the PDF producer tool of a PDF file using the coding style. Therefore, creating PDF files without metadata is not enough to hide this information.

3.8.4 Application of coding style rules to detect other PDF files

We have also applied our rules on the sanitized files downloaded from security agency websites and PDF files created using online PDF producer tools of several websites. It is important to note that compared to the previous results, we do not know how the PDF files were actually created.

Sanitized PDF files of Security dataset: We have applied our rules on both Level-2 and Level-3 sanitized PDf files. The results presented in Table 3.36 cannot be verified with the ground truth. Still, our tool has detected many times `Acrobat Distiller` and `Microsoft Office Word`. We got significant detection rate for Level-2 sanitized PDF files (62% and 35%). The results obtained for Level-3 sanitization need to be interpreted differently. We are not detecting the PDF producer but the software used to sanitize the file. Therefore, it is not surprising to find that we only detect Acrobat Distiller which is the recommendation of the NSA [106, 80] for PDF file sanitization.

Producer tool	# Level-2 (exiftool)	# Level-2 (without metadata)	# Level-3
Acrobat Distiller	384 (31%)	952 (19%)	893 (27%)
Microsoft Office Word	215 (17%)	684 (14%)	33 (1%)
LibreOffice	19 (1%)	20 (0.4%)	0
Ghostscript	34 (3%)	34 (0.6%)	0
Mac OS X Quartz	17 (1%)	25 (0.5%)	0
PdfTeX	100 (8%)	2 (0.04%)	0
SKia/PDF	0	5 (0.1%)	0
Cairo	0	0	0
xdviPDFmx	0	0	0
LuaTeX	0	0	0
PDFLaTeX	0	0	0
Unknown	468 (39%)	3237 (65%)	2387 (72%)
# PDFs Detected	769 (62%)	1722 (35%)	925 (28%)
# PDFs	1237	4959	3313

Table. 3.36: Detection of PDF producer tools on sanitized files.

PDF files created using online tools: There are many websites which propose users to create or optimize `PDF` files online. These websites support conversion of documents from one file format in to another, like `doc/docx` to `PDF`. Along with conversion of documents, most of these websites also support `PDF` optimization. We have created PDF files using these online PDF tools (29 compressors and 22 producer tools) that are freely available. Then, we have applied our detection tool on the obtained PDF files and we have compared the producer detected to the value of the producer in the metadata. Table 3.37 and 3.38 shows the list of online tools we have evaluated and the results we have observed (Table 3.39 and Table 3.40 shows results for each section of the PDF file using our detection tool). We found some inconsistencies for 6 PDF compressor tools (in red in Table 3.38). A producer is advertised in the metadata but it is actually anthoer software that has been used. It should noted that some of the online tools advertise that they use their own PDF producer software but it is actually a generic software that has been used. It means that most of the time a user can install locally the PDF

creation tool instead of uploading his/her sensitive PDF files on an online service which has questionable trust.

Online tool	Producer name in Metadata	Tool detected
pdf.io	LibreOffice 6.0	LibreOffice
Hipdf	Microsoft Word 2013	Microsoft Office Word
PDFyeah	LibreOffice 6.1	LibreOffice
google docs	Skia/PDF m76	qpdf
pdfconvertonline	LibreOffice 5.4	LibreOffice
toPDF	LibreOffice 4.2	LibreOffice
small pdf	Microsoft Word for Office 365	Microsoft Office Word
jinapdf	Microsoft Word 2016	Microsoft Office Word
PDF24 Tools	LibreOffice 6.1	LibreOffice
pdf2go	LibreOffice 6.2	LibreOffice
lightpdf	lightpdf.com...	No match
altocompress	None	Microsoft Office Word
pdfaid.com	pdfaid using ABCpdf...	No match
doc2pub	Neevia PDFcompress...	No match
VeryPDF	http://www.verypdf.com	No match
Sejda	Apache FOP Version 2.3	No match
online2pdf	Online2PDF.com	No match
I love pdf	www.ilovepdf.com	No match
Free PDF Editor	FreePDF.net PDFill	No match
PDFcandy	PDF Candy	No match
ZonePDF	zonepdf.com	No match
sodapdfonline	Soda PDF Online	No match

Table. 3.37: Analysis of online PDF creator tools.

3.8.5 Observation

To the best of our knowledge, detection of PDF producer tool using their coding style is never done before, we are the first to work on this topic. Although we found *arXiv* tool that uses metadata and fonts present in the PDF file to detect the PDF producer tool.

e-Print archive arXiv hosts around 1.6 million e-prints in different science fields. All the submissions are controlled to check that the material provided is appropriate, topical and meets arXiv's guidelines. LaTeX, AMSLaTeX, PDFLaTeX sources, PDF and HTML with JPEG/PNG/GIF images formats are accepted by arXiv. ArXiv accepts only PDF files that are produced by Microsoft Word compatible software. If a PDF file created using LaTeX is submitted to arXiv, it is rejected. The authors must submit their sources and ArXiv will produce the PDF file. ArXiv has a tool to detect (Figure 3.8) if a PDF has been produced using Word or LaTeX sources. *As a*

Online tool	Producer name in Metadata	Tool detected
altocompress	PDFfiller	Ghostscript
compress/zipfile	Same as original file	pdfTeX
VeryPDF	VeryPDF	Ghostscript
pdfcompressor	3-Heights(TM)...	Acrobat Distiller
small pdf	3-Heights(TM)	Acrobat Distiller
jinapdf	GPL Ghostscript 9.26	Ghostscript
PDFill	GPL Ghostscript 9.23	Ghostscript
pdfzipper	GPL Ghostscript 9.21	Ghostscript
pdf2go	GPL Ghostscript 9.26	Ghostscript
PDFcandy	GPL Ghostscript 9.10	Ghostscript
p2w compresspdf	GPL Ghostscript 9.22	Ghostscript
PDFyeah	GPL Ghostscript 9.26	Ghostscript
youcompress	GPL Ghostscript 9.26	Ghostscript
wecompress	Same as original file	Acrobat Distiller
pdfconvertonline	Aspose.Pdf for .NET 17.1.0	No match
lightpdf	lightpdf.com	No match
pdfaid	pdfaid using ABCpdf	No match
PDF resizer	itext-paulo-155...	No match
Neevia PDFcompress	Neevia PDFcompress...	No match
Sejda	SAMBox 1.1.50...	No match
PDF-online.com	3-Heights(TM)...	No match
online2pdf	Online2PDF.com	No match
I love pdf	www.ilovepdf.com	No match
ZonePDF	Aspose.PDF...	No match
PDF24 Tools	GPL Ghostscript 9.26	No match
image resize	GPL Ghostscript 9.23	No match
foxit	Same as original file	No match
sodapdfonline	Same as original file	No match
Hipdf	Same as original file	No match

Table. 3.38: Analysis of online PDF compressor tools.

Online Tool	Producer	Header	Trailer	Xref	Body
pdf.io	LibreOffice 6.0	LibreOffice	LibreOffice	LibreOffice	LibreOffice
altocompress	None	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word
pdfconvertonline	LibreOffice 5.4	LibreOffice	LibreOffice	LibreOffice	LibreOffice
lightpdf	lightpdf.com...	No match	No match	No match	Luatex/pdfTeX (image rule*)
pdfaid.com	pdfaid using ABCpdf...	No match	No match	No match	No match
doc2pub	Neevia PDFcompress...	Acrobat Distiller	No match	No match	Microsoft Office Word
VeryPDF	http://www.verypdf.com	Acrobat Distiller	No match	No match	LibreOffice
Sejda	Apache FOP Version 2.3	No match	No match	No match	Mac Quartz/pdfTeX (image rules*)
toPDF	LibreOffice 4.2	LibreOffice	LibreOffice	LibreOffice	LibreOffice
small pdf	Microsoft Word for Office 365	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word
online2pdf	Online2PDF.com	No match	No match	No match	pdfTeX (image rules*)
I love pdf	www.ilovepdf.com	Acrobat Distiller	No match	No match	pdfTeX (image rules*)
jinapdf	Microsoft Word 2016	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word
Free PDF Editor	FreePDF.net PDFill	Acrobat Distiller	No match	No match	pdfTeX (image rules*)
pdf2go	LibreOffice 6.2	LibreOffice	LibreOffice	LibreOffice	LibreOffice
PDFcandy	PDF Candy	Acrobat Distiller	No match	No match	No match
PDF24 Tools	LibreOffice 6.1	LibreOffice	LibreOffice	LibreOffice	LibreOffice
ZonePDF	zonepdf.com	No match	No match	No match	pdfTeX (image rules*)
sodapdfonline	Soda PDF Online	No match	No match	No match	No match
Hipdf	Microsoft Word 2013	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word	Microsoft Office Word
PDFyeah	LibreOffice 6.1	LibreOffice	LibreOffice	LibreOffice	LibreOffice
google docs	Skia/PDF m76	OverLeaf	OverLeaf	No match	OverLeaf

Table. 3.39: List of onnline PDF creator tools (* image encoding style used by the producer tools).

result, the majority of the PDF available on arXiv are produced by arXiv. This is why we have not created a dataset based on arXiv files.

We were curious to understand how the PDF detection tool of arXiv works, in order to know how close it is to our work. There are many threads online that explain how to bypass arXiv detector. We provide several instructive techniques used to *bypass arXiv PDF detection tool*.

It was possible until 21st April 2010 to fool the detector by creating two PDF files[34]. The first PDF file `F1.pdf` is created using Microsoft Word. It is an empty page. The second file `F2.pdf` is created using LaTeX and it contains the content of the paper. The malicious authors append the pages of `F2.pdf` into `F1.pdf`. Then, the authors delete the first empty page. These modification are done using an PDF modification software.

Another technique[35] used that worked until Nov 2019 is to convert the PDF file in to a Postscript file with `pdf2ps` and then convert it back again with `ps2pdf`.

It was not possible to find a thread still able to fool arXiv detector. We tried a guess and determine method to understand how arXiv PDF detection tool works and how it can be bypassed. We made four attempts to transform a PDF file created using LaTeX and then submit

[34] `http://www.hrstc.org/node/62`

[35] `https://tex.stackexchange.com/questions/186068/how-to-upload-latex-generated-pdf-paper-to-arxiv-without-latex-sources`

Tools	Producer	Header	Trailer	XREF	Body
altocompress	PDFfiller	Ghostscript	Ghostscript	Ghostscript	Ghostscript
pdfconvertonline	Aspose.Pdf for .NET 17.1.0	No match	No match	No match	LuaTeX/pdfTeX
lightpdf	lightpdf.com	No match	No match	No match	LuaTeX/pdfTeX
pdfaid	pdfaid using ABCpdf	No match	Acrobat Distiller	No match	Microsoft Office Word/pdfTeX
PDF resizer	itext-paulo-155...	Acrobat Distiller	No match	No match	pdfTeX
Neevia PDFcompress	Neevia PDFcompress...	Acrobat Distiller	No match	No match	Microsoft Office Word
VeryPDF	VeryPDF	Ghostscript	Ghostscript	Ghostscript	Ghostscript
Sejda	SAMBox 1.1.50...	No match	No match	No match	pdfTeX
pdfcompressor	3-Heights(TM)...	Acrobat Distiller	Acrobat Distiller	No match	Microsoft Office Word/pdfTeX
PDF-online.com	3-Heights(TM)...	Acrobat Distiller	No match	No match	pdfTeX
small pdf	3-Heights(TM)	Acrobat Distiller	Acrobat Distiller	No match	Microsoft Office Word/pdfTeX
online2pdf	Online2PDF.com	No match	No match	No match	pdfTeX
I love pdf	www.ilovepdf.com	Acrobat Distiller	No match	No match	pdfTeX
ZonePDF	Aspose.PDF...	No match	No match	No match	LuaTeX/pdfTeX
jinapdf	GPL Ghostscript 9.26	Ghostscript	Ghostscript	No match	Ghostscript
PDFill	GPL Ghostscript 9.23	Ghostscript	Ghostscript	Ghostscript	Ghostscript
pdfzipper	GPL Ghostscript 9.21	Ghostscript	Ghostscript	Ghostscript	Ghostscript
pdf2go	GPL Ghostscript 9.26	Ghostscript	Ghostscript	Ghostscript	Ghostscript
PDFcandy	GPL Ghostscript 9.10	Ghostscript	Ghostscript	Ghostscript	Ghostscript
PDF24 Tools	GPL Ghostscript 9.26	OverLeaf	No match	No match	Overleaf/pdfTeX/ LuaTeX
image resize	GPL Ghostscript 9.23	Acrobat Distiller	No match	No match	pdfTeX
p2w compresspdf	GPL Ghostscript 9.22	Ghostscript	No match	Ghostscript	No match
PDFyeah	GPL Ghostscript 9.26	Ghostscript	Ghostscript	Ghostscript	Ghostscript
youcompress	GPL Ghostscript 9.26	Ghostscript	Ghostscript	Ghostscript	Ghostscript
wecompress	Same as original file	Acrobat Distiller	No match	No match	Acrobat Distiller
foxit	Same as original file	No match	No match	No match	No match
compress/zipfile	Same as original file	pdfTeX/LuaTeX	pdfTeX	pdfTeX	pdfTeX
sodapdfonline	Same as original file	No match	No match	No match	No match
Hipdf	Same as original file	No match	No match	No match	pdfTeX/Mac Quartz

Table. 3.40: List of online PDF compressor tools.

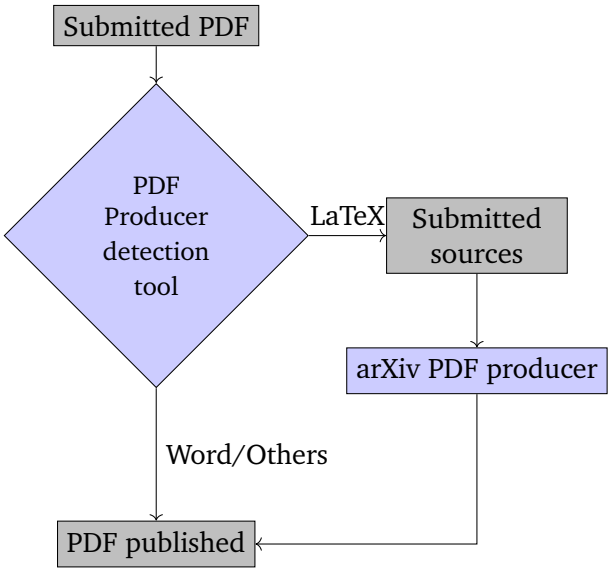

Figure. 3.8: ArXiv: PDF producer tool detection.

Attempt	Outcome
Remove the metadata (using Listing 3.14)	Fail
Compress the PDF file (using qpdf)	Fail
Re-encode using printer's driver	Fail
Use specific the fonts (True Type Fonts)	Fail
Use specific the fonts + remove metadata	Succeed

Table. 3.41: Attempts made to fool arXiv PDF detector.

the result to arXiv. All our attempts failed individually (see Table 3.41). Hence, we tried combinations of them to further test arXiv. *We succeed to fool arXiv by modifying the fonts used in the file and by removing the metadata.* arXiv accepts to publish sanitized PDF files using True Type Font/Open Type Font and created with `pdfLaTeX/XeLaTeX/LuaLaTeX`. The LaTeX code for the fonts is given in Listing 3.23 and we used Listing 3.14 to remove the metadata.

```
\documentclass{article}
\usepackage[T1]{fontenc}
\usepackage{newtxmath,newtxtext}
\usepackage{lipsum}
% just to generate dummy text
\begin{document}
\lipsum
\end{document}
```

Listing 3.23: PDF compiled using `pdfLaTeX` with Microsoft Word style font.

Arxiv PDF detector can be fooled. Therefore, arXiv should not have different policies for the publication of PDF files depending on the software used to create them. It is more consistent for arXiv to systematically request the sources and then produce the PDF file for the publication.

3.9 Extensions to PDF coding style and sanitization

Two group of students in CYSEC[36] master program have extended our work as their master projects. Coding style methodology was implemented by group 1 (Ferreol De Lepinau & Pierre Gertner) for four online PDF producer tools and PDF sanitization method was implemented by group 2 (Benjamin Bihanic & Christophe Néraud) on 40 financially significant companies in France (CAC40[37])

3.9.1 Coding style to detect PDF producer tools

Many user tend to create their PDF files online using the online PDF producer tools. We choose four such tools *Open Office, pdflib, pdfkit* and *PDFFactory* and wrote the rules to detect them using their coding style. Ferreol De Lepinau & Pierre Gertner observed that these four tools have many objects that are tool specific and could be used to identify the PDF files created using them. Table 3.42 shows number of rules written for each of these tools. Due to the contribution of this work, our methodology to detect PDF producer tools can detect 15 producer tools.

Document Section	Number of rules
PDFFactory	10
Open Office	9
pdflib	14
pdfkit	7

Table. 3.42: Number of rules for online tool's coding style.

3.9.2 Implementation of PDF file sanitization

Benjamin Bihanic & Christophe Néraud downloaded around 1041 PDF files from CAC40 companies and analyzed the different levels of sanitization followed by companies (Table 3.43).

[36] `https://casys.gricad-pages.univ-grenoble-alpes.fr/cybersec/index_en.html`
[37] https://markets.businessinsider.com/index/components/cac_40?op=1

Level of sanitization	# PDF
Level-0	625 (60%)
Level-1	370 (36%)
Level-2	1 (0.1%)
Level-3	45 (3.9%)

Table. 3.43: Different levels of sanitization of PDF documents in our database.

3.9.3 Metadata analysis of CAC40 companies

Along with checking the level of sanitization followed by the companies in CAC40, we have also analyzed different information leaked within the metadata of PDF files. Through our analysis we observed that at least 33 companies reveal the OS used, Table 3.44 summarizes our findings.

OS used	# companies	# PDFs
Microsoft Windows	33	411 (39%)
Mac OS	33	270 (26%)
Linux	2	4 (0.4%)

Table. 3.44: OS used by companies in CAC40.

Table 3.45 shows different PDF producer tools used by the employees of CAC40, *Adobe PDF Library* comes first in the list. This producer tool is actually an API that is widely used by C, Java and .NET developers. It is integrated in a lot of software that need to handle the PDF file format.

Producer tool	# PDF	# Companies
Adobe PDF Library	404 (38%)	35
Microsoft Office Word	252 (24%)	30
Acrobat Distiller	91 (9%)	19
Mac OS X Quartz	42 (4%)	7
Microsoft Office PowerPoint	37 (3.5%)	11
Ghostscript	31 (3%)	11
Microsoft Office Excel	30 (2.8%)	7
www.ilovepdf.com	19 (1.8%)	8
SKia/PDF	15 (1.4%)	3
Other tools	204 (19%)	27

Table. 3.45: Repartition of PDF producers

Online tools: We have also discovered that some PDF files published by companies were created using online tools. It implies that employees working in these companies have the habits of uploading files on these websites to create or optimize PDF files. This is dangerous if the company's sensitive data are uploaded in a third party website hosted anywhere around the globe. Such a behaviour is clearly unacceptable in a company. Table 3.46 summarizes our finding concerning the use of third party PDF producer/optimizer website. We found 38 (3.6%) PDF files produced by online third party website and at least 15 companies are concerned. For each online tool, we also looked it's IP address to identify in which country they are located, most of them are in USA.

Type of tools	Tools	Host country	# PDF	# Companies
Producer tools	www.ilovepdf.com	USA	19 (1.8%)	8
	Google Docs	USA	15 (1.4%)	3
Compressor tools	Sejda	Canada	1 (0.1%)	1
	pdfconvertonline	USA	1 (0.1%)	1

Table. 3.46: Online PDF producers/compressor tools used by companies of CAC40.

3.9.4 Analysis of software update policies

Since Microsoft Office Word is one of the most popularly used tool by companies of CAC40, we are going to focus software usage behavior analysis on the use of the Microsoft Office suite tools. We found five different versions as summarized in table 3.47. All in all, 252 (24%) PDF documents were produced using those tools.

In-use version	# PDF	# Companies
Microsoft Word 2016	107 (10%)	17
Microsoft Word for Office 365	97 (9%)	21
Microsoft Word 2013	26 (2.5%)	6
Microsoft Word 2010	19 (1.8%)	7
Microsoft Office Word 2007	3 (0.28%)	3

Table. 3.47: PDF files produced using Microsoft office tools

We focused on the number of PDF documents produced by each outdated editions of Microsoft Office software. Table 3.48 summarizes our findings.

Even if there are not many PDF files that are produced using such outdated software, it is very concerning to discover that in 2020 there are some employees still using software that has not been updated for ten years (3 different employees of the same company are concerned here). Those software are no longer maintained and supported and might contain vulnerabilities that will not be patched by their editor. Having those software still installed and using them on

Software	2020	2019	2018	$\leq$ 2017
Microsoft Word 2013	6	5	0	15
Microsoft Word 2010	3	2	1	13
Microsoft Excel 2013	1	0	0	0
Microsoft Excel 2010	0	1	1	0
Microsoft PowerPoint 2013	1	0	0	0
Microsoft PowerPoint 2010	0	3	3	1

Table. 3.48: Number of PDF files published each year using outdated software (Microsoft office tools).

the workstations of a company constitutes a security threat and is not acceptable. Moreover, if those kind of software has not been updated for so long, one might wonder whether it is the same situation for the operating systems installed on the employees' computers. We also observed some interesting behavior, where some companies updated their software but not properly: we noticed that a user has been creating PDF files using Microsoft PowerPoint 2010 until 2020, when they finally decided to update to Microsoft PowerPoint 2013. This behavior is really concerning as it shows the lack of interest towards security of the company.

The analysis done on the PDF files of companies of CAC40 clearly shows that the employees in the company are unaware of the information they leak within their PDF files. These companies need to enforce regular software updates and proper sanitization methods on their employees.

3.10 Conclusion

In the first part of our work, our study shows that the PDF files published by different security agencies are not sanitized to the level expected by such organizations. Many PDF files published by these agencies contained hidden information which can be used to target their employees. Some agencies care about sanitization but only 3 agencies out of 7 are doing it properly. Whereas, our analysis on PDF files of researchers in the scientific community showed that they do not sanitize their PDF files and distribute it along with many sensitive information which could be used against them. Either all these PDF creators are ignorant or they don't consider the leakage of information within PDF files a serious issue. We believe that the issue is that popular PDF producer tools are keeping **metadata by default** with many other information while creating a PDF. They provide no option for sanitization or it can only be achieved by following a complex procedure. Software producing PDF files need to enforce **sanitization by default**. The user should be able to add metadata only as an option.

Our study also demonstrates why Level-2 sanitization is not enough to protect a file. We were successful in the detection of 11 popular PDF producer tools. Analyzing additional PDF producer tools can extend and improve the detection rate of our tool.

In the later part of our study, our study shows that coding style can be exploited to identify which software has been used to create a PDF file. The results obtained with our tool shows that it is working with high accuracy. Our methodology is more robust than just looking at the metadata fields of the file, which is highly unreliable (it can be altered or removed from the file). We have also shown that complete metadata sanitization is much more difficult than it sounds: our work shows that the analysis of the coding style defeats sanitization.

3.10.1 Impact of our results

We have contacted 63 security agencies out of 75 and informed them about our study on the PDF files published on their websites. We provided them the number of PDF files that are not sanitized on their respective websites. We received seven acknowledgement from *customs.gov.hk, bund.de, ministers.govt.nz, defensa.gob.es, international.gc.ca, gov.si* and *receita.gov.br* security agencies. fmu.gov.pk agency replied to us saying our e-mail might be spam and on further contact they ignored our e-mails. Three agencies: *fra.se, ssi.gov.fr and interior.gob.es* replied to our e-mail and thanked us for letting them know about the un-sanitized PDF files. *ssi.gov.fr* further mentioned that they will address this issue and take necessary steps to handle it.

3.10.2 Open questions

Is it possible to sanitize a PDF file such that an adversary can never learn which software has been used to produce/modify/sanitize the file? To answer this question, we have sanitized PDF files using Adobe Acrobat and Ghostscript tools (see Listing 3.10). 100% of the time, our tool detects correctly the sanitization software. Sanitization only shifts the problem: instead of collecting information on the PDF producer software, an adversary will collect information on the sanitization software.

A countermeasure could be to create a software that changes the signature of PDF files in order to fool our PDF producer detection tool. Different strategies are available to design this countermeasure. The first strategy will be to combine many different coding style features. Our tool will not be able to detect the original PDF producer tool. However, it creates a new

PDF producer signature and identification is still possible. A similar observation was made on countermeasures used to defeat browser fingerprinting [3, 113, 112, 62]. A second strategy will be to modify the PDF file to obtain a new consistent PDF signature from another PDF producer tool. Even if the modifications are perfect, the adversary will know that the PDF file was produced with the real PDF producer or with the imitating software. The adversary still learns something.

Is it possible to have a universal PDF producer and sanitizer tool? This could eliminate the problem of identifying the PDF producer/sanitizer tool.

3.10.3 Possible extensions

We have considered just 15 PDF producer tools, an interesting question is that is it possible to extend our tool to other PDF producers? In other words, is it possible to automatize the creation the Yara rules to identify a producer tool? An exhaustive approach would consist in taking a string of fixed length in a PDF file. This string can be included in a regular expression and the new rule can be tested to check if it is accurate. If it is not, the strings can be extended until it is accurate. Unfortunately, this approach is time consuming: the complexity depends on the file size. Applying machine learning techniques could be a promising future work.

In the near future we also plan to work on the implementation of steganography. While we were working on the coding style to detect the PDF producer tools, we noticed that the escape sequences used within a PDF file and also the key-values used within an object of the PDF file could be used to implement the steganography. We plan to design a tool to perform steganography on PDF files.

4 Conclusions

This dissertation contributes to the research work associated to the personal data of online users. It starts by defining what is personal data in GDPR, then by defining different contexts where some identifiers constitute to be personal data, building profiles using different information associated to the user and finally the analysis of PDF files. In this chapter, we summarize all the contributions and the possible extensions that will be addressed in the near future.

Personal data

Personal data is at the heart of the GDPR, but many people are still unsure exactly what `personal data` refers to. Personal data is an important concept for both *data subjects* and *data controllers*. For data subjects it is important to know what personal data is in order to exercise their rights and also know the impact of collection and processing of their data by organizations. And for the data controllers it is important in order to be complaint with GDPR regulations while collecting and processing the user data.

Our work provides the prospects on the confusion around the technical and the legal aspects of personal data. Our analysis shows that, it is hard for users to understand the technical and legal aspects of personal data as it is massive and complex. Hence, users tend to share a lot information online without knowing the sensitivity of a particular piece of information.

Firstly, we provide a detailed description of personal data under GDPR and its territorial scope in EU and outside (Section 2.1). Our analysis mainly focuses on two identifiers `Name` and `IP address` of users. We demonstrate different circumstances when these two identifiers constitute to be personal data, legal and technical ambiguities associated with these identifiers, difficulties faced by data subjects and data controllers to submit/respond to a subject access request etc..

A name is perhaps the most common means of identifying someone. We describe the context when a specific individual is identified uniquely and when it is necessary to use quasi identifiers (date of birth, gender etc..) for the identification. To show that in some context a name is direct identifier, we worked on the uniqueness of names in France. Using the statistics on

the french population (INSEE dataset), we developed a tool to determine the uniqueness of a name in a particular department in France. This tool is the first step towards educating online users about the sensitivity of their names and the privacy risk associated. Using the same statistics of INSEE data, we also demonstrated the complexities of identification when a name is shared by several people. We showed how other quasi identifiers such as gender, date of birth, address etc.. could help to identify an individual (Section 2.2).

We also introduced the problem of extra territorial scope of GDPR in the application of regulations in different countries. Considering the work done in [2], we compare the naming system in China and Europe, to show different challenges faced by both data subjects and data controllers for the SAR request. Our work on the identifier *name* shows the vagueness around the definition of personal data and the complications faced in the implementation of GDPR regulations in different countries.

To demonstrate that users are unaware of the sensitivity of different information shared online and to check if usernames are personal data or not, we worked on the subject of usernames (Section 2.3). We analyzed over 20,661 usernames provided by MOOC to check if users tend to include any personal information such as first name, last name, address etc.. within their username. Our analysis shows that, many users include their names and other personal information. We were able to extract first name, last name, gender, nationality, ethnicity, postal code, department code etc.. During our analysis, we also found several online tools that provide the information of the existence of a particular username in popular websites. Since some users tend to use same username across different websites, such tools can be used by an attacker to gain information to target users. Our analysis shows that the websites that provide options to create a personalized account, do not warn or educate users on the personal information shared while creating a username and hence many users tend to think that only password is sensitive data and not the username. Our work on name and username has shown that there is a strong need to spread awareness, educate and warn users about information leakage and possible privacy risks associated to their names and sharing them online.

The second part of the chapter 2 includes, analysis of **IP addresses**. Our analysis showed that, though IP addresses are considered as identifiable personal data in the GDPR, in practice, organizations processing IP addresses handle it very differently. Our work showed how the legal ambiguity surrounding the nature of an IP address has been misused by companies. This hypothesis was tested throughout a case study wherein subject access requests containing IP addresses were sent to 109 organizations (companies and Data Protection Authorities). We found out that many of these organizations do not respond properly to data access requests

with IP address even if they specify the use of IP address as a personal data in their privacy policies.

Our work shows that, *IP addresses as personal data* is only a theoretical statement for Internet users because in practice there is no practical means for them to exercise their rights. Internet users cannot prove that they have used an IP address. Therefore, it is easy for websites to deny IP-based subject access requests. **The GDPR is not usable in this context.**

Denial of IP-based subject access requests is not going to change unless drastic modifications are made to change the way IP addresses are allocated to users. The current scheme is device-oriented and it simplifies the task for ISPs. GDPR-friendly IP address allocation needs to be user-oriented. Changing the situation is complicated because it requires to change a core Internet mechanism: IP addresses allocation. However, it is interesting to see that an IPv6 address allocation scheme [78, 77] was GDPR-friendly and considered for standardisation [12]. This modification might be possible but it is difficult and may take years. The results obtained during our analysis emphasize on the possibilities to change the scheme of allocation of IP addresses at least for the research purposes so as to understand how IP addresses are collected and processed by organizations and if it is legitimate or not.

Analysis on PDF files

Under the GDPR, an identifiable person is someone who can be identified either directly or indirectly by their name, an identification number, or their geolocation data etc.. These kinds of information could be embedded within the metadata of documents shared online (PDF files, doc files etc.). Such information are not immediately visible unless someone looks for it. Our second contribution in this dissertation is on the subject of PDF files. In particular we have worked on the hidden information in the PDF file, exploitation of this information, sanitization and forensics using coding style (Chapter 3).

`Metadata` are often considered as a threat to privacy. Often users are not aware of the metadata and that they can contain more information than the actual content they are describing. Our work takes on the metadata of PDF files, we have conducted a large scale study of 595529 PDF files published by the security agencies and scientific community. Our study on the PDF files shows that metadata provides a lot of information on the authoring process used to create PDF files. This is true when the authors directly produces a PDF file (security agencies and Cryptology ePrint Archive) and also true when a third party (open archive HAL) modifies the PDF file. Our study focuses on the different information found in the PDF files of two

different entities and there sensitivity: *security agencies* around the world and the *scientific community*.

Security agency PDF files: Our study on 39664 PDF files published by 75 security agencies of 47 countries shows that, 30155 (76%) PDF files contain metadata information and 9509 (24%) PDF files are sanitized. When the metadata are present, along with some other information, we learn the PDF producer tool and the Operating System used by the authors. Collecting and analyzing PDF files from the same source over several years can reveal the habits of a given employee. It is possible to learn if he/she update/change (or not) software regularly. For instance, we found at least one author who has never changed or updated his/her software during a period of 5 years. This kind of information is particularly interesting for a hacker to target an individual with bad software habits. By analyzing the PDF files published by several employees of the same agency, it is also possible to learn the software policies of an agency. We found at least 19 security agencies in our dataset who are using the same software over a period of 2 years or more. Around 38 security agencies have better practices and are regularly changing or updating their software. When we consider the analysis done on the sanitized PDF files, these files are not sanitized to the level expected by such organizations. Many PDF files published by these agencies contained hidden information which can be used to target their employees. Some agencies care about sanitization but only 3 agencies out of 7 are doing it properly. Our study shows that complete metadata sanitization is much more difficult than it sounds.

Scientific community PDF files: We got access to a very large dataset of 555865 PDF files from HAL and IACR pre-prints. 99.85% PDF files in IACR and 99.30% PDF files in HAL dataset contained the producer tool information in the metadata. Our extraction of the metadata has shown that 23% of the PDF files contains sensitive or valuable information on the authoring process. It included information on the organization of the authors, information on the Operating System, the tool managing the references etc.. We have been able to test if certain metadata fields can be connected to an author or other patterns. These information are very similar to those exploited for browser-fingerprinting [3, 113, 112, 45] and our results showed that just the metadata field names can be used to extract information on the author. On the practices of PDF file sanitization, we observed that PDF sanitization is not popular in the scientific community. We found only one (out of 555865) Level-2 sanitized PDF file. Most of the PDF files published by scientists are not sanitized. Since scientists publish many PDF files, they are particularly exposed to an adversary. A better practice to share publications would be changing the policies of all the scientific publishers. It may be difficult but not impossible. Each scientific community can attempt to change the policy of its main publishers. We have proposed a indirect PDF producing scheme where the authors can provide their sources to a submission & review systems, a preprint server or a camera-ready version system

and then these systems create PDF files. In this way, authors avoid leakage of any sensitive information.

Application of PDF sanitization: Many tools like Adobe Acrobat propose to erase the metadata of a file (including PDF) and it is possible to find scripts or command lines to do it yourself. Moreover, NSA has provided guidelines to sanitize PDF files in [80, 109, 81]. From a technical and user point of view, the problem of PDF files sanitization seems to be solved: technical issues have been identified and they have been implemented in widely used software. There is no obstacle left to prevent users from sanitizing their PDF files. But the issue is that popular PDF producer tools are keeping metadata by default with many other information while creating a PDF. They provide no option for sanitization or it can only be achieved by following a complex procedure. Software producing PDF files need to **enforce sanitization by default**. The user should be able to add metadata only as an option.

In particular, our study urges for security agencies and scientific community to take measures that should enforce stronger sanitization methods to limit the risk of information leak in their PDF files.

In the second part on the study of PDF files, we worked on the **PDF files forensics using coding style**. Using coding style, we implemented a tool that can be exploited to identify which software has been used to create a PDF file. Currently our tool identifies 11 popular PDF producer tools. Our tool is able to detect certain producers with an accuracy of 100% and its over all detection is still high (74%). It is more robust than just looking at the metadata fields of the file, which is highly unreliable (it can be altered or removed from the file). In our work, we exploit patterns in different sections of PDF files. It is more difficult for an adversary to manipulate the content of a PDF file to fool our tool.

Our study on the coding style also demonstrates why sanitization is not enough to protect a file. *Is it possible to sanitize a PDF file such that no one can ever learn which software has been used to produce/modify/sanitize the file?* To answer this question: we have sanitized PDF files using Adobe Acrobat and Ghostscript tools. 100% of the time, our tool detects correctly the sanitization software used. Sanitization only shits the problem: instead of collecting information on the PDF producer software, one can collect information on the sanitization software. A countermeasure could be to create a software that changes the signature of PDF files in order to fool our PDF producer detection tool. Different strategies are available to design this countermeasure. The first strategy will be to combine many different coding style features. Our tool will not be able to detect the original PDF producer tool. However, it creates a new PDF producer signature and identification is still possible. A second strategy will be to modify the PDF file to obtain a new consistent PDF signature from another PDF producer

tool. Even if the modifications are perfect, the adversary will know that one tool was used to produce the PDF and other tool to modify it.

4.1 Future work

The proposals made in this dissertation have scope for improvements and also some extensions. In this section, we mention a few extensions that we plan to do in the near future.

4.1.1 Personal data

GDPR empowers data subjects with certain rights, one of them being the **right to be forgotten**. The right to be forgotten is also known as the right to erasure. This right allows individuals to ask for their personal data to be deleted from the organization. We intend to examine this particular right. In near future we plan to access this right by sending a SAR request to be forgotten and determine if companies and organizations actually erase all the content on the requested data subject or retain some information. This is an interesting topic to explore to check if GDPR rights are properly implemented or not. This also provides more information on different kinds of information retained by organizations and there purposes.

4.1.2 Usernames

On the subject of usernames, currently some online tools (Ethena, genderizer...) allow to check gender, nationality... but this requires the user to send queries on distant servers. This leads to information leakage, which is not acceptable. Our future work will involve developing a tool that can be installed locally by the users and check different kinds of information leaked in their usernames. The tool would inspect if the username includes different informations like first name, last name, DOB, nationality, gender etc.. and warn users on sensitivity of the information, different ways the information can be perceived and exploited.

The second extension would include the analysis of **Websites revealing the information on the registered users**. While creating an username to have a personalized account in some online services, often websites show the availability of that username. Websites display a message if a particular username is already taken by one of there existing customer. This is also true when an user uses his/her email addresses to register to an service. This kind of information can reveal private information on the users, especially if the websites are related to the subjects such as finance, medical, websites that could reveal sexual orientation

or preferences etc.. In the near future, we plan to examine different websites that have high privacy impacts when there customer information are revealed and other information associated to this leakage.

4.1.3 PDF files

PDF sanitization: Discussion on PDF files in Chapter 3, shows that using only 11 popular PDF producer tools (Ghostscript, Acrobat Distiller, Microsoft Office Word, Mac OS X Quartz, LibreOffice, SKia/PDF, LuaTeX, PDFLaTeX, xdviPDFmx, Cairo and PdfTeX), many PDF files (48% of PDF files from Security agency dataset and 96% from IACR dataset and 91% fro HAL dataset) were created. At least these tools can implement sanitization by default and this could be a starting step towards the implementation of sanitization for safer distribution of PDF files.

Coding style to detect producer tool: An interesting question on the work done on coding style of the PDF files is that *is it possible to extend our tool to other PDF producers?* In other words, is it possible to automatize the creation the Yara rules to identify a producer tool? An exhaustive approach would consist in taking a string of fixed length in a PDF file. This string can be included in a regular expression and the new rule can be tested to check if it is accurate. If it is not, the strings can be extended until it is accurate. Unfortunately, this approach is time consuming: the complexity depends on the file size. Applying machine learning techniques could be a promising future work.

PDF steganography: During our analysis of coding style of producer tools, we found that the coding style can be used to implement steganography. The escape sequences used within the PDF files, keys-values used in the PDF objects are good sources that open means for the steganography. In the near future, we plan to implement a tool that could be used to implement steganography on different PDF files created using different producer tools.

4.2 Publications and Dissertation impact

1. The work on the personal information inferred on username was accepted as a short paper: **Mooc and Privacy** in APVP 2018 - l'Atelier sur la Protection de la Vie Privée.

2. Analysis of IP address lead to the submission of a full paper: **Why IP-based Subject Access Requests Are Denied?** to ACM Internet Measurement Conference 2021 (currently under review). Link: `https://arxiv.org/pdf/2103.01019.pdf`

3. The analysis done on the PDF files of security agencies was accepted as a full paper: **Exploitation and Sanitization of Hidden Data in PDF Files** in the ACM conference: Information Hiding and Multimedia Security (IH&MMSec 2021 June 22-25, 2021, Virtual Event, Belgium). Link: `https://arxiv.org/pdf/2103.02707.pdf`

4. Coding style and detection of PDF producer tools lead to a submission of a full paper: **Robust PDF Files Forensics Using Coding Style** to European Symposium on Research in Computer Security (ESORICS) 2021 (currently under review). Link: `https://arxiv.org/pdf/2103.02702.pdf`

Bibliography

[1] *29 Working Party, Opinion 1/2oo8 on Data Protection Issues Related to Search Engines, 8, oo737/EN/WP 148 (Apr. 4, 2008)* (cit. on p. 39).

[2] *Abusing the GDPR for Surveillance and Censorship*. Personal article, not published.. (cit. on pp. 22, 138).

[3] Gunes Acar, Christian Eubank, Steven Englehardt, et al. „The Web Never Forgets: Persistent Tracking Mechanisms in the Wild“. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. Scottsdale, AZ, USA: ACM, 2014, pp. 674–689 (cit. on pp. 2, 136, 140).

[4] Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. „Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising“. In: `https://doi.org/10.1145/2501604.2501612`. Association for Computing Machinery, 2013 (cit. on p. 2).

[5] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery. „Scoring Users' Privacy Disclosure Across Multiple Online Social Networks“. In: *IEEE Access* 5 (2017), pp. 13118–13130 (cit. on p. 2).

[6] Istemi Akkus, Ruichuan Chen, Michaela Hardt, Paul Francis, and Johannes Gehrke. „Non-tracking Web Analytics“. In: (Oct. 2012) (cit. on p. 2).

[7] Chema Alonso and José Palzon. „Tactical Fingerprinting using Foca“. In: *DEF CON 17 Hacking Conference*. Las Vegas, VA, USA, 2009, pp. 41–50 (cit. on p. 102).

[8] Chema Alonso, Enrique Rando, Francisco Oca, and Antonio Guzmán. „Disclosing Private Information from Metadata, hidden info and lost data“. In: *Black Hat Europe 2009*. Amsterdam, The Netherlands, 2009 (cit. on p. 75).

[9] Chema Alonso, Enrique Rando, Francisco Oca, and Antonio Guzmán. „Disclosing Private Information from Metadata, hidden info and lost data“. In: *Black Hat Europe 2009*. Amsterdam, The Netherlands, 2009 (cit. on p. 102).

[10] *ARTICLE 29 DATA PROTECTION WORKING PARTY, Opinion 4/2007 on the concept of personal data Adopted on 20th June* (cit. on p. 14).

[11] *Article 4 EU GDRP "Definitions*. `http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm` (cit. on pp. v, vii).

[12] Tuomas Aura. „Cryptographically Generated Addresses (CGA)“. In: *RFC* 3972 (2005), pp. 1–22 (cit. on pp. 63, 139).

[13]Tuomas Aura, Thomas A. Kuhn, and Michael Roe. „Scanning electronic documents for personally identifiable information“. In: *Proceedings of the 2006 ACM Workshop on Privacy in the Electronic Society, WPES 2006*. Ed. by Ari Juels and Marianne Winslett. Alexandria, VA, USA: ACM, 2006, pp. 41–50 (cit. on pp. 4, 7, 73, 76, 77, 99).

[14]Jef Ausloos and Pierre Dewitte. „Shattering one-way mirrors – data subject access rights in practice“. In: *International Data Privacy Law* 8.1 (2018), pp. 4–28 (cit. on p. 45).

[15]Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. „Security Analysis of Subject Access Request Procedures How to authenticate data subjects safely when they request for their data“. In: *2019 - Annual Privacy Forum*. Rome, Italy, June 2019, pp. 1–20 (cit. on p. 45).

[16]Matteo Cagnazzo, Thorsten Holz, and Norbert Pohlmann. „GDPiRated - Stealing Personal Information On- and Offline“. In: *Computer Security - ESORICS 2019 - 24th European Symposium on Research in Computer Security*. Vol. 11736. Lecture Notes in Computer Science. Luxembourg: Springer, 2019, pp. 367–386 (cit. on p. 45).

[17]Curtis Carmony, Xunchao Hu, Heng Yin, Abhishek Vasisht Bhaskar, and Mu Zhang. „Extract Me If You Can: Abusing PDF Parsers in Malware Detectors“. In: *23rd Annual Network and Distributed System Security Symposium, NDSS 2016*. San Diego, CA, USA: The Internet Society, 2016 (cit. on pp. 4, 73).

[18]*Case C-101/01 Criminal proceedings against Bodil Lindqvist*. ECLI:EU:C:2016:779. 2003 (cit. on pp. 3, 38).

[19]*Case C-582/14 Breyer v Bundesrepublik Deutschland*. ECLI:EU:C:2016:779. 2016 (cit. on pp. 3, 38–40).

[20]Aniello Castiglione, Alfredo De Santis, and Claudio Soriente. „Security and privacy issues in the Portable Document Format“. In: *Journal of Systems and Software* 83.10 (2010), pp. 1813–1822 (cit. on p. 79).

[21]*Catch Up on Privacy Around the World on Data Privacy Day 2021!* `https://www.mofo.com/resources/insights/210127-data-privacy-day.html`. 2021 (cit. on p. 2).

[22]Abdelberi Chaabane, Mohamed Ali Kaafar, and Roksana Boreli. „Big Friend is Watching You: Analyzing Online Social Networks Tracking Capabilities“. In: `https://doi.org/10.1145/2342549.2342552`. Association for Computing Machinery, 2012 (cit. on p. 2).

[23]Stephen Checkoway, Hovav Shacham, and Eric Rescorla. „Are text-only data formats safe ? Or, use this LaTeX class file to pwn your computer“. In: *Proceedings of LEET 2010. USENIX*. San Diego, California, USA, 2010 (cit. on p. 106).

[24]Stephen Checkoway, Hovav Shacham, and Eric Rescorla. *Don't take LaTeX files from strangers*. 2011 (cit. on p. 106).

[25]Andrew Cormack. „Is the Subject Access Right Now Too Great a Threat to Privacy?“ In: *European Data Protection Law Review* 2.1 (2016), pp. 15–27 (cit. on p. 45).

[26]M. Cotton and L. Vegoda. *Special Use IPv4 Addresses*. RFC 5735. RFC Editor, 2010, pp. 1–11 (cit. on p. 35).

[27] Court of Justice of the European Union. *Case 582/14 – Patrick Breyer v Germany*. ECLI:EU:C:2016:779. 2016 (cit. on pp. 3, 34).

[28] Court of Justice of the European Union. *Case C-70/10, – Scarlet Extended v SABAM*. ECLI:EU:C:2011:771. 2011 (cit. on p. 38).

[29] Samuel D. Warren and Louis D. Brandeis. "*The Right to Privacy*". *Harvard Law Review. IV*. December 1890 (cit. on p. 25).

[30] *Directive 2009/136/ec of the european parliament and of the council.* `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:0036:en:PDF` (cit. on p. 2).

[31] *ECHR 2 December 2008, K.U. v. Finland, Application no. 2872/02*. 2008 (cit. on p. 39).

[32] European Data Protection Board (EDPB). *Article 29 Working Party, WP 37: Privacy on the Internet – An Integrated EU Approach to On-line Data Protection, at 21, adopted on Nov. 21, 2000.* 2000 (cit. on p. 40).

[33] European Data Protection Board (EDPB). *EDPB Opinion 4/2007 on the concept of personal data (WP 136), adopted on 20.06.2007*. 2007 (cit. on pp. 38, 44, 45).

[34] European Data Protection Board (EDPB). *Opinion 1/2008 on data protection issues related to search engines, Adopted on 4 April 2008 (WP 148)*. 2008 (cit. on p. 40).

[35] Steven Englehardt and Arvind Narayanan. „Online Tracking: A 1-Million-Site Measurement and Analysis“. In: Association for Computing Machinery, 2016 (cit. on p. 2).

[36] Court of Justice of the European Union. *Case C-70/10, – Scarlet Extended v SABAM*. ECLI:EU:C:2011:771. 2011 (cit. on p. 3).

[37] Sebastian Dabkiewicz Fahimeh Alizadeh Nicolas Canceill and Diederik Vandevenne. *Using Steganography to hide messages inside PDF files*. 2012 (cit. on p. 77).

[38] Jeremy Faircloth. „Chapter 2 - Reconnaissance“. In: *Penetration Tester's Open Source Toolkit (Fourth Edition)*. Ed. by Jeremy Faircloth. Fourth Edition. Syngress (cit. on p. 101).

[39] Yun Feng, Baoxu Liu, Xiang Cui, et al. „A Systematic Method on PDF Privacy Leakage Issues“. In: *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018*. New York, NY, USA: IEEE, 2018, pp. 1020–1029 (cit. on pp. 4, 7, 73, 77).

[40] Ricard Fogues, Jose Such, Agustín Espinosa, and Ana García-Fornes. „Open Challenges in Relationship-Based Privacy Mechanisms for Social Network Services“. In: *International Journal of Human-Computer Interaction* 31 (Feb. 2015), pp. 0–0 (cit. on p. 30).

[41] Imane Fouad, Cristiana Santos, Arnaud Legout, and Nataliia Bielova. „Did I delete my cookies? Cookies respawning with browser fingerprinting“. working paper or preprint. May 2021 (cit. on p. 36).

[42] S. L. Garfinkel. „Leaking Sensitive Information in Complex Document Files–and How to Prevent It“. In: *IEEE Security Privacy* 12.1 (2014), pp. 20–27 (cit. on pp. 4, 7, 73, 75).

[43] *GDPR. Issues: Personal Data*. `https://gdpr-info.eu/issues/personal-data/` (cit. on p. 13).

[44]"*General Data Protection Regulation*". https://gdpr-info.eu/art-3-gdpr/ (cit. on p. 11).

[45]Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. „Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale“. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*. Lyon, France: ACM, 2018, pp. 309–318 (cit. on p. 140).

[46]*Guidelines 3/2018 on the territorial scope of the GDPR(Article 3) Version 2.1*. November 2019 (cit. on p. 11).

[47]*Guidelines on transparency under Regulation2016/679, (WP260),” 2018* (cit. on p. 44).

[48]Tim Hickman, Matthias Goetz, Deltev Gabel, and Chris Ewing. „IP addresses and personal data: Did CJEU ask the right questions?“ In: *Privacy Laws & Business International Report* (2017) (cit. on p. 38).

[49]Jockum Hildén. *Am I my IP address's keeper? Revisiting the boundaries of information privacy, The Information Society, 33:3, 159-171, DOI: 10.1080/01972243.2017.1294127*. 2017 (cit. on p. 38).

[50]R. Hinden and S. Deering. *Internet Protocol Version 6 (IPv6) Addressing Architecture*. RFC 3513. RFC Editor, 2003, pp. 1–26 (cit. on pp. 3, 35, 37).

[51]Peter Hustinx. *Nameless Data Can Still be Personal, OUT-LAW.COM, Nov. 6, 2008, http://www.out-law.com/page-9563* (cit. on p. 39).

[52]*Internet advertising revenue report*. https://www.iab.com/wp-content/uploads/2019/10/IAB-HY19-Internet-Advertising-Revenue-Report.pdf (cit. on p. 11).

[53]*Internet Assigned Numbers Authority (IANA)* (cit. on p. 35).

[54]*Internet Protocol*. Tech. rep. 791. Sept. 1981. 51 pp. (cit. on pp. 3, 37).

[55]Digital Rights Ireland. *Cases C-293/12 and C-594/12 Digital Rights Ireland [2014] EU:C:2014:238, para 26.* 2014 (cit. on p. 3).

[56]ISO. *Document management—Portable document format—Part 2: PDF 2.0*. ISO ISO 32000-2:2017. Geneva, Switzerland: International Organization for Standardization, 2008 (cit. on pp. 9, 67, 108, 111, 113).

[57]David M. Martin Jr., Hailin Wu, and Adil Alsaid. „Hidden surveillance by Web sites: Web bugs in contemporary use“. In: *Commun. ACM* 46.12 (2003), pp. 258–264 (cit. on pp. 4, 73, 79).

[58]Brian W. Kernighan and P. J. Plauger. *The elements of programming style (2. ed.)* McGraw-Hill, 1978 (cit. on pp. 9, 109).

[59]Alessandro KHOURY. „Dynamic IP Addresses Can be Personal Data, Sometimes. A Story of Binary Relations and Schrödinger's Cat“. In: *European Journal of Risk Regulation* 8 (Mar. 2017), pp. 191–197 (cit. on p. 38).

[60]David M. Kristol. „HTTP Cookies: Standards, privacy, and politics“. In: *ACM Trans. Internet Techn.* 1.2 (2001), pp. 151–198 (cit. on pp. 3, 37).

[61]Frederick Lah. „Are IP Addresses "Personally Identifiable Information"?“ In: *Journal of Law and Policy for the Information Society* 4.3 (2008) (cit. on p. 38).

[62] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. „Browser Fingerprinting: A Survey“. In: *ACM Trans. Web* 14.2 (2020), 8:1–8:33 (cit. on pp. 2, 3, 37, 136).

[63] Pavel Laskov and Nedim Srndic. „Static detection of malicious JavaScript-bearing PDF documents“. In: *Twenty-Seventh Annual Computer Security Applications Conference, ACSAC 2011*. Orlando, FL, USA: ACM, 2011, pp. 373–382 (cit. on p. 79).

[64] Kai Li, Zhangxi Lin, and Xiaowen Wang. „An empirical analysis of users' privacy disclosure behaviors on social network sites“. In: *Information & Management* 52.7 (2015), pp. 882–891 (cit. on p. 2).

[65] Aleksandr V. Litvinov. „The Data Protection Directive as Applied to Internet Protocol (Ip) Addresses: Uniting the Perspective of the European Commission with the Jurisprudence of Member States“. In: *The George Washington International Law Review* 45 (2013), p. 579 (cit. on p. 34).

[66] Aleksandr V. Litvinov. „The Data Protection Directive as Applied to Internet Protocol (Ip) Addresses: Uniting the Perspective of the European Commission with the Jurisprudence of Member States“. In: *The George Washington International Law Review* 45 (2013), p. 579 (cit. on p. 38).

[67] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. „M.: On Dominant Characteristics of Residential Broadband Internet Traffic“. In: *In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. 2009, pp. 90–102 (cit. on pp. 36, 37, 41).

[68] Davide Maiorca and Battista Biggio. „Digital Investigation of PDF Files: Unveiling Traces of Embedded Malware“. In: *IEEE Security & Privacy* 17.1 (2019), pp. 63–71 (cit. on pp. 4, 73).

[69] Davide Maiorca and Battista Biggio. „Digital Investigation of PDF Files: Unveiling Traces of Embedded Malware“. In: *IEEE Security & Privacy* 17.1 (2019), pp. 63–71 (cit. on p. 79).

[70] Davide Maiorca, Battista Biggio, and Giorgio Giacinto. „Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks“. In: *ACM Comput. Surv.* 52.4 (2019), 78:1–78:36 (cit. on p. 79).

[71] Mariano Di Martino, Pieter Robyns, Winnie Weyts, et al. „Personal Information Leakage by Abusing the GDPR 'Right of Access'“. In: *Fourteenth Symposium on Usable Privacy and Security, SOUPS 2018*. Santa Clara, CA, USA: USENIX Association, 2019 (cit. on p. 45).

[72] J. McIntyre. „Balancing Expectations of Online Privacy: Why Internet Protocol (IP) Addresses Should Be Protected as Personally Identifiable Information“. In: *Depaul Law Review* 60 (2010), p. 895 (cit. on p. 38).

[73] Michael Meli, Matthew R. McNiece, and Bradley Reaves. „How Bad Can It Git? Characterizing Secret Leakage in Public GitHub Repositories“. In: *26th Annual Network and Distributed System Security Symposium, NDSS*. To appear. San Diego, California, USA: The Internet Society, 2019 (cit. on p. 106).

[74] Karl Mendelman. „Fingerprinting an Organization Using Metadata of Public Documents“. MA thesis. Estonia: University of Tartu, 2018 (cit. on pp. 7, 75, 76).

[75]Vikas Mishra, Pierre Laperdrix, Antoine Vastel, et al. „Don't Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem". In: *WWW '20: The Web Conference 2020*. Taipei, Taiwan: ACM/IW3C2, 2020, pp. 808–815 (cit. on pp. 3, 37, 41).

[76]J. M. Mittman. *German court rules that IP addresses are not personal data. Proskauer, October 10*. `https://www.pinsentmasons.com/out-law/news/german-court-says-ip-addresses-in-server-logs-are-not-personal-data`. 2008 (cit. on p. 40).

[77]Gabriel Montenegro and Claude Castelluccia. „Crypto-based Identifiers (CBIDs): Concepts and Applications". In: *ACM Trans. Inf. Syst. Secur.* 7.1 (2004) (cit. on pp. 63, 66, 139).

[78]Gabriel Montenegro and Claude Castelluccia. „Statistically Unique and Cryptographically Verifiable (SUCV) Identifiers and Addresses". In: *Proceedings of the Network and Distributed System Security Symposium, NDSS 2002*. San Diego, CA, USA: The Internet Society, 2002 (cit. on pp. 63, 66, 139).

[79]Paul Norris Cliveand de Hert, Xavier L'Hoiry, and Antonella Galetta. *The Unaccountable State of Surveillance Exercising Access Rights in Europe*. Springer International Publishing, 2017 (cit. on p. 45).

[80]NSA. *Hidden Data and Metadata in Adobe PDF Files: Publication Risks and Countermeasures*. Tech. rep. National Security Agency, 2008 (cit. on pp. 80, 92, 93, 125, 141).

[81]NSA. *Redaction of PDF Files Using Adobe Acrobat Professional X*. Tech. rep. I73-025R-2011. `https://apps.nsa.gov/iaarchive/library/ia-guidance/security-configuration/applications/redaction-of-pdf-files-using-adobe-acrobat-professional-x.cfm`. National Security Agency, 2015 (cit. on pp. 80, 92, 93, 141).

[82]*Official Journal of the European Union. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. 2016 (cit. on p. 2).

[83]Ramakrishna Padmanabhan, Amogh Dhamdhere, Emile Aben, kc claffy, and Neil Spring. „Reasons Dynamic Addresses Change". In: *Proceedings of the 2016 ACM on Internet Measurement Conference, IMC 2016, Santa Monica, CA, USA, November 14-16, 2016*. ACM, 2016, pp. 183–198 (cit. on p. 36).

[84]*Paris Appeal Court decision - Anthony G. vs. SCPP (27.04.2007)*. `http://www.legalis.net/jurisprudence-decision.php3?id_article=1954`. 2007 (cit. on pp. 3, 40).

[85]*Paris Appeal Court decision - Henri S. vs. SCPP (15.05.2007)*. `http://www.legalis.net/jurisprudence-decision.php3?id_article=195`. 2007 (cit. on pp. 3, 40).

[86]James Pavur. „GDPArrrrr: Using Privacy Laws to Steal Identities". In: *Blackhat USA 2019*. Las Vegas, NV, USA, 2019 (cit. on p. 45).

[87]*PDF Reference: Adobe Portable Document Format Version 1.4 with Cdrom*. 3rd. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001 (cit. on pp. 68, 69, 72, 112).

[88]Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. „How Unique and Traceable Are Usernames?" In: *Privacy Enhancing Technologies*. Ed. by Simone Fischer-Hübner and Nicholas Hopper. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–17 (cit. on p. 28).

[89] Emmanuel Pernot-Leplay. *A Third Way Between the China's Approach on Data Privacy Law: Between the U.S. and the E.U.?* May 2020 (cit. on p. 21).

[90] *Personal information online code of practice*. Tech. rep. 2010 (cit. on p. 65).

[91] *Publications Office of the EU, Study of case law on the circumstances in which IP addresses are considered personal data*. `https://op.europa.eu/en/publication-detail/-/publication/d7c71500-75a3-4b1c-9210-96c74b6fa2be/language-en`. 2011 (cit. on pp. 38, 39).

[92] Elie Raad, Richard Chbeir, and Albert Dipanda. „User Profile Matching in Social Networks". In: Sept. 2010 (cit. on p. 28).

[93] *Regional Internet Registries (RIRs)* (cit. on p. 35).

[94] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016 (cit. on pp. 13, 38).

[95] Alan Stewart Reid. „The European Court of Justice case of Breyer". In: *Journal of Information Rights, Policy and Practice* 2.1 (2017) (cit. on p. 38).

[96] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. „Detecting and defending against third-party tracking on the web". In: 2012 (cit. on p. 2).

[97] Manuel Campos Sanchez-Bordona. *Opinion of Advocate General Campos Sanchez-Bordona in Case C-582/14 Breyer v Bundesrepublik Deutschland*. ECLI:EU:C:2016:339. 2016 (cit. on p. 38).

[98] Zain Shamsi, Ankur Nandwani, Derek Leonard, and Dmitri Loguinov. „Hershel: single-packet os fingerprinting". In: *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '14*. Austin, TX, USA: ACM, 2014, pp. 195–206 (cit. on p. 102).

[99] Richard M. Smith. *Microsoft Word Documents that Phone Home*. Retrieve on Internet Archive: Wayback Machine `https://web.archive.org/web/20010203194100/http://www.privacyfoundation.org/advisories/advWordBugs.html`. 2000 (cit. on pp. 4, 73, 79).

[100] Charles Smutz and Angelos Stavrou. „Malicious PDF detection using metadata and structural features". In: *28th Annual Computer Security Applications Conference, ACSAC 2012*. Orlando, FL, USA: ACM, 2012, pp. 239–248 (cit. on pp. 4, 73, 75).

[101] Pavol Sokol, Jakub Mísek, and Martin Husák. „Honeypots and honeynets: issues of privacy". In: *EURASIP Journal on Information Security* 2017 (2017), pp. 1–9 (cit. on p. 34).

[102] Nedim Srndic and Pavel Laskov. „Detection of Malicious PDF Files Based on Hierarchical Document Structure". In: *20th Annual Network and Distributed System Security Symposium, NDSS 2013*. San Diego, CA, USA: The Internet Society, 2013 (cit. on p. 79).

[103] Didier Stevens. „Malicious PDF Documents Explained". In: *IEEE Security & Privacy* 9.1 (2011), pp. 80–82 (cit. on pp. 4, 73, 79).

[104] Katherine Strater and Heather Richter. „Examining Privacy and Disclosure in a Social Networking Community". In: New York, NY, USA: Association for Computing Machinery, 2007 (cit. on p. 2).

[105]Latanya Sweeney. „Simple Demographics Often Identify People Uniquely“. In: *Health* 671 (Jan. 2000) (cit. on p. 28).

[106]Systems and Network Attack Center. *Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF*. Tech. rep. National Security Agency, 2005 (cit. on pp. 74, 80, 93, 125).

[107]*The ePrivacy Directive*. `https://ec.europa.eu/digital-single-market/en/news/eprivacy-directive` (cit. on p. 11).

[108]Vetle Torvik and Sneha Agarwal. *Ethnea—an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database*. Mar. 2016 (cit. on p. 31).

[109]Unified Cross Domain Capabilities Office. *Inspection and Sanitization Guidance for Portable Document Format*. Tech. rep. National Security Agency, 2011 (cit. on pp. 92, 141).

[110]Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. „A Study on Subject Data Access in Online Advertising After the GDPR“. In: *Data Privacy Management, Cryptocurrencies and Blockchain Technology - ESORICS 2019 International Workshops, DPM 2019 and CBT 2019*. Vol. 11737. Lecture Notes in Computer Science. Luxembourg: Springer, 2019, pp. 61–79 (cit. on p. 45).

[111]Tavish Vaidya, Eric Burger, Micah Sherr, and Clay Shields. „Where art thou, Eve? Experiences laying traps for Internet eavesdroppers“. In: *Workshop on Cyber Security Experimentation and Test, CSET 2017*. Vancouver, BC, Canada: USENIX Association, 2017 (cit. on p. 79).

[112]Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. „Fp-Scanner: The Privacy Implications of Browser Fingerprint Inconsistencies“. In: *27th USENIX Security Symposium, USENIX Security 2018*. Baltimore, MD, USA: USENIX Association, 2018, pp. 135–150 (cit. on pp. 2, 136, 140).

[113]Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. „FP-STALKER: Tracking Browser Fingerprint Evolutions“. In: *2018 IEEE Symposium on Security and Privacy, SP 2018*. San Francisco, CA, USA: IEEE Computer Society, 2018, pp. 728–741 (cit. on pp. 2, 136, 140).

[114]*Web tracking protection*. `http://www.w3.org/Submission/web-tracking-protection/` (cit. on p. 2).

[115]Weilin Xu, Yanjun Qi, and David Evans. „Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers“. In: *23rd Annual Network and Distributed System Security Symposium, NDSS 2016*. San Diego, CA, USA: The Internet Society, 2016 (cit. on pp. 4, 73).

[116]Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martín Abadi. „Host Fingerprinting and Tracking on the Web: Privacy and Security Implications“. In: *19th Annual Network and Distributed System Security Symposium, NDSS 2012*. San Diego, California, USA: The Internet Society, 2012 (cit. on p. 36).

[117]Frederik Zuiderveen Borgesius. „Breyer Case of the Court of Justice of the European Union: IP Addresses and the Personal Data Definition (Case Note)“. In: *European Data Protection Law Review* 3.1 (2017) (cit. on p. 38).

List of Figures

List of Tables