UNIVERSITÀ
DEGLI STUDI DI
SALERNO

Ministero dell'Istruzione,
dell'Università e della Ricerca

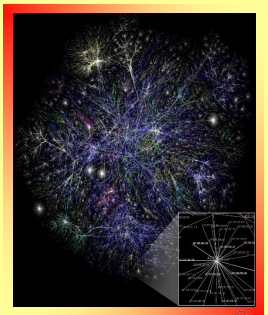
Dottorato di Ricerca in Informatica
e Ingegneria dell'Informazione

Domenico Desiato – Enhancing the Sharing and the Management of Personal Data in the Big Data Era

Tesi di Dottorato/Ph.D. Thesis

Enhancing the Sharing and the Management of Personal Data in the Big Data Era

Domenico Desiato



Supervisor: **Prof. Giuseppe Polese**

Ph.D. Program Director: **Prof. Pasquale Chiacchio**

Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata
Dipartimento di Informatica

Ciclo 33 – AA 2020/2021

Università degli Studi di Salerno

Dottorato di Ricerca in Informatica e Ingegneria dell'Informazione
Ciclo 33 – a.a 2020/2021

TESI DI DOTTORATO / PH.D. THESIS

Enhancing the Sharing and the Management of Personal Data in the Big Data Era

DOMENICO DESIATO

SUPERVISOR: **PROF. GIUSEPPE POLESE**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Dipartimento di Ingegneria dell'Informazione ed Elettrica
e Matematica Applicata
Dipartimento di Informatica

Abstract

Nowadays, thanks to the digitalization of business processes and public administrations, many significant Big data collections are available. Users are direct suppliers of data when publishing contents on social networks. However, when using a service on the web, users must often provide their data, which will become property of the company running the service. To this end, users need to be aware of the privacy issues related to the management of their data, whereas companies need to ensure the protection of users' personal data, also according to new laws and regulations issued by governments. On the other hand, there exists the necessity not to limit the processing of data by companies and other public institutions. Thus, it is necessary to devise methods devoted to the identification of possible privacy threats during users' online activities, and to develop privatization strategies that possibly do not downgrade the significance of data.

This dissertation provides experimental evidence of several threats for users when providing their personal data for accessing online services, aiming to increase their awareness, and it describes new methodologies and tools to support companies when processing personal data of their users. In particular, the proposed methodologies exploit data correlations expressed in terms of relaxed functional dependencies (RFDs) to define privatization strategies, aiming to safeguard user's privacy, and to detect malicious accounts in social networks. Finally, two automatic tools have been designed and implemented to help users better understand privacy threats during their online activities.

Abstract

Nell'era corrente, grazie alla digitalizzazione dei processi aziendali e delle pubbliche amministrazioni, sono disponibili grandi collezioni di dati. Gli utenti sono diventati fornitori diretti dei dati, ad esempio quando pubblicano contenuti sui social network o utilizzano servizi sul web, i quali diventeranno proprietá delle aziende che gestiscono tali servizi. Per questo motivo, gli utenti devono essere consapevoli delle problematiche di privacy dei propri dati e le aziende devono garantire la protezione dei dati personali, in conformitá con le nuove regolamentazioni. Tuttavia, é necessario non limitare il trattamento dei dati da parte di aziende/enti pubblici. Pertanto, é necessario sviluppare metodologie dedicate all'identificazione di possibili minacce alla privacy durante le attivitá di utilizzo dei servizi online e sviluppare strategie di privatizzazione che non degradino l'informazione intrinseca dei dati.

Questa tesi fornisce prove sperimentali inerenti problematiche riscontrabili quando gli utenti forniscono dati personali per accedere ai servizi fruibili online. Allo scopo di rendere gli utenti consapevoli riguardo la privacy dei dati, la tesi descrive nuove metodologie/strumenti per supportare le aziende nel trattamento dei dati personali. In particolare, le metodologie proposte sfruttano le correlazioni di dati, espresse in termini di dipendenze funzionali rilassate (RFD), per definire strategie di privatizzazione volte a salvaguardare la privacy dell'utente e per discriminare account malevoli nel dominio dei social network. Infine, sono stati progettati e implementati due strumenti automatici per offrire supporto, in termini di privatizzazione, quando gli utenti utilizzano servizi fruibili online.

Indice

Capitolo 1

Introduction

"Big data" is a general term introduced for referring to the growing of information sources, and is characterized by different properties, the main of which are high volume, heterogeneity, and variability over time.

Nowadays, users are the main providers of these data through their online activities. However, although users' data feed advanced analytics processes that organizations can use for developing innovative insights, products, and services, the management of personal information and their application in such processes can yield severe privacy threats. In such scenario, it is necessary to identify any sensitive data and handle the related privacy issues.

Sensitive data refer to those data that uniquely identify individuals or sensitive information about them. This kind of data is mostly spread over the Internet, and is often provided for accessing online services or to share contents over social network platforms. Sometimes, users are not aware of the privacy threats related to inadequate management of their sensitive data. Typical scenarios involve users that want to easily access online services without thinking about privacy issues concerning their sensitive data. In particular, if users grant consent to process their personal data, they usually have no awareness concerning who and how manage such data [1]. This could yield different privacy issues, such as the possibility to deal with malicious accounts that

might jeopardize users' identities [2]. Therefore, the application of privacy-preserving policies in the virtual world represents a relevant problem. This is because it is challenging to implement forms of control policies for the Web, considering its information-sharing nature.

Another example of risk exposure is the creation of a virtual life strictly coupled to the physical one, which can be often found on social networks. The latter represent a significant source of information, and also in this context sensitive data are massively spread. Most of the social platforms like Facebook, Twitter, and Instagram permit people to share emotions, ways of thinking, points of view, and so on. Thus, social networks play a fundamental role to simplify and promote interactions among people. Users tend to use social networks to share information massively; in most cases, they do not care about privatizing data and are unaware of the privacy threats they can be exposed to. Thus, in a context in which a social network profile contains detailed information that uniquely refers to a specific individual, preserving his/her privacy becomes a fascinating challenge.

The safe management of sensitive data also has a significant impact on business processes. In fact, companies need to manage users data with parsimony, by providing privatization strategies for safeguarding them. Companies have many difficulties in determining how they can use the data to avoid legal issues related to data privacy violations. Standard privacy preservation techniques, such as cryptography and obfuscation, could lead to the impossibility of using the data, even if some of them are not sensitive. Thus, it is necessary to distinguish between sensitive and non-sensitive data in an effective way. In general, the management of sensitive data becomes even more critical in complex application domains, such as IoT environments [3], Smart Grids [4], Social Networks [5], and so on. As an example, the necessity to manage sensitive data arises when hospitals adopt sensor networks to monitor patients, and in particular, disabled patients. Moreover, guaranteeing privacy preservation becomes an increasingly complex task when accomplishing advanced data processing operations, like for

instance, data integration [6], and record-linkage [7]. In fact, such processes could yield privacy violations when the compared data sources contain sensitive data: even if sensitive data are obscured to meet users privacy requirements, such data processing operations could introduce new sensitive data due to the generation of new identification patterns.

Particular interest should also be devoted to automatic techniques for extrapolating knowledge from data, like for example, in classification or data mining. Also such techniques can potentially compromise users privacy when analyzed data are referred to users. In particular, machine learning results can potentially disclosure sensitive data and jeopardize users privacy when they are released for public accesses. Thus, also in this application domain, there is the need to develop possibly automated solutions for preservation, by guaranteeing the possibility to analyze data and results.

With this in mind, this thesis presents several methodologies to support companies in the management of privacy-preserving issues, together with experimental evaluations and tools to stimulate users to increase their awareness when exposing their personal data to possible privacy threats. In particular, referring to the user awareness problem, two automatic tools applied in the Web browsing and social networks scenarios are presented. The first one represents a visual analytics tool that allows users to understand how their sensitive data are exchanged or shared among different network services. The tool visualizes the communication flow generated during Web browsing activities, highlighting the providers tracking their data. It draws a real-time summary graph showing the information tracked from the service providers to which the user connects over the internet [8]. Instead, with the aim of evaluating privacy threats when users share information on different social networks and making the users aware of these issues, the second tool exploits image-recognition techniques to recognize a user from his/her picture, aiming to collect his/her personal data accessible through social networks where s/he has a profile. Finally, the last proposal in the social network context is represented by

a novel methodology devoted to the detection of fake accounts. It aims to support the discrimination of "anomalous" accounts that can compromise the trustability of users' activities. In particular, the methodology exploits correlations holding on the data stored in the social networks, and a new heuristic to derive a predictive profile for fake accounts [2].

Concerning the problem of correctly safeguarding sensitive data, a new methodology is presented. It exploits data profiling strategies to automatically evaluate possible implications among data that could disclose the values of sensitive ones. Thus, the proposed methodology permits to increase the confidentiality of a dataset, while reducing the number of values to be obscured in order to admit analysis processes over the not obscured ones [9]. A second proposal represents a novel methodology for identifying data anonymization strategies for guaranteeing data privacy. Also in this case, data profiling strategies have been used, but for defining suitable generalization rules to anonymize data. This methodology also exploits a multi-objective optimization strategy (i.e., the Pareto frontier), to help data owners in releasing anonymized datasets on which classification activities can be performed.

Thesis outline. The thesis starts by introducing privacy problems in the big data context in Chapter 2. Then, Chapter 3 introduces background definitions and well-known privacy preserving techniques to permit a better understanding of the methodologies described in later chapters of this dissertation. Chapter 4 introduces the new proposals to improve users' privacy awareness during Internet Web browsing, statistical analysis of users' privacy in social networks, and fake account identification by using data profiling. Chapter 5 describes a new methodology to preserve information confidentiality in big data processing tasks, like data integration. Instead, in Chapter 6, a new data anonymization methodology for guaranteeing privacy in machine learning contexts is presented. Finally, Chapter 7 concludes the thesis and provides directions for future works.

Capitolo 2

Privacy Issues in the Big Data context

Recently, organizations working in different areas, such as government, banking, medicine, insurance, and so forth, are striving to make their data electronically available. Since these organizations collect data of their users for exploration, analysis, research, or any other purposes, they should also take care of possible privacy-preserving issues.

In general, although various definitions have been provided to define data privacy, there is no accepted standard definition of this concept [10, 11]. Privacy was established as a right in the Universal Declaration of Human Rights [11] in 1948. Data privacy concerns the gathering and the management of personal data, and it can be categorized into content privacy and interaction privacy. Content privacy refers to the prevention of disclosing individuals' identities from anonymized or encrypted databases, such as extracting information from their credit card records stored in a national level database. By contrast, interaction privacy refers to the prevention against the disclosure of a given content concerning an individual, such as checking victims' encrypted web traffic or using voice fingerprints to access services [12].

Many current studies adopt the definition of content and interaction privacy [10, 11] in order to provide privacy-preserving

applications, even when they process and analyse personal data. This can be valuable in boosting the effectiveness of organizations or support prospective plans. However, since big data collections may contain some sensitive data, such applications can threaten the privacy of individuals. Transforming data or anonymizing individuals' data may minimize the utility of the processed data and lead to inaccurate analysis [10]. Hence, numerous endeavours have been devoted to the preservation of privacy, even by considering the application contexts in which they are used.

In this chapter, some privacy-preserving application contexts are surveyed, analysing problems and recent solutions for each of them. Later chapters will describe the new methodologies and tools proposed in this dissertation to tackle them.

2.1 Users privacy awareness in sharing data

Nowadays, data are massively spread on the world wide web. Sensitive data are quickly released by users also because, in most cases, control policies defined over data are not easy to understand or are sidelined. Thus, it arises the necessity to develop methodologies and tools capable of improving the awareness of users concerning the privacy of their personal data. This section describes three real-life scenarios in which users share sensitive data without taking care of privacy threats, and it analyzes some solutions described in the literature. The application contexts described in the following will be those for which new solutions will be described in later chapters of this dissertation.

The application of privacy-preserving policies in the virtual world represents a relevant issue. This is due to the fact that it is difficult to implement forms of control policies on the Internet, considering its information-sharing nature. For example, users have access to many network services, and in order to have complete access to all of their features, they must sign an agreement to share their own sensitive information. Moreover, if users grant consent

to process their data, they usually have no awareness on who and how will manage their data [1]. This could yield different privacy issues, such as the possibility to deal with malicious accounts that might jeopardize users' identities [2]. In this context, several works have been proposed. In [13], authors describe 13 network visualization tools, outlining their advantages and disadvantages. They employ qualitative coding as part of their research design to extract some metrics from the advantages and disadvantages of the described tools. Their purpose is to facilitate the analysts during the construction of evaluation methodologies, which they use to measure the effectiveness of visualization tools through usability studies. In [14], a tool named NetMod is presented. It uses simple analytical models providing designers of large interconnected local area networks with an in-depth analysis of system performances. The tool can be used in environments consisting of thousands of websites. The analytical models and the user interface of NetMod have been tested on a campus-wide network. MVSec is a visual analytics system supporting analysts to understand how to manage information flows over secure networks [15]. The system permits to perform data fusion activities on multiple heterogeneous datasets, aiming to reduce the effort of the data fusion process by means of several visual metaphors. The authors defined multiple coordinated views, providing analysts with multiple visual perspectives to characterize loud events, dig out subtle events, and investigate relations among events in the datasets. Several case studies have been used to demonstrate how the system helps analysts draw an analytical storyline of networking and understand network changes. Finally, in [16] a prototype 3D visualization system for real-time monitoring of both the status of networked devices (wired, wireless, IoT devices) and the network's dynamics (e.g. configuration, load, traffic, abnormal events, suspicious connections, failed IoT devices, etc.) is presented. Through this prototype, users can visualize the current status of the networks of interest simply and intuitively, and from anywhere on the Internet (even from a mobile device). Furthermore, users can receive alerts through short text or instant messages whenever something significant occurs on

the network.

Users tend to use social networks to share information massively; in most cases, they do not care about privatization of their data and are unaware of the privacy threats they can be exposed to. Moreover, users registered on several social networks are even more exposed to privacy disclosure. In a context in which a social network profile contains detailed information that uniquely refers to a specific individual, preserving his/her privacy becomes a fascinating challenge. In this context, several works have been defined. In [17], the authors define a new approach for helping social media users to evaluate their privacy disclosure score (PDS). They assess PDS by taking into account user data shared across multiple social networking sites. Moreover, they highlight sensitivity and visibility as the main points that significantly impact user privacy to derive the PDS for each user. The proposed approach exploits the statistical and fuzzy systems for specifying potential information loss derived from the PDS. In [18], a study based on the "Likes" of users is conducted. It highlighted how a simple "Like" is sensitive content that can be used by both social media and the marketing area to steal information on the users' interests, to propose his/her targeted advertising, and to capture and reconstruct his/her data. In [19], a survey analyzes aspects related to tracking community evolution over time in dynamic social networks. The authors provide a classification of various methods to track community evolution in dynamic social networks. They describe four main approaches by using as a criterion the working principles: i) based on independent successive static detection and matching; ii) based on dependent successive static detection; iii) based on the simultaneous study of all stages of community evolution, and iv) concerns methods directly working on temporal networks. In [20], the authors define two modes of users' private information disclosure behaviour: voluntary sharing and mandatory provision. They exploit the Communication Privacy Management theory to build a framework for explaining the impact of individual characteristics, context, motivation, and benefit-risk ratio on the user's willingness to share their data. They highlight that perceived risk

has less impact on voluntary sharing than previously suggested studies. Finally, a recent study used data from members of social networks to find Multi-SIM subscribers within the same operator or between operators, in order to improve campaigns and churn prediction models of Telecom customers [21].

Another fundamental aspect to be monitored over a social network is the popularity of a profile, witnessed by the number of its followers. A profile with many followers is considered to be influential, and hence it can have a better reputation, attract better paid advertisements, and so on. As a consequence, a common practice of several social network users is to buy fake followers to increase the popularity of their profile, also because they can be bought at a low price. This practice might only aim to support individual vanity, which would be harmless, but if it aims to make an account more reliable and influential, then it might be dangerous. Similarly, spammers could adopt the practice of buying fake followers, aiming to increase their popularity, influence, and to better promote products, trends, fashions, and so on. In this context, several works have been produced. In [22], a framework for fake account detection in online social networks is presented, which relies on Support Vector Machine, Naive Bayes, and Decision tree classification methods. The detection process starts with the selection of the profile to analyze. The second step consists in choosing the suitable attributes (i.e., features) for classifying the selected profile, using the above mentioned machine learning models. The framework extrapolates information useful for creating the training and test datasets, and it compares the classification accuracy of the three above mentioned classification methods, choosing the best performing one. In [23], a methodology for detecting fake accounts on Instagram platforms is proposed. The authors have enumerated the main characteristics to discriminate a fake account from a genuine one. In particular, by manually examining different types of accounts, they have extracted a set of features to highlight malicious accounts' characteristics. Moreover, they have analyzed the liking behaviour of each account to build an automatic mechanism to detect fake likes on Instagram. Finally, they

report an evaluation accuracy of their methodology, highlighting an achieved precision of 83.5%. In [24], authors propose several algorithms and techniques to detect fake profiles, most of which exploit the large volumes of unstructured data generated from social networks. They also provide an exhaustive survey on the existing and latest fake profile detection techniques. In [25], machine learning techniques are used to derive better predictions on fake account identification, based on posts and statuses involving them. The authors experimentally evaluated their methodology by using an SVM classifier on data extracted from the Twitter platform. In [26], authors have performed a qualitative survey that analyses the pros and cons of different detection techniques concerning malicious activities on social networks. They classified different detection approaches with specific analysis of their applicability. Finally, in [27], a survey of recent advancements in the fake account detection methodologies on social networking websites is presented. In particular, the authors summarize the recent development of fake account detection technologies and discuss the challenges and limitations of the existing models. Moreover, they categorize the existing works by focusing on the specific social network analyzed and the technology/methodology employed to discriminate fake accounts. This survey is, undoubtedly, prominent to help future researchers identify the gaps in the current literature and develop a generalized framework for fake profile detection on social networking websites.

This section has discussed problems and solutions concerning privacy issues linked to web browsing activities, social network data sharing, and detection of malicious accounts over social networks. In the next section, the data privacy preservation problem will be analyzed in the data integration context.

2.2 Privacy preserving data integration

Nowadays, there is a need of collecting and integrating a huge amount of data from multiple sources, storing information concerning different real-world domains, such as hospitals, banks, insurance agencies, pharmaceutical companies, government agencies, and so on. Often, in such domains it is required to perform data linkage and integration, aiming at enriching data with additional valuable information, and enabling data analysis processes that are impossible on individual datasets. However, in the task of data integration, the collection of sensitive data at one place makes them vulnerable to re-identification or linking attacks [6]. In general, data integration is a complex process due to: i) the different schema design of the involved data sources, ii) the absence of unique entity identifiers, iii) the necessity of solving attribute level, structure level, and data value conflicts, and iv) the necessity to identify and remove dirty data. Moreover, it is meaningful to solve the schema and data conflicts in a privacy-preserving manner, in order to provide an integrated view of protected data. Privacy-preserving data integration (PPDI) involves the use of schema information, ontology alignments, and personally-identifying information (PII) of individuals to link and match data [28]. Such data need to be protected from re-identification or record linkage attacks [28, 29]. The linking of records from multiple datasets, namely record linkage, has attracted a lot of attention in the PPDI area [30]. Thus, privacy-preserving record linkage (PPRL) mechanisms are used to find correspondences between similar values in multiple datasets and to generate the matching pairs in a secure way [28, 31, 32]. PPRL involves using exact matching and approximate matching paradigms in two-party and multi-party data integration techniques, with or without the use of a trusted third party. Additionally, several efforts have been done in the context of privacy-preserving semantic data integration, which deals with semantic web-enabled record linkage attacks [33]. However, there are still many concerns related to possible attacks to be addressed

in privacy-preserving data integration scenarios [34].

In the PPRL context, several works have been performed. A more recent proposal is the general framework presented in [35], which concerns normalised measures to practically evaluate and compare privacy-preserving record linkage (PPRL) solutions. Furthermore, the authors aimed at comparing the state of the art PPRL techniques, by considering widely used numerical measures, such as scalability, linkage quality, and privacy. Instead, Schnell *et al.* propose the Bloom Filter-based protocol, which aims to guarantee privacy preserving record linkage [36]. It uses encrypted identifiers, and similarity computations of Bloom filters with Hash Message Authentication Codes (HMACs) on a q-grams similarity function [37]. Bit Vectors (BV) is another approach for representing numerical data values in privacy-preserving record linkage [38]. It represents an accurate distance preserving encoding schema for embedding numerical values into a privatization space, in a way that preserves the initial distances. In particular, in order to compute hash values, BV uses extremely simple and computationally cheap operations, instead of expensive cryptographic hash functions. Finally, a Privacy-Preserving Probabilistic Record Linkage (P3RL) methodology has been proposed in [39]. It facilitates the linkage of existing datasets in health-related research settings, and provides a different solution w.r.t. classical cryptographic methods.

This section has discussed problems and solutions concerning privacy preservation in the data integration context. Since this process is often preparatory to machine learning tasks, even adopting privacy preserving techniques at this stage does not always guarantees that the machine learning tasks will not provide some clues that might yield disclosure of sensitive information. Thus, in the next section, the privacy preserving machine learning problem will be analyzed.

2.3 Privacy preserving machine learning

Nowadays, machine learning (ML) has become a core discipline for many popular applications, such as image classification, speech recognition, natural language translation, and so on. In particular, a machine learning model needs data to extrapolate knowledge, like for example in classification activities. Therefore, more quality data are fed to the ML model, more precisely, will be the classification results. In particular, a training dataset is necessary to train and generate a machine learning model, but the dataset itself can contain sensitive samples, e.g., personal medical records, employee information, financial data, etc. An ML model can yield the disclosure of sensitive data, jeopardizing the users' privacy. To this end, it arises the necessity to develop privacy preservation techniques that guarantee the non-disclosure of sensitive data through the application of ML models. To this end, the first step is to develop anonymization strategies preventing a user's identification even if ML is used, for example, to classify his/her personal data.

In this section, anonymization strategies exploiting generalization are analyzed. This will provide the basis for the definition of a new anonymization technique exploiting generalization strategies to anonymize data, which guarantees the possibility to use ML for classification activities. A number of approaches target anonymity specifically within classification contexts. In [40] a k-anonymization algorithm is presented, which aims to find a generalization strategy, not necessarily optimal, in the sense of minimizing data distortion, but preserving the classification structure. In particular, this algorithm masks a dataset through a sequence of refinements, starting from the most generalized state for each attribute and iteratively refines a generalized value until the anonymity requirement over it is violated. Information gain and anonymity loss are used as selection criteria for guiding the process, aiming to heuristically maximizing the classification performances. Similarly, in [41] a Top-Down Specialization (TDS)

approach is presented, aiming to generalize a dataset to achieve anonymity while preserving its usefulness for classification. To this end, anonymity is restricted to a virtual identifier, a combination of attributes for which generalizations are applied in a top-down fashion, by using a generalization taxonomy for categorical attributes and intervals for continuous attributes. TDS employs information gain and anonymity loss as data quality measures to evaluate the achieved generalizations. An approach relying on both suppression and swapping to preserve anonymity in the context of classification is proposed in [42]. It leverages an existing classification tree induction algorithm (i.e., C4.5) trained on the quasi-identifiers of the original dataset and entropy as the data utility measure. k-anonymity is achieved by manipulating its leaves. A more general approach to achieve anonymization within various data mining problems, such as classification, association rule mining, and clustering, is proposed in [43]. This approach constructs a data mining model similarly to the well-known ID3 decision tree induction algorithm [44]. In particular, it iteratively splits the data by selecting, among all attributes, the one maintaining the highest gain (for a specific gain function, e.g., Information Gain or the Gini Index) after such a process. However, independently from the gain, the split is not applied whenever it causes a breach of k-anonymity.

In the next chapter, theoretical foundations concerning data profiling will be provided. They will be helpful to understand the novel solutions described in this dissertation to tackle the privacy problems described in this chapter.

Capitolo 3

Background

This chapter introduces well-known strategies defined in the literature to guarantee information confidentiality and anonymity, which represent the general requirements claimed in a privacy-preserving scenario. Successively, the chapter introduces the theoretical foundations of the data profiling research area, since it provides automated tools to extract metadata from big datasets, which could be potentially used for malicious activities aiming to disclose sensitive data. In particular, the chapter will focus on metadata such as functional dependencies and relaxed functional dependencies, since they are the basis for deriving the new methodologies discussed in later chapters of this dissertation.

3.1 Privacy-preserving Techniques

There are many classical methods to guarantee anonymization and information confidentiality over data. However, in the context of big data they suffer from scalability issues [45]. Moreover, some typical big data processing tasks might yield privacy problems. For instance, data analytics and data integration processes could jeopardize the user privacy. More specifically, when data contain information referring to persons, not all information can be processed without manipulating them through anonymization strategies.

Protecting personally identifiable information (PII) is increasingly difficult, because the data are shared too quickly.

To eliminate privacy concerns, the agreement between the company responsible for managing data and the data owner must be determined by policies. Personal data must be anonymized (de-identified), and transferred into secure channels [46]. However, the user identity can be uncovered through the application of complex analysis processes, such as the ones performed through artificial intelligence techniques. In fact, the predictions resulting from these analysis processes could lead to unethical issues. In the following sections, the anonymity and information confidentiality issues are presented, by also describing several well-known strategies that can be applied into different application scenarios.

3.1.1 Information confidentiality

Data privacy concerns several aspects. Among them, this section focuses on information Confidentiality (IC). The latter is a general privacy-preserving concept, which requires to preserve the confidentiality of specific users' data, also referred to as sensitive data, aiming to protect them against unauthorized accesses [47].

In general, cryptography and perturbation models are the most frequently used techniques to guarantee information confidentiality. Nevertheless, confidentiality issues can become specific depending on the application domain on which data are used. Thus, different solutions can be found in the literature w.r.t. the application domain they consider. For instance, in the data mining context, the Framework for Accuracy in Privacy-Preserving mining (FRAPP) represents a generalized model for random perturbation-based methods, operating on categorical data under strict privacy constraints [48]. Instead, guaranteeing privacy preservation in the context of big data processing tasks adds further complexity, since there might be several complex operations potentially yielding privacy threats. A typical example is represented by data integration, since data integrated from several data sources might partially or totally imply obscured data [6, 29]. An essential pioneering work

in this context was carried out by Dalenius T. in 1986 [49]. The author shows how it is possible to identify unique records in a dataset by merely sorting them. This simple idea turns out to be highly effective, because it permits to discover unique records. A more recent proposal is the general framework presented in [35], which concerns normalized measures to practically evaluate and compare privacy-preserving record linkage (PPRL) solutions. Furthermore, the authors aimed at comparing the state of the art PPRL techniques, by considering widely used numerical measures, such as scalability, linkage quality, and privacy. Instead, Schnell *et al.* propose the Bloom Filter-based protocol, which aims to guarantee privacy-preserving record linkage [36]. It uses encrypted identifiers, and similarity computations of Bloom filters with Hash Message Authentication Codes (HMACs) on a q-grams similarity function [37]. Bit Vectors (BV) is another approach for representing numerical data values in privacy-preserving record linkage [38]. It represents an accurate distance preserving encoding schema for embedding numerical values into a privatization space, in a way that preserves the initial distances. In particular, in order to compute hash values, BV uses extremely simple and computationally cheap operations, instead of expensive cryptographic hash functions. Finally, a Privacy-Preserving Probabilistic Record Linkage (P3RL) methodology has been proposed in [39]. It facilitates the linkage of existing datasets in health-related research settings, and provides a different solution w.r.t. the classical cryptographic methods.

3.1.2 Anonymization

The collected data should be treated as a private table that encompasses multiple records [50]. Each record (row) represents a single user and comprises several attributes that are specific to a particular individual [51]. These attributes can be categorized into three classes [52]: Identity attributes (IA) that explicitly identify the records of a user (e.g., name, mobile phone number, social security number, and driver's license number); quasi-identifier (QI)

attributes, denoting a sequence of non-explicit attributes referring to individuals (e.g., race, age, date of birth, ZIP code, and gender), which can potentially identify records related to them; and finally, sensitive attributes (SA) that contain confidential data of individuals (e.g., salary and disease) [53].

In general, strategies aiming to solely eliminate attributes (IAs), which explicitly identify users from the table before disclosing them, have been demonstrated to be inefficient [52]. In fact, identification attributes (QIs), representing a set of non-explicit attributes referred to individuals, should be anonymized before releasing the data [10].

The following sections describe one of the most frequently used technique to anonymize QIs, i.e., k-anonymity, together with its extensions, namely L-diversity and T-closeness.

3.1.2.1 K-anonymity approach

k-anonymity is an extensively recognized privacy-preserving technique [54]. The idea underlying the k-anonymity strategy is based on the modification of the values of the QI attributes, in order to limit an attacker in detecting the identity of persons stored in a particular dataset, while the released data remain as useful as possible [55]. The k value is used as a measure of privacy. Thus, the lower the k value, the higher will be the probability of deanonymizing the data. Conversely, if the k value is high, then an attacker will have more difficulty unraveling the identity of individuals. However, increasing the k value will simultaneously lower the usefulness of the data [10].

Although the k-anonymity strategy permits to guarantee a certain level of privacy, it also has some limitations. Firstly, k-anonymization-based techniques will have difficulty in identifying the QI attributes over external datasets, and determining the extent to which information can be disclosed to others [56]. In fact, recent studies [51] have shown that approximately 87% of the population can be distinctly recognized by using the seemingly innocuous QI attributes. Moreover, in previous studies [12], a mobility

dataset has been collected for 1.5 million people, and a basic anonymization operation has been applied (eliminating apparent ID attributes). Nonetheless, these studies were able to identify a person with 95% of precision by only using four spatio-temporal points. This result has been confirmed by a recent study [12], which analyzed a data set of 90-day financial dealings of over 1 million persons. In particular, it demonstrated that four spatio-temporal points permit to effectively re-identify approximately 90% of the persons.

The k-anonymity approach attempts to work on the attributes of QI, which involves identifying a person's age, gender, and ZIP code, with no investment in the sensitive attributes [12]. Hence, the k-anonymity-based method is subjected to indirect attacks that enable the possibility of precisely deducing the features of an individual, thereby leading to the disclosure of identity. Examples of such an attack are homogeneity attack and background knowledge attack, which are based on the following aspects: an opponent has sufficient background knowledge from the relationship between sensitive and QI attributes to conduct probabilistic attacks [12], or when the QI attributes are connected with other public datasets, it is possible to aid an adversary in disclosing the identities and other sensitive attributes of individuals [12, 56]. In addition, with k-anonymity-based techniques information loss is unavoidable when attempting to attain a high level of privacy [57]. Thus, such techniques can possibly affect the use of data, like for example yield the production of imprecise or even impractical knowledge throrugh data mining processes. In general, providing a good balance between privacy and utility is essential in big data applications.

3.1.2.2 L-diversity approach

L-diversity was designed by Machanavajjhala et al. [53] to protect individuals from possible disclosure of sensitive attributes after the application of a k-anonimity strategy [54]. This approach is considered an extension of k-anonymity. The primary aim of L-

diversity is to preserve privacy by increasing the diversity of sensitive values. This technique entails to treat the values of a specific attribute declared as sensitive, regardless of its distribution in the data. Thus, for each sensitive attribute it prescribes to have at least l distinct values within each equivalence class provided by the application of a k-anonymity strategy [58].

The major drawback of the L-diversity approach lies in the distribution of values of such sensitive attributes, because different values can have extremely different levels of sensitivity. This poses serious privacy risks, e.g., enabling an opponent to deduce a value according to the probability of choosing it from the attribute distribution. This kind of attack is referred to as skewness attack [50]. In addition, this approach is inadequate to prevent the disclosure of attributes against similarity attacks (in an equivalence class, although the values of the sensitive attribute are different, they can be semantically similar). An opponent can easily have access to the sensitive attribute because the knowledge of the global distribution of this attribute is available to the opponents. L-diversity guarantees the diversity of sensitive values in every group but does not consider their semantic proximity. This drawback motivated the development of the T-closeness approach [50, 58].

3.1.2.3 T-closeness approach

T-closeness was presented by Li et al. [50] as an extension of the l-diversity group-based anonymization, which is commonly used to protect privacy in datasets. In this approach, the value distribution of sensitive attributes in any equivalence class should be similar to the attribute distribution in the whole table. For example, the distance between the two distributions should not exceed the threshold t [50].

3.2 Data Profiling

Data Profiling is an important and frequent activity performed by IT professionals and researchers. It represents the process of examining the data available in an existing data source and collecting statistics and information about that data [59].

Data profiling encompasses a vast set of methods to examine datasets and produce metadata. Among the simpler tasks there are statistics, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its values. Metadata might involve multiple columns, like in the case of inclusion or functional dependencies. More advanced techniques detect approximate properties or conditional properties of the dataset at hand. Figure 3.1 shows a classification of data profiling tasks.

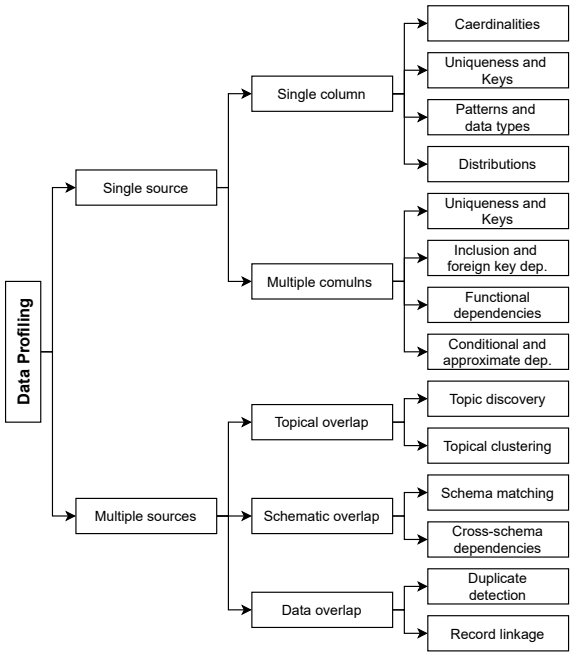


Figura 3.1: Data profiling tasks.

Systematic data profiling, i.e., profiling beyond the occasional

exploratory SQL query or spreadsheet browsing, is usually performed by dedicated tools or components, such as IBM's Information Analyzer [60], Microsoft's SQL Server Integration Services (SSIS) [61], or Informatica's Data Explorer [62]. Their underlying approaches follow the same general procedure: A user specifies the data to be profiled and selects the types of metadata to be generated. Next, the tool computes in batch mode the metadata by using SQL queries and/or specialized discovery algorithms. Depending on the volume of the data and the selected profiling results, this step can last minutes to hours. The results are usually displayed in a vast collection of tables, charts, and other visualizations to be explored by the user. Typically, the profiling results can then be translated into constraints or rules that are then enforced in a subsequent cleansing/integration phase. The need to profile a new or unfamiliar set of data arises in many situations. Often, data profiling is preparatory for some subsequent tasks. In what follows, a summary of use cases for data profiling activities are is provided.

Query optimization. Most database management systems perform basic profiling task to support query optimization with statistics about tables and columns. These profiling results can be used to estimate the selectivity of operators and the cost of a query plan.

Data cleansing. Probably, this is one the most typical use cases for data profiling. In particular, the metadata extracted with data profiling can help reveal data errors, such as inconsistent formatting within a column, missing values, or outliers. Profiling results can also be used to measure and monitor the general quality of a dataset.

Data integration. Often the datasets to be integrated need to be inspected to understand how to perform the data integration process. For example, for improving the data integration process it is necessary to know the size of the dataset, the data types, the semantics of columns and tables, the dependencies holding over the dataset, and so on. Moreover, the vast abundance of open data that can potentially be integrated with enterprise data has

amplified this need.

Scientific data management. The management of data that is collected during scientific experiments or observations has created additional motivations for the efficient and effective usage of data profiling. When importing raw data (e.g., from scientific experiments or extracted from the Web) into a DBMS, it is often necessary and helpful to profile the data, aiming to derive a suitable schema for them.

Data analytics. A profiling step should precede almost any statistical analysis, or data mining task, in order to help the analyst understand the data at hand and appropriately configure tools, such as SPSS [63] or Weka [64]. Pyle describes the detailed steps that must be performed for analyzing and subsequently preparing data for data mining [65].

The knowledge about data types, keys, foreign keys, and other constraints can support data modelling and help keeping data consistent, improve query optimization, and reap all the other benefits of structured data management. Query formulation and indexing [66], scientific discovery [67], and database reverse engineering [68] provide further motivations for data profiling.

An exciting aspect of data profiling is that its activities, especially those searching data correlations, i.e., functional dependencies [69] and their extensions [70], can be used to detect possible privacy threats, yielding the possibility of exploiting them to improve privacy in data mining and data integration processes. The next section introduces concepts strictly linked to data profiling, which will be helpful to understand the privatization strategies underlying the solutions proposed in this dissertation.

3.2.1 Functional dependencies & Relaxed functional dependencies

As said above, functional dependencies (FDs) and relaxed functional dependencies (RFDs) represent some of the most relevant profiling metadata. This section details concepts concerning func-

tional FDs and relaxed RFDs. To this end, in what follows we first need to recall some basic concepts of relational databases.

A relational database schema $\mathcal{R}$ is defined as a collection of relation schemas (R_1,..., R_n), where each R_i is defined over a set $attr(R_i)$ of attributes (A_1,..., A_m). Each attribute A_k has associated a domain $dom(A_k)$, which can be finite or infinite. A relation instance (or simply a relation) r_i of R_i is a set of tuples such that for each attribute $A_k \in attr(R_i)$, $t[A_k] \in dom(A_k)$, $\forall$ $t \in r_i$, where $t[A_k]$ denotes the projection of t onto A_k. A database instance r of $\mathcal{R}$ is a collection of relations (r_1,...,r_n), where r_i is a relation instance of R_i, for $i \in [1, n]$.

In the context of relational databases, data dependencies have been mainly used to define data integrity constraints, aiming to improve the quality of database schemas and to reduce manipulation anomalies. There are several types of data dependencies, including functional, multivalued, and join dependencies. Among these, functional dependencies (FDs) are the most commonly known, mainly due to their use in database normalization processes. Since RFDs extend FDs, let us recall the definition of FD.

Definition 1. (Functional dependency). *A functional dependency (FD) φ, denoted as $X \rightarrow Y$, between two sets of attributes $X, Y \subseteq attr(\mathcal{R})$, specifies a constraint on the possible tuples that can form a relation instance r of $\mathcal{R}$; requiring that $X \rightarrow Y$ holds on r iff for every pair of tuples (t_1, t_2) in r, if $t_1[X] = t_2[X]$, then $t_1[Y] = t_2[Y]$. The sets X and Y are also called Left-Hand-Side (LHS) and Right-Hand-Side (RHS), resp., of φ.*

RFDs extend FDs by relaxing some constraints of their definition. In particular, they might relax on the *attribute comparison* method, and on the fact that the dependency must hold on the entire database.

Relaxing on the attribute comparison method means to adopt an approximate operator to compare tuples, instead of the "equality" operator. In order to define the type of attribute comparison method used within an RFD, the concept of *constraint* is used [71].

Definition 2. (Constraint). *A constraint ϕ is a predicate evaluating whether the similarity/distance, or the order relation, between two values of an attribute A falls within a predefined interval.*

Thus, a constraint depends on a similarity/distance function, or an order relation, defined on the attribute domain, plus one or more comparison operators with associated threshold values defining the feasible intervals of values.

An example of constraint ϕ defined on the attribute `Address` and the edit distance ED function could be: $0 \leq ED(addr1, addr2) \leq \varepsilon$, where $addr1$ and $addr2$ are two address values, whereas 0 and ε are two given threshold values.

Definition 3. (Set of constraints). *Given a set of attributes $X = \{A_1, \ldots, A_k\}$, a set of constraints $\Phi = \{\phi_1, \ldots, \phi_k\}$ defined on them represents a collection of constraints that are applied to $\{A_1, \ldots, A_k\}$, respectively.*

A functional dependency holding on "almost" all tuples or on a "subset" of them is said to relax on the extent [69]. In the case of "almost" all tuples, a *coverage measure* should be specified to quantify the degree of satisfiability of the dependency. Whereas, in the case of "subset" (*constrained domain* in the following), conditions on the attribute domains should be specified to define the subset of tuples satisfying the dependency.

Definition 4. (Coverage measure). *A coverage measure Ψ on φ, $\Psi : dom(X) \times dom(Y) \rightarrow \mathbb{R}^+$, quantifies the amount of tuple pairs in r satisfying φ.*

As an example, the *confidence measure* introduced in [67] evaluates the cardinality of the greatest set of tuples $r_1 \subseteq r$ for which φ holds in r_1.

Several coverage measures can be used to define the satisfiability degree of an RFD, but usually they return a value normalized on the total number of tuples n, with n cardinality of r, so producing a value $v \in [0, 1]$.

Definition 5. (Constrained domain). *Given a relation database schema $\mathcal{R}$ with attributes $\{A_1, \ldots, A_k\}$ defined on domains $\{dom(A_1), \ldots, dom(A_k)\}$ respectively, $dom(A_1) \times dom(A_2) \times \cdots \times dom(A_k) = dom(\mathcal{R})$, respectively, and let c_i be a condition on $dom(A_i)$, $i = 1 \ldots k$, the constrained domain $\mathbb{D}_c$ is defined as follows*

$$\mathbb{D}_c = \left\{ t \in dom(\mathcal{R}) | \bigwedge_{i=1}^{k} c_i(t[A_i]) \right\}.$$

Constrained domains enable the definition of tuple "subsets" on which a functional dependency holds.

Then, a general definition of RFD can be given:

Definition 6. (Relaxed functional dependency). *Let us consider a relational schema $\mathcal{R}$. An RFD ϱ on $\mathcal{R}$ is denoted by*

$$\left[X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} Y_{\Phi_2} \right]_{\mathbb{D}_c} \tag{3.1}$$

where

- $\mathbb{D}_c$ is the constrained domain that filters the tuples on which ϱ applies;
- $X, Y \subseteq attr(\mathcal{R})$, with $X \cap Y = \emptyset$;
- Φ_1 and Φ_2 are sets of constraints on attribute sets X and Y, respectively;
- Ψ is a coverage measure defined on $\mathbb{D}_c$;
- ε is a threshold, with $0 \leq \varepsilon \leq 1$.

Given $r \subseteq \mathbb{D}_c$, a database instance r on $\mathcal{R}$ *satisfies* the RFD ϱ, denoted by $r \models \varrho$, if and only if: $\forall\ (t_1, t_2) \in r$, if Φ_1 is true for each constraint $\phi \in \Phi_1$, then *almost always* Φ_2 is true for each constraint $\phi' \in \Phi_2$. Here, *almost always* means that $\Psi(X, Y) \geq \varepsilon$.

In other words, if $t_1[X]$ and $t_2[X]$ agree with the constraints specified by Φ_1, then $t_1[Y]$ and $t_2[Y]$ agree with the constraints

specified by Φ_2 with a degree of certainty (measured by Ψ) greater than ε.

Based on definition (3.1), the canonical FD $X \rightarrow Y$ can also be written as:

$$\left[X_{\text{EQ}} \xrightarrow{\Psi_1} Y_{\text{EQ}}\right]_{\mathbb{D}_{\text{TRUE}}} \tag{3.2}$$

where TRUE is a sequence of tautologies, $\mathbb{D}_{\text{TRUE}} = dom(\mathcal{R})$, EQ is the equality constraint, and Ψ_1 represents the fact that the dependency must hold on all tuples of the instance r (i.e., $\Psi(X, Y) = 1$, and $\varepsilon = 1$).

Example 1. *Let us consider a database of the census income, containing the following data of citizens:* `Name`*,* `Surname`*,* `SSN`*,* `Age`*,* `Address`*,* `Native-Country`*,* `Occupation`*, and* `Sex`*. According to this, it is likely to have the same* `Native-Country` *for costumers having the same* `Name` *and* `Surname` *thus, an FD* `Name, Surname` $\rightarrow$ `Native-Country` *might hold. However, the names, surnames, and countries might be stored by using different abbreviations/-variations, and/or typos may have been introduced during tuple insertion operations. Thus, the following RFD might hold:*

$$\left[\texttt{Name}_{\approx}, \texttt{Surname}_{\approx} \xrightarrow{\Psi_1} \texttt{Native-Country}_{\approx}\right]_{\mathbb{D}_{\text{TRUE}}}$$

where $\approx$ is the string similarity function. On the other hand, few cases of homonyms for the customers have to be considered. For this reason, the previous RFD should also admit exceptions. This can be modeled by introducing a different coverage measure to make the RFD relax on the extent:

$$\left[\texttt{Name}_{\approx}, \texttt{Surname}_{\approx} \xrightarrow{\psi(X,Y)\geq 0.90} \texttt{Native-Country}_{\approx}\right]_{\mathbb{D}_{\text{TRUE}}}$$

Capitolo 4

Privacy awareness in Social Networks and Web browsing

This chapter highlights several privacy issues related to social network participation and web browsing. In particular, an experiment shows how it is possible to derive sensitive data when users are registered on different social network platforms. Moreover, a visual metaphor is implemented in a tool showing how user data are tracked during web browsing.

Nowadays, information is spread over different network channels, and in most cases, users are unaware of how their sensitive data are managed and shared. For example, users have access to many network services, and in order to have a complete access to all their features, they must sign an agreement to share their sensitive information. Moreover, if users grant the consent to process their personal data, they usually have no awareness concerning who and how will manage these data. This also occurs with online social platforms, such as WeChat, Facebook, LinkedIn, and Twitter, which have become extremely popular, involving many real social relationships. In this context, effective privacy-preserving methods should ensure both the privacy of data and their availability, by also limiting the possibility for users to deal with malicious

accounts, since they might jeopardize their identities. In fact, the influence and popularity of users play a fundamental role, because many people follow influential accounts and tend to trust them. However, what happens if the followed trusted account is fake? Indeed, although the practice to artificially "amplify" the number of followers by means of fake accounts could be used to make an account more trustable and influential, it could be also adopted by spammers aiming to create disinformation, to promote advertising campaigns, and to steal personal data of individuals.

To overcome these problems, in what follows new approaches for improving users' awareness concerning privacy threats in Web browsing or in social network activities are proposed. In particular, the first work presents a visual metaphor based tool supporting users in understanding how their data are shared among network service providers while they perform internet browsing. The second work aims to analyse users' data extracted from several social networks to evaluate privacy preservation. Finally, a novel methodology is proposed to detect fake accounts in social networks, by understanding the reliability of users with social network profiles.

4.1 Enhancing user awareness during internet browsing

With the advent of e-commerce and social networks, people often unconsciously disseminate their sensitive data through different platforms such as Amazon, eBay, Facebook, Twitter, Instagram, and so on. In this scenario, it would be useful to support users with tools increasing their awareness on how their sensitive data are exchanged.

This section presents a visual analytics tool enabling users to understand how their sensitive data are exchanged or shared among different network services. In particular, the proposed tool visualizes the communication flow generated during Web browsing activities, highlighting the providers tracking their data. The tool provides a real-time summary graph showing the information ac-

quired from the network. A user study is presented to highlight how the proposed tool improves the user's perception of privacy issues.

Web browsers are the primary tools to access the Internet, and can also be extended by adding new features in order to facilitate users during Web searches, or to monitor the content of Web pages, i.e., to block malicious content (see Section 2.1). However, the monitoring of data that users spread over the Internet is a complex task, especially when sensitive data are involved. In addition, making users aware of possible privacy threats related to the spreading of their sensitive data requires the definition of user-friendly approaches. These should intuitively highlight all aspects related to privacy, and dynamically support users in the comprehension of possible privacy threats. To this end, in the proposed visual analytics tool, an interactive graph structure is exploited to enable users to easily track their data. Furthermore, this visual metaphor allows users to have a visual chronology of their browsing web activities, letting them understand how different network services communicate with each other by sharing user data.

4.1.1 Overview of the system components

This section presents the technologies used for building the proposed visual analytics tool, by also explaining the design choices and how the technologies used are connected. Then, the network sniffer underlying the tool is described, how it interacts with data streams through ReThinkDB, and how the sniffed packets have been used to compose the proposed visual analytics tool.

4.1.1.1 Network sniffer

This component relies on Scapy, which is a Python program enabling users to forge, dissect, emit, or sniff network packets, probes, scans, or network attacks [72]. It provides a DSL (Domain Specific Language) that enables a powerful and fast description of any kind of packet. Scapy allows users to describe a package or a set

of packages as layers that overlap on each other. The fields of each level have useful default values that can be overloaded. Scapy does not force the user to employ predefined scenarios or templates.

Scapy is involved in the receiving phase only, which means that every time a user wants to send packets, a different tool must be involved. When a user probes a network, s/he will send many stimuli, and some of them will be answered. If the user chooses the right stimuli, it is possible to obtain the necessary information from the responses or the lack of responses. Unlike many tools, Scapy will give the user all the information, i.e. all the stimuli s/he sent, and all the responses s/he got. For example, it is possible to probe a TCP port scan, visualize the data in terms of results of a port scan, and then decide whether to visualize the TTL of the response packet. It is not necessary to perform a new analysis every time the user wants to view other data. A common problem in network probing tools is that they try to interpret the answers they got instead of only decoding and giving facts. Saying something like "I received a TCP Reset on port 80" is not subject to interpretation errors. Saying "The port 80 is closed" is an interpretation that can be right most of the times, but wrong in some specific contexts. For instance, some scanners tend to report a filtered TCP port when they receive an ICMP destination unreachable packet. This may be right, but in some cases, it means that the packet was not filtered by the firewall, and there was no host to forward the packet to.

Scapy has been used to create an acquisition network script, which is always listening to all possible actions that the user performs during his/her network activities. The script intercepts all network packets and applies a filter on the communication protocol, selecting only the https and HTTP protocols, and their cookies.

4.1.1.2 Managing Data Streams

Data streams are managed by means of ReThinkDB, which is a flexible, and scalable database for Web and server development. Like other NoSQL databases, it keeps user data in-sync across client apps through realtime listeners and offers offline support for mobile and Web applications so that the user can build responsive apps, working regardless of network latency or Internet connectivity.

The ReThinkDB data model supports flexible, hierarchical data structures. The information is stored in JSON documents, organized into collections. Documents can contain complex nested objects in addition to sub-collections. In this type of database, the user can submit queries to retrieve specific documents or all the documents in a collection matching matches his/her query parameters. User queries can include multiple chained filters, and combine filtering and sorting. In particular, ReThinkDB uses data synchronization to update data on any connected device. However, it's also designed to make simple and efficient one-time fetch queries. It offers automatic, multi-region data replication, strong consistency guarantee, atomic batch operations, and realtime transaction support. It is a local and/or cloud-hosted NoSQL database, providing access to iOS, Android, and Web apps employing native SDK.

4.1.1.3 Visualization

This section presents the technologies used for implementing the interactive visual interface of the proposed tool. In particular, the Web application uses D3.js for building interactive graphics. The latter is a JavaScript library for manipulating data-oriented documents[1]. It achieves visual representations through data loading, data binding, and analytic transformation elements. Unlike graphs generated using Excel, the ones obtainable through D3.js

[1]https://d3js.org/

enable developers to define customized mapping rules. Also, physics and force simulations provide a more dynamic representation for the user.

Depending on their needs, developers can determine the mapping values to the graphics, such as display color, size, and so on. A D3-based graph is graphically displayed on a Web page by using CSS3, HyperText Markup Language, and Scalable Vector Graphics. It allows dealing with SVG, which is the World Wide Web Consortium specification providing a network vector graphics standard [73]. SVG strictly abides by XML syntax and uses a textual description language to annotate image contents. It is a resolution-independent vector and an image graphic format.

4.1.1.4 The overall Architecture

The architecture of the proposed visual analytics tool is shown in Figure 4.1. In detail, at a lower level there are the *Network Sniffer Module* and the *Browsermod Proxy*, which are both responsible for capturing the network traffic. The modules communicate with both the *Data Parser* and the *Database Connector Module*, in order to reorganize and filter the raw data and store them into the *Real-Time Local Database*, which has been implemented by using ReThinkDB. In particular, although the latter is only used to store data caught on the Web, it offers security mechanisms to prevent unauthorized access. Moreover, Web browsing is provided through the *Selenium Driver Module*, which embeds several useful functionalities to manage data retrieved from websites. Furthermore, a Node.js server queries the *Real-Time Local Database* and represents the received data in terms of the proposed metaphors, by using the real-time graphic library (D3.js) presented above. Finally, the structured visual metaphors are presented to the user via the *UI Data Visualization* module.

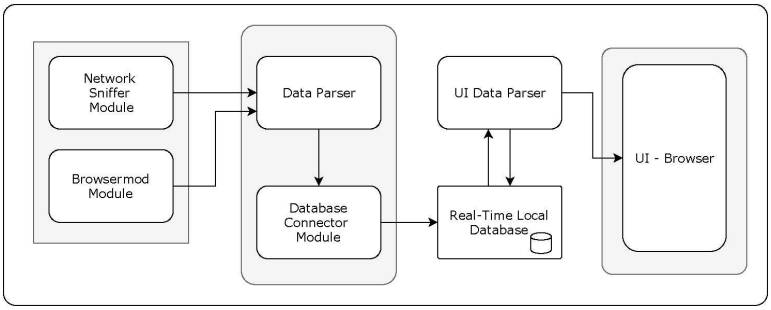


Figura 4.1: System Architecture.

4.1.2 The Vipat tool

This section describes the proposed visual analytics tool, named VIPAT, showing its main functionalities. Figure 4.2 shows a sample of network graph that VIPAT generates to make the user aware of the providers that are collecting his/her sensitive data. The graph is constructed in real-time so that the user can analyze the changes occurred when interacting with different providers. In other words, when a user browses websites, with different network providers managing his/her personal information, VIPAT captures such information by creating a new node in the graph, as shown in Figure 4.2.

The azure node within the connected graph represents the user node, the green nodes represent providers using the HTTPS protocol, the red ones represent providers using the HTTP protocol, and the yellow ones represent advertising hosts. Moreover, it is possible to navigate the graph in order to understand how the data are shared by different network providers.

On the bottom of the visual interface other two interactive charts are provided:

(i) The first one (Figure 4.3) represents an interactive bar chart showing the top 10 network providers that are taking more advantage of user information, in terms of amount of packets in which they are involved;

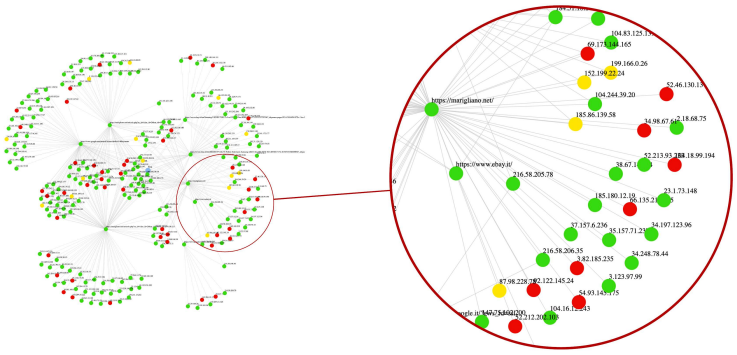


Figura 4.2: VIPAT network graph representing the information spread on the network during user's browsing sessions.

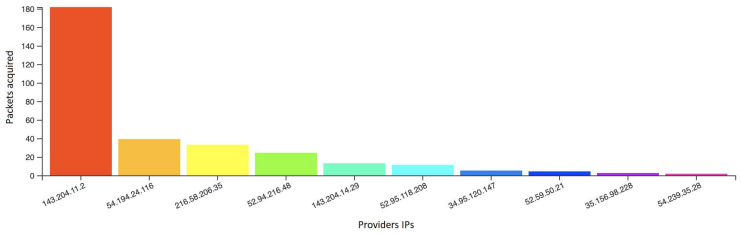


Figura 4.3: Most contacted providers during user's session browsing.

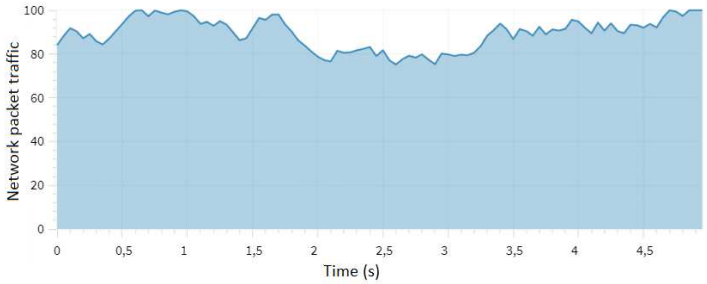


Figura 4.4: Frequency of the network packets exchanged each 0.5s.

(ii) The second one (Figure 4.4) shows a chart representing the frequency rate at which packets are transferred by service providers on the network each half a second.

Figura 4.5: Geolocation of the packet stream worldwide.

These information allow users to increase their awareness concerning the frequency by which network providers manage their information. The interaction with the charts allows users to retrieve additional information related to the selected network provider. Finally, Figure 4.5 shows the interface representing a linked globe, in which the connections established in the different parts of the world during user Web browsing activities are drawn. In particular, this additional information enables the user to track network communications, so that is possible to understand where the network providers are localized.

As it can be seen the the presented interface, simple visual metaphors have been adopted, so that the user can have a complete overview of the network environment during his/her interaction with VIPAT, and be aware of how the different network providers manage his/her information. In fact, one of the most important characteristics of VIPAT is that the user can interact with all the graphs, by obtaining more specific and/or additional information based on the performed selections.

4.1.3 Experimental Evaluation

This section presents the results of a user study whose aim was to evaluate how VIPAT increases the users' awareness about the automatic acquisition of personal data by external providers.

4.1.3.1 Experimental Settings

The evaluation session has been performed by involving twenty users of different ages in a supervised experiment. Users had different education levels, including high school diploma, master degree, and Ph.D. The user evaluation has been performed in a research laboratory, where users accessed to pre-configured computers having VIPAT installed.

In the first step, users filled an initial questionnaire, after which they started performing several Web browsing tasks. In particular, the initial questionnaire contained questions concerning the user's ability to interact with Web applications and other questions related to personal information like age, gender, social occupation, and so on. These questions aimed to asses the following aspects: (1) how much users concern about security and privacy issues, (2) what behavior the users adopt for managing cookies while visiting a Web page, and (3) what data generated during the Web browsing activities users think they own.

Web browsing tasks have been designed to make a user aware of the spread of his/her own data while performing browsing activities. Five tasks have been defined to be executed by each user, all including actions to perform on Web sites. As an example, a user task included browsing a shopping site and simulating a purchase. Another task required to browse a site of news, and to read the daily news. All users have been supervised during their Web browsing tasks. Moreover, in order to monitor the users' interaction with VIPAT, the following behavioural statistics have been monitored: (*i*) how many times each user opened VIPAT, and (*ii*) how many times s/he interacted with it. After users completed their Web browsing tasks, it has been asked them to fill out a further questionnaire. The latter included questions concerning user awareness on how his/her personal information is acquired by external providers during Web browsing. In particular, the questions permitted to evaluate the opinion of the users about whether browsing data are (*i*) safely handled, (*ii*) exploited for commercial purposes, and (*iii*) frequently exchanged among providers.

For each question, a five-point Likert-type answer format has been provided, ranging from "strongly disagree" (coded as 1) to "strongly agree" (coded as 5). The initial and the post-task questionnaires share several questions, aiming to monitor whether the users' privacy perception changed after using VIPAT.

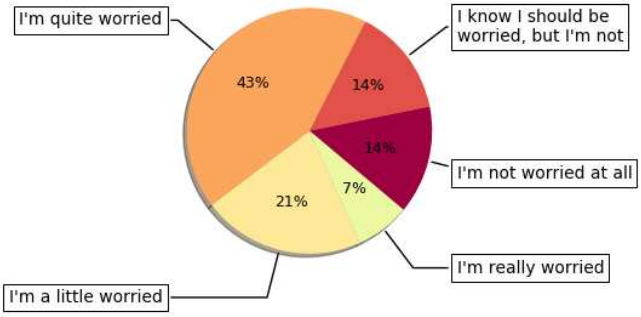


Figura 4.6: Statistics about the concerns of users on security and privacy issues.

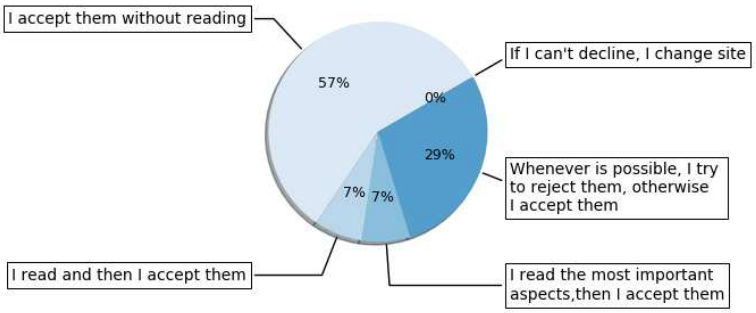


Figura 4.7: Statistics about the behaviour of the users when dealing with cookies' policies.

4.1.3.2 Results

Figure 4.6 shows the users' statistics collected about their concerns on security and privacy issues. It is worth noting that most of the

users are quite worried about them and that three of them indicated that they were not worried, although they have knowledge about possible risks. Figure 4.7 shows the users' statistics about the behavior they adopt with cookies' policies. Most of the users accept them without reading their content. Moreover, no user is willing to leave the website when it is not possible to decline the policy.

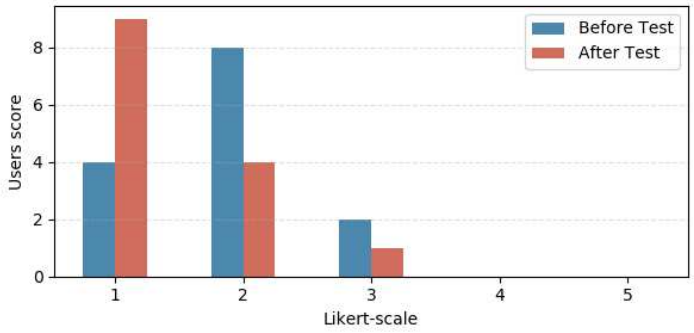


Figura 4.8: Statistics about users' awareness on the actual ownership of Web browsing data.

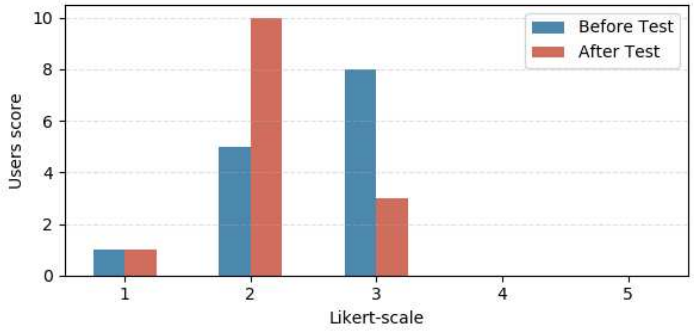


Figura 4.9: Statistics about users' opinions on how much Web browsing data are safely managed.

Figure 4.8 shows the statistics about the opinions of the users on the exclusive ownership of the data generated during Web brow-

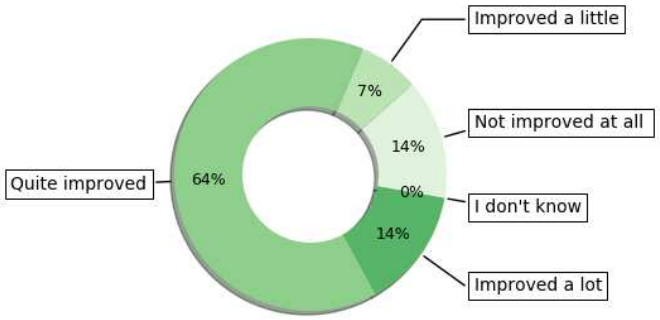


Figura 4.10: Answers related to the users' awareness on data privacy.

sing. This question is contained in both questionnaires, hence the figure highlights how the users' opinions changed after using VIPAT. Figure 4.9 shows the users' answers regarding their opinion about the safe management of navigational data on the Web. Also, this question is contained in both questionnaires, and it is possible to observe that VIPAT increased the perception of insecurity about Web data management.

Figure 4.10 shows the users' answers to the question of how VIPAT changed their data privacy awareness. Most of the users considered VIPAT quite helpful in improving their perception of data privacy. In particular, the majority of users, especially those without computer science degrees, gained consciousness on how their data were exchanged without them being aware of that. What they considered really surprising was the fact that during navigation many different providers were communicating with each other, via a third ads-labeled provider, trying to acquiring information on what the user did previously, in order to show a more incisive (and profitable) advertisement.

Figure 4.11 shows statistics about the users' interactions with VIPAT. The plot highlights how many times users need to get information on the navigation data, looking quickly at or interacting with VIPAT. As can be observed from the results, the number

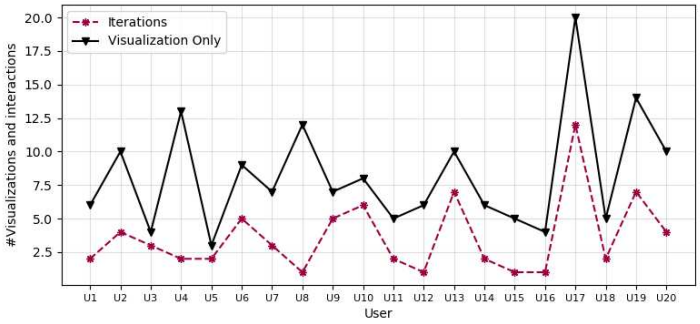


Figura 4.11: Users' interactions with VIPAT.

of visualizations is quite different among users. There have been users that very rarely looked at VIPAT (they looked at VIPAT a number of times equal to the number of tasks assigned to them) and others (e.g., User 17) that felt in need to observe the tool a conspicuous amount of times. On average the users interacted with VIPAT 3.6 times during the experimental session, usually at the beginning of the experiment, as the first Web page was visited, and after all tasks were completed.

4.2 Social network data analysis to highlight privacy threats

Social networks are a vast source of information, and they have been increasing impact on people's daily lives. They permit to share emotions, passions, and interactions with other people around the world. While enabling people to exhibit their lives, social networks guarantee their privacy. The definitions of privacy requirements and default policies for safeguarding people's data are the most difficult challenges that social networks have to deal with.

Aiming to analyse privacy requirements offered by social networks, in what follows an evaluation is performed by collecting

data concerning people who have different social network profiles. To this end, a tool exploiting image-recognition techniques to recognise a user from his/her picture has been built, with the aim of collecting his/her personal data accessible through social networks where s/he has a profile. A dataset of 5000 users has been composed, by combining data available from several social networks, and by comparing social network data that must be mandatorily provided in the registration phase, publicly accessible data, and those retrieved by the performed analysis. The goal is to analyse the amount of extrapolated data, aiming to highlight privacy threats when users share information on different social networks, in order to help them be aware of these aspects.

This work shows how users data on social networks can be easily retrieved, which represents a clear privacy violation. This study aims to improve the user's awareness concerning the spreading and managing of social networks data. To this end, all the statistical evaluations made over the gathered data unleashing privacy issues have been highlighted.

4.2.1 Methodology

This section describes the proposed methodology by summarising it in two meaningful steps: the single- and the cross-social data extrapolation steps.

In the single social data extrapolation step (Figure 4.12), the picture and the name of the user are exploited as the input of the Social Module. The latter performs specific operations only over a single social network, beginning with a search of the user target, by exploiting the photos and the name associated with his/her profile. The face recognition module tries to find a match between the discovered photos and the initial user's picture. If a match is found, the social network module yields all the user profile data available on his/her specific social network.

The idea of exploiting a face recognition module is justified from the fact that it is used to avoid the homonymies on user's

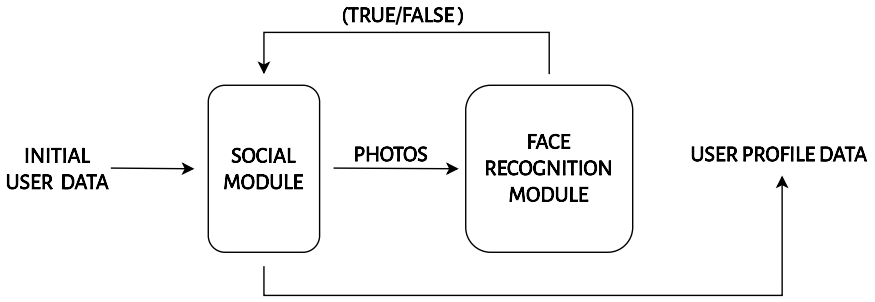


Figura 4.12: Single social data extrapolation step.

names. In Figure 4.12, it is possible to see the general process of the single social data extrapolation step.

In the cross-social data extrapolation step (Figure 4.13), the way in which the inputs are exploited, and the interaction procedure of the face recognition module are the same as in the single social data extrapolation step. The main difference is the exploitation of multiple social network modules for extrapolating several user profile data. In particular, each module can extract user profile data from a specific social network. In this way, it is possible to collect several user profile data from different social networks. Obviously, the only limitation is that the target user needs to own a registered profile on each social network. Finally, all the user profile data associated with each social network feed the integration module for aggregating all collected user profile data. In Figure 4.13, it is possible to see the general process of the cross social data extrapolation step.

In the proposed methodology, a single analysis over a specific social network is differentiated from a cross-analysis over multiple social networks. In this way, it is possible to estimate the minimum amount of user data that is possible to extrapolate from a single social network, and evaluate the maximum number of users data that can be aggregated from different social networks.

The following section describes in-depth all sub-modules included in the user data extrapolation tool, by explaining how they interact with each other.

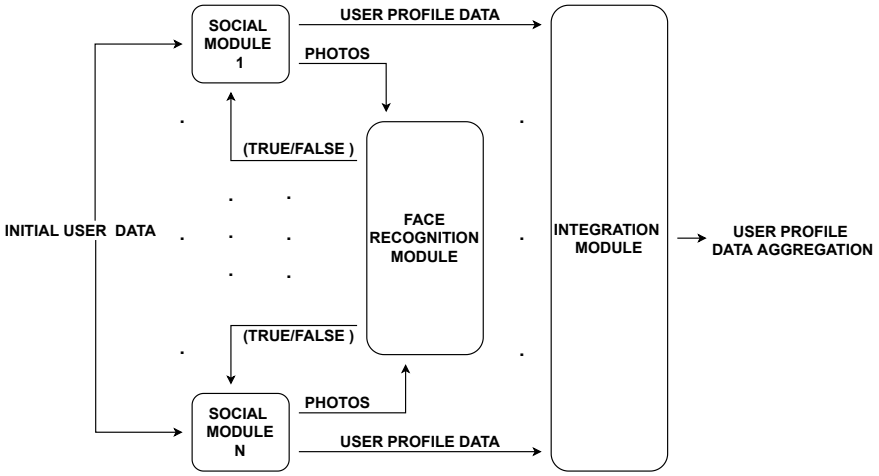


Figura 4.13: Cross-social data extrapolation step.

4.2.2 Social data analyzer

Extracting user data from multiple social networks is a complex task. There are several issues related to the extraction that yield specific choices for the components of the social data analyzer (SODA) tool: i) the number of users involved in the analysis process can be large, ii) each social network relies on different implementation technologies, and iii) continuous upgrades of the social network platforms require continuous maintenance of system components. To this end, the tool SODA is built on the top of the existing system Social Mapper[2], extending several of its components, aiming to tackle the issues mentioned above.

4.2.2.1 A tool for analyzing user data

As said above, SODA has been built on the top of Social Mapper, an open-source tool exploiting face recognition techniques to find social media profiles across different social network platforms. In particular, Social Mapper is capable to search user profiles on the social network platforms, such as Facebook, Linkedin, Instagram,

[2]https://github.com/Greenwolf/social_mapper

VKontakte, Twitter, Pinterest, Weibo, and Douban. It is essential to notice that, since SODA is an extension of Social Mapper, it can search people by only considering an image and at least one of the following data: name, surname, city, email, or the company in which the user works. From these, SODA is capable of browsing the Web by exploiting Selenium[3], a framework that is generally used for activities such as testing, browsing, and scraping Web content.

SODA provides means to automate the navigation on any Web page, by creating a bot to perform operations, and simulating the behaviours of a real user during a Web browsing session. It is important to note that the bot can exploit the search engines behind each social network platform. Since it simulates the operations of a real user, SODA can search for users registered over different social network platforms by quickly filling the search bars and accomplishing the search. In this way, the search is computationally feasible and permits analysing only a subset of users that match the search parameters. The combination of these strategies with a powerful recognition algorithm allows SODA to achieve accurate results. In particular, among the many facial recognition algorithms proposed in the literature [74], Social Mapper relies on the Viola-Jones [75], one of the most frequently used facial recognition algorithms. It uses *Haar feature-based cascade* filters to extract meaningful features of an individual's face [76].

With respect to Social Mapper, the proposed tool SODA provides several novel functionalities that allow to perform an in-depth analysis of the data shared by users, and extend the search on a large scale. The first new functionality enables the system to find people that work in a specific company. To this end, SODA exploits the search mechanism of Linkedin to select the users working in a given company and returns information on their public profile as an output. The idea of starting with Linkedin for selecting users to analyse is due to the fact that they were unlikely to be fake. In fact, it has been demonstrated in [77] that the amount of faker users registered over Linkedin is extremely

[3]https://www.selenium.dev/

small. Moreover, since SODA starts from the list of people working for companies, the probability of finding a fake user is also extremely low. In fact, Linkedin provides each company with a tool to monitor the users registered on them [78]. In particular, this task is generally entrusted to the human resources managers, who periodically check the users affiliated with the company, in order not to damage the seriousness and professional attitude of the company. To this end, exploiting Linkedin for selecting users permits to work with consistent initial data that belong to real users.

Most of the remaining extensions provided by SODA affect the crawling components. In fact, Social Mapper is limited to only extracting the URLs of the different user profiles. Thus, in SODA have been redesigned to add several new navigation features. In particular, due to the various structures of Web pages, it was necessary to design different targeted changes to facilitate the data acquisition phase of each crawling module. More specifically, it has been added support to Web selectors to perform more accurate Web searches. In fact, the selectors are one of the most robust technologies for manipulating Web content, since they support the most frequently used Web browsers. Thanks to these extensions, SODA is able to perform large-scale searches and extract users' data. However, there might be cases of homonymy between users registered in a social network. To this end, SODA combines the information of each user with the results of the face recognition algorithm to only extract a person who best matches the search. More specifically, the face recognition algorithm compares the image taken as input with those of all possible users registered in a specific social network. A user profile is returned as output if and only if the image is at least 60% compatible with the input one and if the data match with it. This threshold ensures that the number of false positives is minimized. In case several users match the search criteria and exceed the threshold of the face recognition module, SODA can extract the data of each user and merge them into a single output. This strategy, combined with the focused search performed by each social network, allows SO-

DA to maximize the amount of extracted information. However, it was necessary to define a threshold value to limit the maximum number of matches and the searches in each social network. This threshold can be set during the configuration step of SODA and is valid for searches in all social networks. Notice that, the choice of this threshold can significantly affect the analysis of SODA. In fact, although a high value for the threshold could maximize information extraction by also capturing users with multiple profiles, this could lead to the extraction of inaccurate user information and significantly lengthen the search time of SODA. To this end, it has been considered a lower threshold value, i.e., 1, in order to speed up the search time and increase the precision of the extracted information.

It is important to notice that SODA is not able to check fake accounts. In fact, it analyzes and extracts information from social network profiles by simulating a real user. Thus, this threshold does not guarantee that the first profile extracted from the social network is a real profile. However, in case the extracted profile was fake, the photo of this profile has already satisfied the similarity threshold of the face recognition module. Thus, this means that this kind of profile is a clone of a real user profile since it would have both the same data and a photo of the real person that SODA was looking for [79, 80].

The following section describes the extended system architecture of SODA, analysing its components, and comparing each with the previous version provided in Social Mapper. Finally, it will be examined the interaction between components of SODA to clearly describe how it extracts information from different social networks.

4.2.2.2 SODA architecture

The architecture of Social Mapper provided no presentation layer, and it relied on a two-tier model, in which it is possible to identify the following two layers:

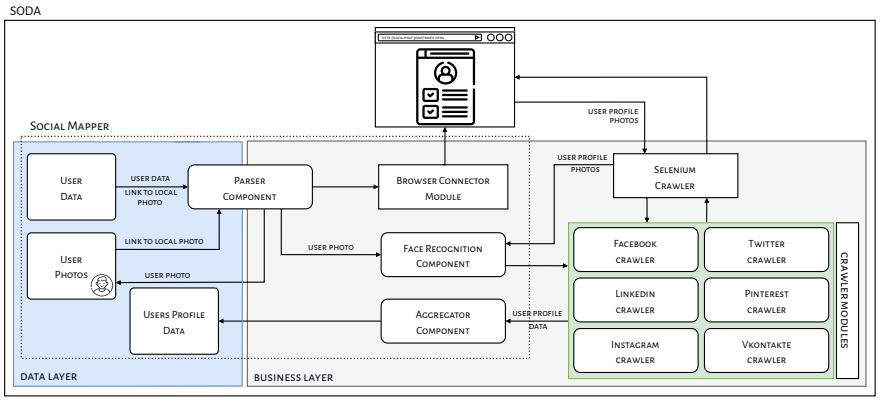


Figura 4.14: Architecture of SODA.

- The *Data layer* containing the initial information necessary for running the system. It consists of all the initial user information, which enables Social Mapper to acquire data for user profiling;
- The *Business layer* containing the modules for extracting information from different social networks.

However, the components of these two layers showed low modularity, making the system difficult to maintain. Thus, part of the work aimed at restructuring each component, in order to derive more a maintainable system, which could be easily upgraded and extended. The first extension of Social Mapper concerns the introduction of a module that enables SODA to manage faults and/or exceptions generated by each component. In fact, to perform a large-scale analysis of people, it is necessary that the system continue operating properly in case of failure of one or more of its components. Moreover, to enable SODA to analyze the information that a user shares, it has been necessary to upgrade each component of Social Mapper in order to add new functionalities for crawling information from different social networks. These new crawling functionalities exploit more general Web selectors, allowing SODA to analyze the contents of the Web pages, regardless of the technologies behind each social network platform.

Figure 4.14 shows the architecture of SODA. In particular, the components within the business layer communicate with those within the data layer through the *Parser* components and the Selenium APIs.

The data are acquired by the *Parser* component, which is responsible for interpreting the system input, trying to understand the execution modes, and for sharing information of each user with the *Face Recognition* module. Moreover, the *Parser* invokes the *Browser Connector* module interface, which enables SODA to execute the local Web browser. After which, it is necessary to interact with the Web pages and extract information. To this end, SODA exploits the functionalities provided by Selenium. More specifically, to extract specific information on each social network, six modules are defined, one for each social network on which it is possible to access user profiles and extract their information. In particular, SODA crawlers search for a user by using the initial information read by the Parser module, and extract all the profile pictures of the users that match the search criteria. The list of pictures is sent to the *Face Recognition* component, which compares the image taken in input with those extracted from the social networks, in order to identify the correct subjects to be analysed. The list of identified subjects is shared with the crawling modules, which acquire all information of each profile, storing them locally. Finally, the *Aggregator* component receives all the data, and groups all the information extracted by the crawlers in a single file.

4.2.3 Experimental Evaluation

This section presents a single-social and a cross-social evaluation, aiming to investigate the sensitivity of the extrapolated data. In what follows, it has been described the collected dataset, the two experimental sessions for evaluating the data of analysed users, and the performances of the proposed tool in terms of extrapolated attributes.

4.2.3.1 Experimental Settings

Dataset. The experimental evaluation required the creation of a dataset of people by randomly extracting them from the Web. In particular, all information is extracted by exploiting the crawler's functionality.

Since social network platforms have different templates for managing the user's information, it has been implemented an ad-hoc crawler to interact with different Web pages and extract only information characterizing the user. To this end, it has been created a dataset containing photos and a few initial information concerning real users, e.g., name, surname, and/or company. The first operation for creating the dataset has been to select people from different parts of the world. In particular, the new features of SODA have been exploited in order to enable searching people working for a specific company. To this end, more than 100 international companies have been randomly selected, from which SODA extracted more than 11000 images of distinct users. The new data have been aggregated into a single structured file and have been used to assess user privacy. It is worth noting that the initial version of the dataset only contained essential information for starting the execution of SODA, whereas all the other data have been added during its execution.

Although the crawler modules try to maximize information extraction from the Web, it might happen that some users do not share enough information, so that the associated tuples in the dataset will contain many null values. Moreover, some user images were not of satisfactory quality or did not show the face. For this reason, in the resulting dataset, only a subset of users has been stored, i.e., those yielding zero or few null values. Thus, the data of 7000 users have been selected, by considering their information as the initial data for the performed evaluation. After the execution of SODA, data from 5000 users have been retrieved, i.e., from users registered on at least one social network platform. For each of them, it was necessary to perform several operations to standardize the extracted data, by removing incorrect values

and cleaning information from outliers, e.g. special characters, and/or emoticons. Finally, all data with proper syntax have been inserted in the initial dataset containing the information of each user already extracted for the search.

Evaluation Metrics. As described in the previous section, the experimental evaluation has involved 7000 people, randomly selected from the Web. Starting from them, the analysis on each social network has been performed, also including Linkedin, aiming to evaluate the effectiveness of SODA.

Among the people involved in the evaluation, 5000 have been found on at least one social network, and have been classified as true positive (TP), 878 people have been classified as false positives (FP), that is, people who have been erroneously found on a social network and with a matching rate greater than 60%; 1122 people have not been found, and therefore were classified as false negative (FN), and finally, the people who were not registered on any social network were classified as (TN), and in the proposed evaluation have been considered as 0, since the initial data was extracted from Linkedin.

It is important to notice that the people who have not been correctly identified, i.e., TP, probably changed their profile photos during the evaluation period. In fact, the experimental evaluation lasted several months. Therefore, it is likely that in the meantime users could change their profile photos, making identification more complex. In addition, other reasons could be due to the posture assumed by the subject in the photos and the lighting conditions. In fact, several studies have shown that these two factors can negatively affect face recognition algorithms by reducing the matching rate [81, 82].

Although these problems could affect evaluation results, the effectiveness of SODA is shown by the following metrics:

$$Precision = \frac{TP}{TP + FP} = \frac{5000}{5000 + 878} = 0.85$$

$$Recall = \frac{TP}{TP + FN} = \frac{5000}{5000 + 1122} = 0.82$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{5000 + 0}{5000 + 0 + 878 + 1122} = 0.72$$

4.2.3.2 Results

Evaluation single-social. This section describes statistics obtained by evaluating data extracted by each considered social network. In particular, the information that are frequently shared by users over every single social network and the way each social network preserves user privacy are highlighted. To this end, starting from the 5000 users contained in the dataset, a single social network evaluation has been performed. This allowed to independently analyze the results obtained by each social network, avoiding to consider whether a user is present on multiple platforms, which will be discussed in the next section.

Figure 4.15 shows the most frequently shared information on Linkedin extracted by 1570 users registered to it. Among the 5000 initial users, only the accounts from which it has been possible to extract sensitive information useful for the proposed analysis have been considered. It is possible to notice that *Employment* and the *City* are the most frequently shared information on Linkedin. In particular, the attribute city can refer to the place of residence or the place of birth, but in most cases, these are equal.

Results in Figure 4.15 highlight even more that Linkedin is a social network for job finding, where users tend to share their employment and city, aiming to find better job opportunities.

Figure 4.16 shows the most frequently shared information on Facebook, extracted by 1161 users registered to it. It is possible to notice that basic information related to the *gender*, *Education* or *Work*, and the *Place where the user lives* are the most frequently shared information on this social network. In particular, as it can be seen in Figure 4.16, no user has shared his/her details on the date of birth, which combined with the other data, could significantly affect privacy. Facebook permits users to hide their data in order to preserve privacy.

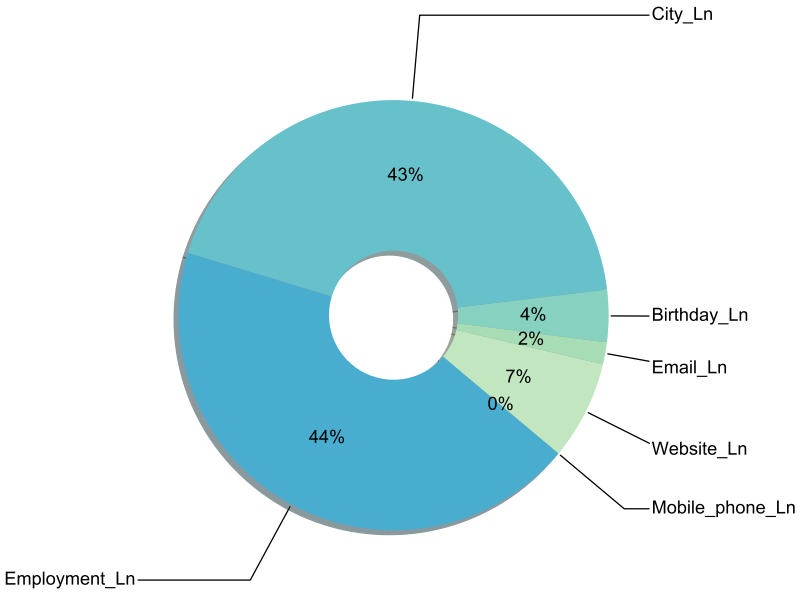


Figura 4.15: Analysis of information shared on Linkedin.

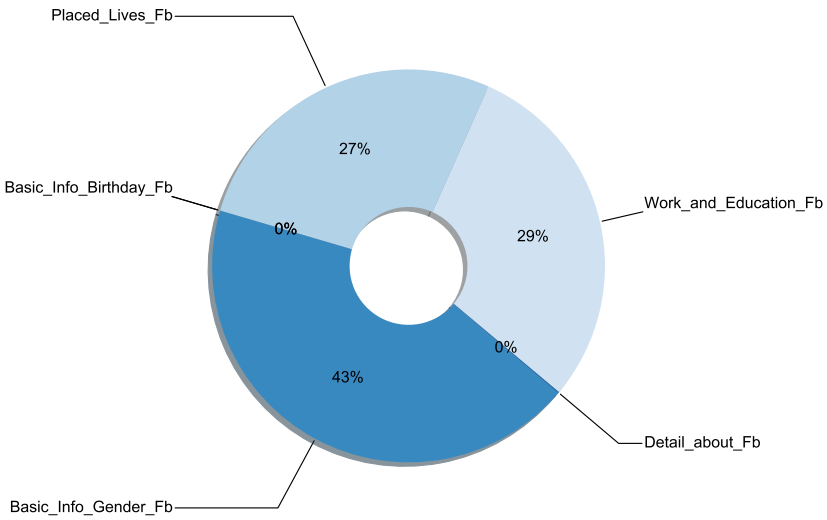


Figura 4.16: Analysis of information shared on Facebook.

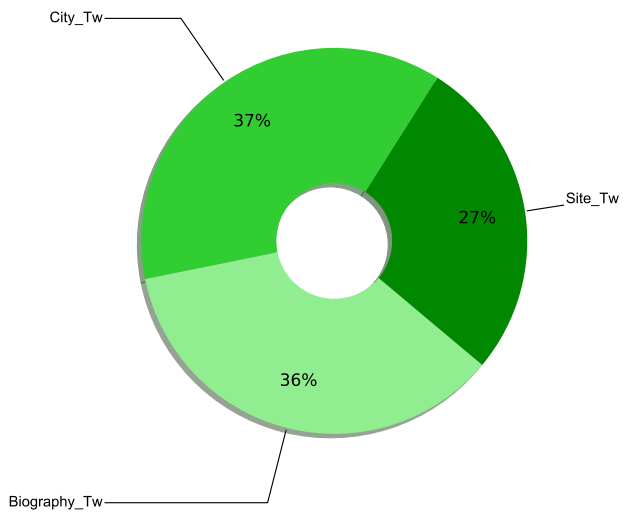


Figura 4.17: Analysis of information shared on Twitter.

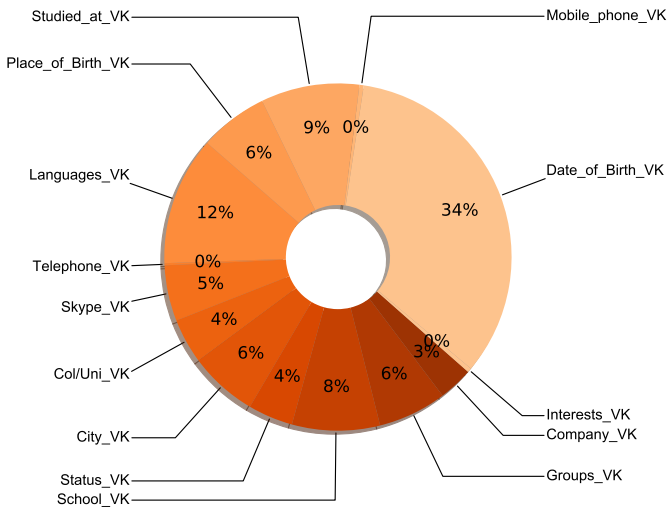


Figura 4.18: Analysis of information shared on VKontakte.

Figure 4.17 shows the most frequently shared information on Twitter, extracted by 86 users registered to it. Despite not many users involved in the analysis, it is possible to notice that the

City, *Website*, and the *Biography* of a user are the most frequently shared information on this social network. In particular, through the biography a user can share additional information, such as his/her telephone number, email, or other information.

Twitter is used by many famous people, but it offers less prevention in terms of privacy, mainly due to the fact that users tend to insert data in their biography, not being aware to disclose them.

Figure 4.18 shows the most frequently shared information on VKontakte, extracted by 251 users registered to it. It is possible to notice that, the *Date of birth*, the *Spoken languages*, and the *Education* information are the most frequently shared data on this social network. In particular, as shown in Figure 4.18, no many users have shared their *Telephone* numbers. As Facebook, also VKontakte is a social network that allows users to share a vast amount of information, and it permits users to hide specific details to preserve privacy.

Concerning Pinterest and Instagram, 1688 and 2845 user profiles have been evaluated, respectively. In particular, these two social networks are massively used for sharing photos, and no other types of data have been found for the performed analysis. Furthermore, the only textual information on Instagram that seemed useful for the proposed analysis was the user biography. Yet, a user can write anything in it, so it has been decided not to take the biography into account for the proposed analysis.

In table 4.1, for each analysed social network, it has been summarised the information it retrieved. Yet, "Required attributes" (i.e. attributes mandatory in the social network's registration phase), "Public attributes" (i.e. attributes public by default), "Attributes extracted" (i.e. attributes gathered by the performed analysis for a specific social network), and "Number of extracted attributes" (i.e. the amount of extracted attributes for each specific social network) have been compared.

As it is possible to notice in Table 4.1, except for Twitter and Instagram, all other social networks permit to retrieve different information, so that starting from the Public attributes it is possible to reconstruct a partial user's profile.

	Required data	Public data	Data extracted
Linkedin	Name & Surname E-mail	Name & Surname City Employment Birthday	Full_Name Mobile_phone_Ln Website_Ln Email_Ln Birthday_Ln City_Ln Employment_Ln
Facebook	Name & Surname Phone Number Birthday Gender	Name & Surname	Full_Name Work_and_Education_Fb Placed_Lives_Fb Contact_Fb Basic_Info_Birthday_Fb Basic_Info_Gender_Fb Detail_about_Fb
Twitter	Name & Surname E-mail Phone Number Birthday	Name & Surname City Biography Website	Full_Name Site_Tw City_Tw Biography_Tw
Instagram	Name & Surname E-mail Phone Number	Name & Surname Biography Website	Full_Name Biography_In
VKontakte	Phone Number Birthday Gender Name & Surname	Name & Surname Place_of_Birth Website Company Languages Mobile_phone Telephone College_or_university Status School Interests	Full_Name Date_of_Birth_VK City_VK Studied_at_VK Place_of_Birth_VK Languages_VK Mobile_phone_VK Telephone_VK Skype_VK College_or_university_VK Status_VK School_VK Groups_VK Company_VK Interests_VK

Tabella 4.1: Single social features extrapolation.

Evaluation cross-social This section describes statistics derived by performing a cross-social analysis on data extrapolated by all the available social networks. In particular, it has been investigated the possibility of aggregating information made publicly accessible by users over different social networks, aiming to perform a more detailed analysis.

Figure 4.19 shows the distribution diagram for the users regi-

stered over the considered social network platforms. In particular, except for the first bar highlighting the number of users not involved in social networks, it is possible to group the other bars in three blocks, representing the users found in one, two, and three social network platforms, respectively. The blue dots under each bar indicate the social networks on which the users have been found after the experimental session. As it is possible to see from Figure 4.19, there are no users discovered in more than three social network platforms, and Instagram represents the most frequently used platform from the users involved in the performed evaluation. However, in Figure 4.20, it is possible to notice that users share a large amount of information on Linkedin. This is mainly due to the registration policies of this social network, which requires to insert various personal data. Since users exploit Linkedin mainly for business purposes, this means that they share a vast amount of data without privatising them.

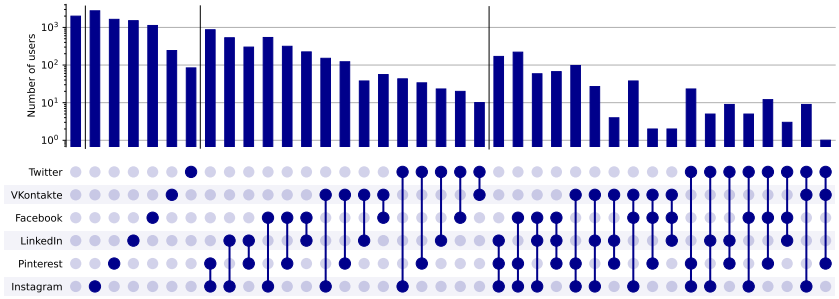


Figura 4.19: Distribution diagram of the analyzed users.

In Figure 4.20, the statistics concerning email sharing over different social networks are shown. By analysing different social networks, it is possible to notice that only Linkedin, Facebook, and VKontakte have a special section for inserting this information. Concerning the email histogram in Figure 4.20, the x-axis represents the attribute *Email* over Linkedin, Facebook, and VKontakte, while the y-axis represents the absolute frequencies of emails shared on each social network. In detail, Linkedin users present

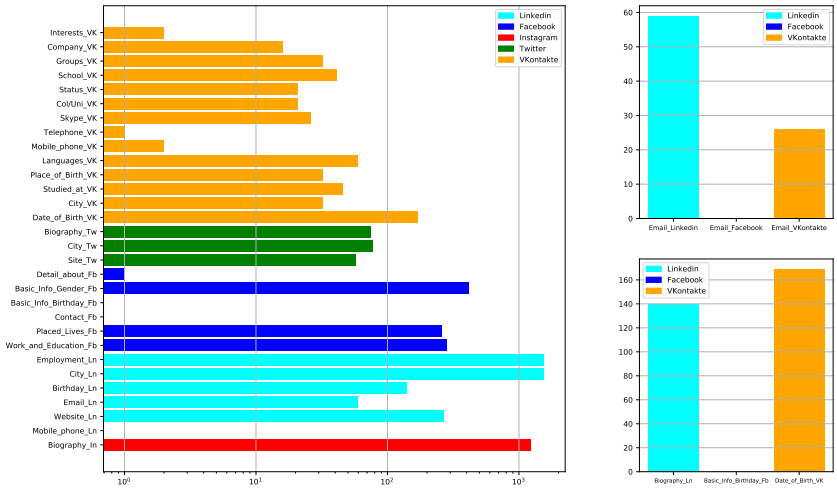


Figura 4.20: Attribute statistics of the entire dataset.

a high frequency for sharing the attribute *Email*, whereas few are the users that shared it on VKontakte, and no one on Facebook.

In Figure 4.20 statistics concerning the *Date of birth* sharing over different social networks are shown. By analysing different social networks, it is possible to notice that only Linkedin, Facebook, and VKontakte have a special section for inserting this information. Concerning the *Date of birth* histogram in Figure 4.20, the x-axis represents the attribute *Date of birth* over Linkedin, Facebook, and VKontakte, while the y-axis represents the absolute frequencies by which this attribute is shared on each social network. In details, users of VKontakte and Linkedin present a high frequency for the attribute *Date of birth*, whereas no one shared it on Facebook. Furthermore, it is possible to notice that before registering on VKontakte, users have to mandatorily insert the date of birth, which is never hidden for the analysed users, even if VKontakte permits handling privacy settings.

Concerning the sharing of *Telephone number*, only the information available in VKontakte was useful for the performed analysis, but it could be possible to retrieve a reduced amount of telephone numbers. The insertion of the telephone number is essential

for registering in VKontakte, but the majority of analysed users maintain this data hidden. Other social networks always hide the telephone number.

In Figure 4.21, statistics concerning the sharing of the *City* over different social networks are shown. It is possible to notice that only Linkedin, Facebook, Twitter, and VKontakte have a special section for inserting this information. Concerning Figure 4.21, the x-axis represents *attributes City*, *Place of living*, and *Place of birth* over Linkedin, Facebook, Twitter, and VKontakte, whereas the y-axis represents the absolute frequencies by which the attribute *city* is shared on each social network. In details, users on Linkedin and Facebook present a high frequency for the attribute *City*, whereas few are the users who have shared it on Twitter and VKontakte. In all analysed social networks, it has been possible to retrieve information related to the city of users.

In Figure 4.22, statistics concerning information about *Training* and *Employment* sharing over different social networks are shown. It is possible to notice that only Linkedin, and Facebook have a special section for inserting this information. The x-axis represents attributes *Employment*, *Work/Education*, *Studied at*, *College/university*, *School*, and *Company*, over Linkedin, Facebook, Twitter, and VKontakte, whereas the y-axis represents the frequencies by which this information shared on each social network. In details, users on Linkedin and Facebook present a high frequency for attributes *Employment* and *Work/Education*, whereas few of them share *College/University*, *School* on VKontakte.

A cross-social analysis permits the reconstruction of information over different social networks. For example, a user registered on several social networks can decide to privatise some information on a specific social network, where s/he can choose to unmask the same information over other social networks. It means that by analysing a specific user over different social networks, it is possible to obtain more detailed information.

In the performed analyses, privatised data, i.e., data that is not publicly available on user profiles, and the data of the users

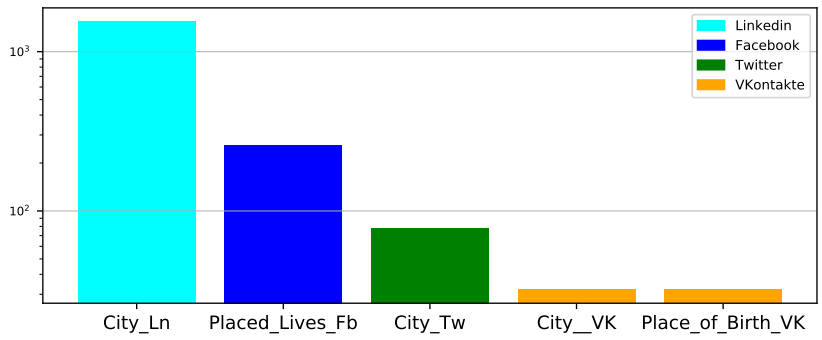


Figura 4.21: City attribute shared over analyzed social networks.

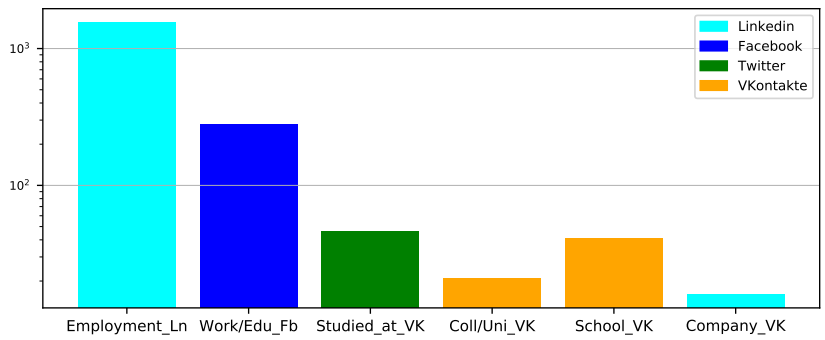


Figura 4.22: Job attribute shared over analyzed social networks.

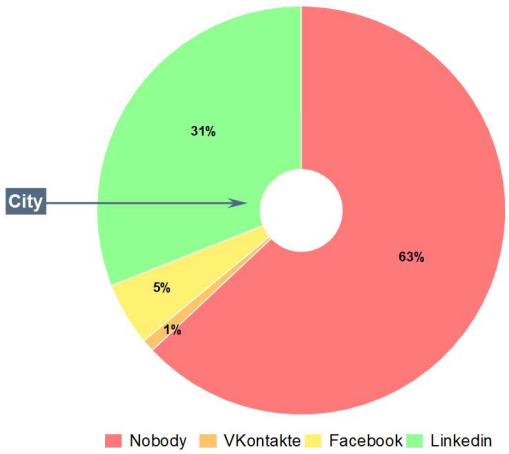


Figura 4.23: Attributes reconstruction by exploring all analyzed social except Twitter.

that is not found on any social networks, are managed in the same way considering them as missing values.

The most frequently accessible information on Twitter is the city, since it can be reconstructed through other social networks. Figure 4.23 shows that 4923 out of 5000 analyzed users are not registered on Twitter or have privatized this information on it. However, 31% out of 4923 users published their city on Linkedin, while 5% on Facebook, and 1% on VKontakte. The remaining 63% out of 4923 users did not share this information over any considered social network, leading to the impossibility of extracting the information concerning their city. Consequently, only in the last case, it is possible to guarantee the confidentiality of the data (e.g., city), by simply requiring the management of its privatization over just one social network (e.g., Twitter).

The information that is most frequently accessible on Facebook is *Mobile phone*, *City*, *Date of birth*, *Email*, and information concerning *Education*, and *Training* or *Work*. For the proposed analysis on Facebook, the last two attributes have been merged. In Figure 4.24, the percentage of information privatised by Facebook users, but published on other social networks, has been shown:

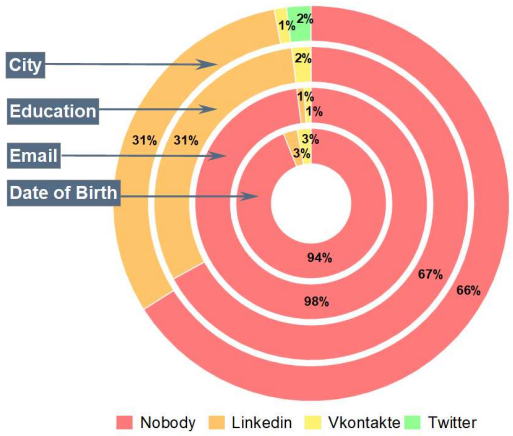


Figura 4.24: Attributes reconstruction by exploring all analyzed social except Facebook.

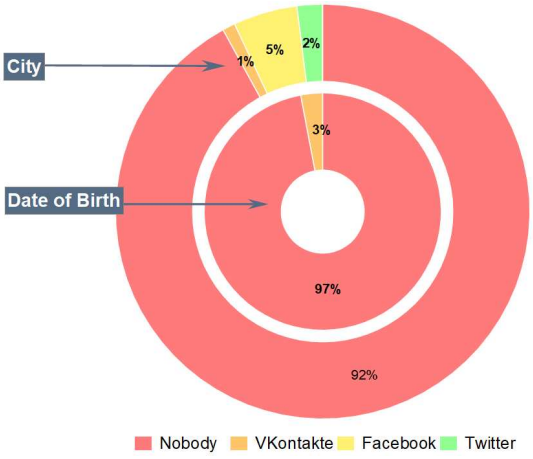


Figura 4.25: Attributes reconstruction by exploring all analyzed social except Linkedin.

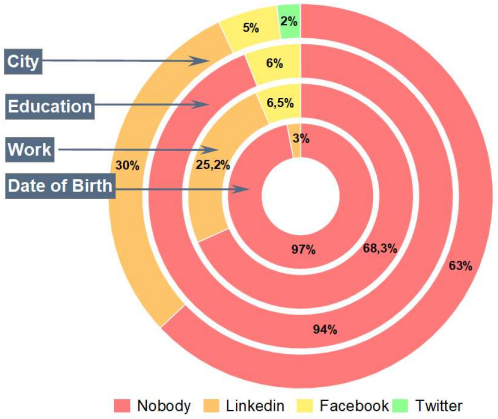


Figura 4.26: Attributes reconstruction by exploring all analyzed social except VKontakte.

- In the figure, no diagram is shown for *Mobile number*, since among the 5000 analyzed users who have privatized their mobile number on Facebook, no one has allowed the reconstruction of this information from other social networks;

- Among the 5000 users analyzed, 4743 have privatised their *Hometown* or *Residence* on Facebook, or are not registered to this social network. Among them, 31% have published this information on Linkedin, 2% on Twitter, and 1% on VKontakte. Thus, 34% of them allow the reconstruction of this information from other social networks;

- Among 5000 analyzed users who have privatized their *Date of birth* on Facebook or are not registered on this social network, 3% shared it on VKontakte, and 3% on Linkedin. In summary, 94% of analyzed users have privatized this information, since 6% of them shared it on other social networks;

- Among the 5000 analyzed users who have privatized the

Email on Facebook or are not registered on this social network, only 1% of them shared it on Linkedin, while 1% on VKontakte. In summary, 2% of analyzed users shared the *Email* on other social networks, so 98% have completely privatized it;

- Among the 5000 users analyzed, 4721 users have privatized *Education* on Facebook, or are not registered on this social network. Among them, 31% published this information on Linkedin, and 2% on VKontakte. In summary, 33% of analyzed users have shared the *Education* on other social networks, so 67% have completely privatized it.

Results show that most of the analysed users who have privatised a given data on Facebook have also privatised it on other social networks. Among all considered social networks, Linkedin has proved to be useful for the reconstruction of user's information.

The information that are most frequently accessible on Linkedin are *Mobile phone*, *City*, *Date of birth*, *Email*, and *Employment*. In Figure 4.25, it has been shown the percentage of information privatised on Linkedin, but published on other social networks:

- Similarly to Facebook, no diagram is shown for *Mobile phone* number, since among the 5000 analyzed users who have privatized their mobile phone number on Facebook, or who are not registered on this social network, no one published it on other social networks;

- Among the 5000 users analyzed, 3450 have privatised their *Hometown* or *Residence* on Linkedin, or are not registered on this social network. Among them, 5% have published it on Facebook, 2% on Twitter, and 1% on VKontakte. In summary, 8% of analysed users shared *Hometown* or *Residence* on other platforms, so 92% have completely privatised it;

- Among the 5000 users analyzed, 4861 have privatized their *Date of birth* on Linkedin or are not registered on this social

network. Among them, only 3% shared it on VKontakte. In summary, 3% of analyzed users shared the *Date of birth* on other social networks, while 97% have completely privatized it;

- Among the 5000 users analyzed, 4942 have privatized their *Email* on Linkedin or are not registered on this social network. Among them, only 1% shared it on VKontakte. In summary, 1% of analyzed users shared the *Email* on other social networks, while 99% have completely privatized it;

- Among the 5000 users analyzed, 3445 have privatized their *Training/Work* on Linkedin or are not registered on this social network. Among them, 6% shared it on Facebook, and 1% on VKontakte In summary, 7% of analyzed users shared the *Training/Work* on other social networks, so 93% have completely privatized it.

Results show that most of the analysed users who have privatised a given data on Linkedin have also privatised it on other social networks. Among all considered social networks, Facebook has proven to be useful for the reconstruction of user's information.

The information that are most frequently shared on VKontakte are *Mobile phone*, *City*, *Date of birth*, *Email*, and information concerning *Training* and *Work*. In Figure 4.26, the percentage of information privatised on VKontakte, but published on other social networks, has been shown:

- Similarly to the previous analysis, no diagram is shown for *Mobile phone* number on VKontakte, since among the 5000 analyzed users who have privatized their mobile phone number on VKontakte, or who are not registered on this social network, no one published it on other social networks;

- Among the 5000 users analyzed, 4990 have privatized their *Hometown* or *Residence* on VKontakte or are not registered on this social network. Among them, 30% of them have published it on Linkedin, 2% on Twitter, and 5% on Facebook.

In summary, 37% of analysed users shared the *Hometown* or *Residence* on other social networks, so 63% have completely privatised it;

- Among the 5000 users analyzed, 4832 have privatized their *Date of birth* on VKontakte or are not registered on this social network. Among them, only 3% of them have published it on Linkedin. In summary, 3% of analysed users shared it on other social networks, so 97% have completely privatised it;

- Among the 5000 users analyzed, 4975 have privatized their *Email* on VKontakte or are not registered on this social network. Among them, only 1% of them shared it on Linkedin. In summary, 1% of analysed users shared it on other social networks, so 99% have completely privatised it;

- Among the 5000 users analyzed, 4997 have privatized their *Education* on VKontakte or are not registered on this social network. Among them, only 6% of them have published it on Facebook. In summary, 6% of analysed users shared it on other social networks, so 94% have completely privatised it;

- Among the 5000 users analyzed, 4998 have privatized their *Work* on VKontakte or are not registered on this social network. Among them, 25.2% of them have published it on Linkedin, and 6.5% on Facebook. In summary, 31.7% of analysed users shared it on other social networks, so 68.3% have completely privatised it.

Results show that most of the analysed users who have privatised a given data on VKontakte have also privatised it on other social networks, except for *Employment*, *City of residence* or *Date of birth*. Among all considered social networks, Linkedin has proven to be useful for the reconstruction of user's information.

Table 4.2 summarises the additional information gathered by performing a cross-social analysis over each analysed social network. In particular, for each social network (rows in Table 4.2), it has been highlighted the additional information retrieved from other ones (columns in Table 4.2). Obviously, the diagonal reports similar information presented in Table 4.1. As it is possible to notice in Table 4.2, Facebook, Twitter, and Vkontakte permit to retrieve beneficial information concerning users for creating more a detailed user's profile.

Finally, Table 4.3 shows a final overview of the user profile information collected through cross-social analysis. It highlights some of the sensitive information of users by merging the extrapolated and reconstructed data, with the aim to create a complete user profile for each subject.

As prescribed in the GDPR: data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, genetic characteristics, biometric information processed solely to identify a human being, health-related information, and concerning a person's sex life or sexual orientation, is considered sensitive[4]. Data reported in Table 4.3, are singularly not sensitive for GDPR, but their aggregation permit to identify a specific user putting at risk his/her privacy.

4.2.4 Ethical discussion

Social networks represent a vast information source in terms of data. However, processing and analysing data gathered by social networks could raise ethical discussions. This section aims to explain the ethical issues related to the presented work.

Concerning the application of the GDPR for research purposes, it states that for meeting *"the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclo-*

[4]https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en

	Linkedin	**Facebook**	**Twitter**	**Instagram**	**VKontakte**
Linkedin	Mobile_phone_Ln Website_Ln Email_Ln Birthday_Ln City_Ln Employment_Ln	Placed_Lives_Fb Basic_Info_Gender_Fb Detail_about_Fb	Biography_Tw	Biography_In	Place_of_Birth_VK Languages_VK Skype_VK College_or_university_VK Status_VK Groups_VK Company_VK Interests_VK
Facebook	Mobile_phone_Ln Website_Ln Email_Ln Employment_Ln	Work_and_Education_Fb Placed_Lives_Fb Contact_Fb Basic_Info_Birthday_Fb Basic_Info_Gender_Fb Detail_about_Fb	Biography_Tw Site_Tw	Biography_In	Place_of_Birth_VK Languages_VK Mobile_phone_VK Telephone_VK Skype_VK Status_VK Groups_VK Company_VK Interests_VK
Twitter	Mobile_phone_Ln Email_Ln Birthday_Ln Employment_Ln	Work_and_Education_Fb Contact_Fb Basic_Info_Birthday_Fb Basic_Info_Gender_Fb Detail_about_Fb	Site_Tw City_Tw Biography_Tw	Biography_In	Date_of_Birth_VK Studied_at_VK Place_of_Birth_VK Languages_VK Mobile_phone_VK Telephone_VK Skype_VK College_or_university_VK Status_VK School_VK Groups_VK Company_VK Interests_VK
Instagram	Mobile_phone_Ln Website_Ln Email_Ln Birthday_Ln City_Ln Employment_Ln	Work_and_Education_Fb Placed_Lives_Fb Contact_Fb Basic_Info_Birthday_Fb Basic_Info_Gender_Fb Detail_about_Fb	Site_Tw City_Tw Biography_Tw	Biography_In	Date_of_Birth_VK City_VK Studied_at_VK Place_of_Birth_VK Languages_VK Mobile_phone_VK Telephone_VK Skype_VK College_or_university_VK Status_VK School_VK Groups_VK Company_VK Interests_VK
VKontakte	Website_Ln Employment_Ln	Basic_Info_Gender_Fb Detail_about_Fb	Site_Tw Biography_Tw	Biography_In	Date_of_Birth_VK City_VK Studied_at_VK Place_of_Birth_VK Languages_VK Mobile_phone_VK Telephone_VK Skype_VK College_or_university_VK Status_VK School_VK Groups_VK Company_VK Interests_VK

Tabella 4.2: Cross social features extrapolation.

	Description
Full name	Name and Surname of the user.
Mobile_phone	Mobile number of the person.
Telephone	Landline number.
Website	Personal or company website.
Email	Personal email.
Birthday	Date of birth.
City_of_Birth	Place of birth, can be the same as current place of residence.
Employment	Job position.
Placed_Lives	Current place of residence, can be the same as place of birth.
Gender	Gender of the individual.
Skype	Skype nickname.
College	Name of the college or university attended.
Status	Professional status or highest level of education.
School	Attended schools.
Groups	Names of groups to which the user is subscribed.
Interests	Interests of the user.
Company	Company name the employee belongs
Biography	Biography written by the user.
Languages	Languages of the user.

Tabella 4.3: User's profile information obtained after cross-social analysis.

sure of personal data in the context of scientific research purposes" (recital 159). The GDPR defines some other bases, for the processing of the personal data to be lawful. When the processing is necessary to protect the vital interests of the data subject or another natural person (Article 6(1)(d)); or when the processing is necessary for the performance of a task carried out in the public interest (Article 6(1)(e)). Moreover, recital 157 identifies the benefits of personal data research, subject to appropriate conditions and safeguards. These benefits include the potential for new knowledge when researchers *" obtain essential knowledge about the long-term correlation of a number of social conditions"*. The results of the research *"obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people, and improve the efficiency of social services"* (recital 157).

According to the claims described above, social network users data have been collected to perform a specific analysis, with the only aim to improve user's awareness concerning privacy threats over different social networks. The analysis has shown that users

are not really aware of privacy threats linked to the dissemination of their data over different social platforms.

To comply with GDPR, only the statistics retrieved from the collected social network data will be made public, without publishing the data itself. From the ethical point of view, user's privacy has not been violated, because this was not the target; data has been collected with the only purpose to emphasise privacy issues related to social network data dissemination.

It is possible to justify the ethical aspects of the proposed work by referring to articles 6(1)(d) and 6(1)(e) defined in the GDPR. This research could be the baseline to improve the user's awareness in terms of data privacy and also help to determine new strategies to privatise social network user's data.

4.3 Malicious Account Identification in Social Networks

Currently, platforms for social interactions on the Web are utilised by people of all ages. In fact, many tasks are being transferred over social networks, like advertisements, political communications, and so on, yielding huge volumes of data disseminated over the network. However, this produces several concerns regarding the truthfulness of such data and of the accounts generating them.

Data are often manipulated by malicious users in order to obtain profit. As an example, malicious users create fake accounts and fake followers to increase their popularity and attract more sponsors, followers, and so on., potentially producing several negative implications that impact society.

The number of strategies for detecting and deleting fake accounts has grown proportionally to the number of new algorithms developed for harmful purposes. As described in Section 2.1, such strategies typically exploit machine learning techniques to classify data patterns that characterize fake accounts. To this end, a new approach has been defined aiming to enhance machine learning

techniques in discriminating fake accounts. It exploits algorithms to extract RFDs from the data stored in the social networks and a new heuristic to derive their application in order for discriminating fake accounts. In addition, according to the defined heuristic, a feature selection strategy has been proposed in order to improve classification results. Experimental results highlight the effectiveness of the proposed technique in distinguishing fake accounts over the Twitter platform, and in improving existing machine learning based techniques. The Twitter platform has been chosen because it is the only social network platform providing publicly available data, due to strict privacy laws.

4.3.1 RFD-based Fake account discrimination

Fake account identification is a hot topic, since the massive usage of social networks has contributed to quickly spread harmful information. However, the manual detection of such accounts requires a big effort by humans to analyse vast volumes of accounts data. To this end, the technique proposed in this paper supports the automatic identification of fake accounts in big social networks, by exploiting the concept of RFD (explained in Section 3.2.1) to emphasize data correlations that are typical of fake accounts.

The RFD formalism is suitable for detecting fake accounts, since it captures similarities among data that are typical of automatic procedures used to create fake accounts. Fake accounts generators usually introduce small differences during the generation of account properties, such as screen name, user data, and account creation time-stamp. Key dependencies have been excluded, since they do not involve pairwise tuple comparisons; hence they do not provide meaningful patterns to discriminate between Fake and Real accounts.

Given a relation Set_{REAL} containing real account data, and a relation Set_{FAKE} containing fake account data. Thanks to the availability of RFDs extraction algorithms [70], it is possible to automatically extract RFDs holding on each dataset. Therefore, by analysing differences between the two sets of RFDs, it is

possible to extract meaningful patterns concerning fake accounts. To this end, the goal is to detect the RFDs highlighted in Figure 4.27, that is, those holding on fake but not on real accounts, as described by the following formula:

$$\Delta_{\mathrm{Set_{FAKE}},\mathrm{Set_{REAL}}} = \mathrm{P}_{\mathrm{Set_{FAKE}}} \setminus \mathrm{P}_{\mathrm{Set_{REAL}}}$$

Where $\mathrm{P}_{\mathrm{Set_{FAKE}}}$ and $\mathrm{P}_{\mathrm{Set_{REAL}}}$ represent the RFDs holding on Fake and Real datasets, respectively.

RFDs that are common to the two sets can be ignored, since they do not permit to discriminate between the two types of accounts.

Analogously, given the relation $\mathrm{Set_{VERIFIED}}$, which contains data of accounts named real accounts, since their genuineness has been certified by Twitter, for which it would be useful to analyse the human behaviour. This can be done by considering RFDs holding on real accounts but not on fake ones. More specifically, the attention has been focused on a subset of RFDs identified in $\mathrm{Set_{REAL}}$ (the grey part of Figure 4.28), which corresponds to the

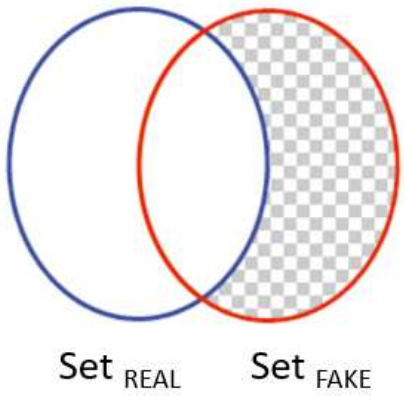


Figura 4.27: Intersection of RFD sets holding on Real and Fake accounts, respectively.

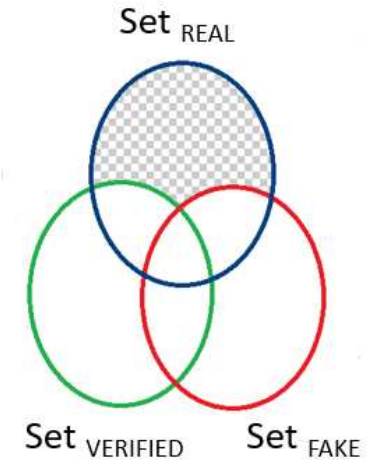


Figura 4.28: Intersection of all RFD sets holding on Real, Verified, and Fake account, respectively.

set of RFDs defined by the following formula:

$$\Delta_{\mathrm{Set_{REAL}},\mathrm{Set_{FAKE}},\mathrm{Set_{VERIFIED}}} = \mathrm{P}_{\mathrm{Set_{REAL}}} \setminus (\mathrm{P}_{\mathrm{Set_{Verified}}} \cup \mathrm{P}_{\mathrm{Set_{FAKE}}})$$

Where $\mathrm{P}_{\mathrm{Set_{FAKE}}}$, $\mathrm{P}_{\mathrm{Set_{REAL}}}$, and $\mathrm{P}_{\mathrm{Set_{Verified}}}$ represent the RFDs holding on Fake, Real, and Verified datasets, respectively. $\Delta_{\mathrm{Set_{REAL}},\mathrm{Set_{FAKE}},\mathrm{Set_{VERIFIED}}}$ contains the RFDs that are only in $\mathrm{Set_{REAL}}$, without considering those that are in $\mathrm{Set_{FAKE}}$ and $\mathrm{Set_{VERIFIED}}$. In other words, the purpose is to find human behavioural patterns that are not replicable by algorithms or bots. Thus, it is possible to ignore RFDs holding on either $\mathrm{Set_{FAKE}}$ or $\mathrm{Set_{VERIFIED}}$. In particular, have been overlooked the RFDs holding on $\mathrm{Set_{VERIFIED}}$, since this set includes Cyborgs, which mix human and bot behaviours, hence they are not useful to characterize typical human behaviour.

At this point, the selected RFDs can be used to discriminate fake accounts from real and verified ones. However, there could be scenarios in which a vast number of RFDs are generated, which would require a significant effort from an expert to exploit them for

discriminating fake accounts. For this reason, in the following, an RFD ranking technique has been introduced, which relies on the concept of *support* to sort RFDs and highlights the most meaningful ones. To this end, the ranking technique exploits the concept of *support*, which represents a well known metric in the context of association rule mining [83].

In this context, the support of an RFD $X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} A_\phi$ indicates how frequently tuple pairs are similar according to the comparison thresholds defined for the LHS X, over the total number of tuple pairs of a dataset. More formally, let r be a relation instance, S be the set of all possible tuple pairs that can be generated from the tuples of r, and $X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} A_\phi$ be a discovered RFD, then the support $supp(X)$ can be defined as:

$$supp(X) = \frac{|\{(t_1, t_2) \in S \text{ s.t. } (t_1, t_2) \text{ satisfies } \Phi_1\}|}{|S|}$$

where t_1 and t_2 are tuples of r, and Φ_1 is the sets of similarity constraints on attributes $B \in X$.

In order to use the concept of support within the proposed ranking technique, a weighted version of the definition above has been considered. In particular, $supp(X)$ has been weighed by considering the domain cardinality of each attribute in X. In this way, it is possible to emphasize RFDs involving LHS attributes with higher domain cardinalities.

4.3.2 Experimental Evaluation

In the evaluation of the proposed technique some difficulties have been encountered, due to the fact that some social networks do not make their data publicly available. In fact, the only social network sharing some of its data is Twitter Inc, hence experiments have been performed on this social network.

4.3.2.1 Experimental Settings

Datasets. The datasets on which experiments have been performed are shown in Table 4.4, together with some profiling data

about them. They contain 9019 accounts, each annotated with one of the following labels: *Real*, *Verified* and *Fake*. Twitter platform has been used for the proposed experiment. Concerning *Real* and *Fake* accounts, the available datasets provided by [84] have been used. Concerning *Verified* accounts, only user profiles validated by Twitter have been considered. In other words, all Twitter profiles with "blue tick" on their Twitter page.

Datasets	**# Columns**	**# Rows**	**Size [KB]**
Verified accounts	15	3949	400
Real accounts	15	1757	168
Fake Accounts	15	3313	316

Tabella 4.4: Profiles of the datasets considered in the evaluation.

Although the Twitter APIs enable to retrieve more than 15 attributes from Twitter accounts, the proposed analysis has been focused only on the 15 most relevant attributes of each considered dataset, since the remaining ones have some correlation with them. The selected attributes are shown in Table 4.5.

Process. The first step has been to automatically extract RFDs from the datasets shown in Table 4.4 by means of the algorithm described [69]. Although the latter is capable of extracting hybrid RFDs, for the purposes of this study is possible to consider only those using relaxation on the comparison method. Such RFDs have been used to discriminate Fake (FK) accounts from Real (RL) and Verified accounts (VRF). Furthermore, according to the technique described in Section 4.3.1, the following parameters have been computed: (i) the cardinality domains of attributes involved in the datasets; (ii) the weights linked to such attributes, using them for computing the support of the discovered RFDs.

4.3.2.2 Results

In order to discriminate fake accounts, the goal of the proposed methodology is to derive the following sets of RFDs:

$$\Delta_{\mathrm{Set_{FAKE}},\mathrm{Set_{REAL}}} = \mathrm{P}_{\mathrm{Set_{FAKE}}} \setminus \mathrm{P}_{\mathrm{Set_{REAL}}}$$

Attribute	Description
name	Name chosen by the account owner. It consists of a maximum of 20 characters.
screen_name	Identifier associated with the account owner. It consists of a maximum of 15 characters.
location	Location defined by the account owner.
url	Boolean value representing the absence or not of the URL set by the account owner.
description	Boolean value representing the absence or not of the description set by the account owner.
followers_count	Number of current followers associated with the account owner.
friends_count	The number of users followed by the account owner profile.
listed_count	Number of public list in which the account owner appears.
created_at	Data and time representing the creation of the account on Twitter.
favourites_count	Number of Tweet generating by the account owner during his/her activities on Twitter.
geo_enabled	Boolean value representing the geotagging of the Tweets related to the account owner.
statuses_count	Number of tweets/retweets that the account owner made.
lang	Code associated with the language specified by the account owner.
default_profile	Boolean value representing changes related to the theme or background of the account owner.
default_profile_image	Boolean value revealing whether the default image of Twitter has been changed by the account owner.

Tabella 4.5: Attributes concerning Twitter user objects.

$$\Delta_{\mathrm{Set_{FAKE}},\mathrm{Set_{VERIFIED}}} = \mathrm{P}_{\mathrm{Set_{FAKE}}} \setminus \mathrm{P}_{\mathrm{Set_{VERIFIED}}}$$

Table 4.6 describes the numbers $|\mathrm{P}_{\mathrm{Set_{VERIFIED}}}|$, $|\mathrm{P}_{\mathrm{Set_{FAKE}}}|$, and $|\mathrm{P}_{\mathrm{Set_{VERIFIED}}} \cap \mathrm{P}_{\mathrm{Set_{FAKE}}}|$, of RFDs extracted from the datasets Verified, Fake, and Verified ∩ Fake accounts, respectively.

Table 4.7 describes the number $|\mathrm{P}_{\mathrm{Set_{REAL}}}|$, $|\mathrm{P}_{\mathrm{Set_{FAKE}}}|$, and $|\mathrm{P}_{\mathrm{Set_{REAL}}} \cap \mathrm{P}_{\mathrm{Set_{FAKE}}}|$ of RFDs extracted from the Real, Fake, and Real ∩ Fake accounts, respectively.

The column *Thrs* shows the thresholds used by the RFD discovery algorithm to compute distances between tuples. A *Thrs* value 0 corresponds to a traditional FD. From what said above, key dependencies have been discarded, since they are not useful to discriminate fake accounts.

Thrs	$\lvert\mathrm{P}_{\mathbf{Set_{VERIFIED}}}\rvert$	$\lvert\mathrm{P}_{\mathbf{Set_{VERIFIED}}} \cap \mathrm{P}_{\mathbf{Set_{FAKE}}}\rvert$	$\lvert\mathrm{P}_{\mathbf{Set_{FAKE}}}\rvert$
0	79	3	65
1	19	8	20
2	31	6	24
3	52	5	44
4	76	5	47
8	85	6	26
12	13	6	13

Tabella 4.6: Number of RFDs extracted from Verified vs Fake datasets.

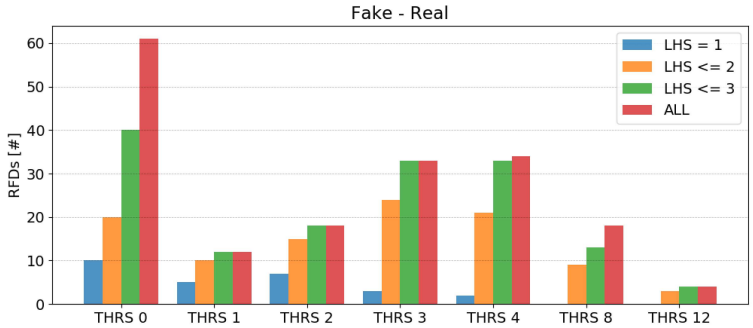
Thrs	$\lvert P_{\mathbf{Set_{REAL}}} \rvert$	$\lvert P_{\mathbf{Set_{REAL}}} \cap P_{\mathbf{Set_{FAKE}}} \rvert$	$\lvert P_{\mathbf{Set_{FAKE}}} \rvert$
0	62	4	65
1	17	8	20
2	17	6	24
3	36	11	44
4	30	13	47
8	41	8	26
12	8	6	13

Tabella 4.7: Number of RFDs extracted from Real vs Fake datasets.

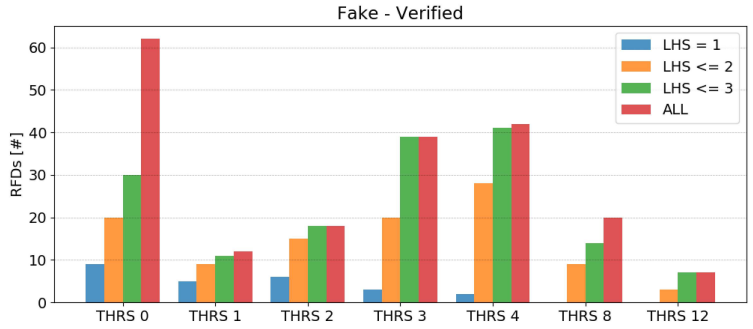
As prescribed by the proposed technique, experiments were accomplished by discarding key dependency and by considering different thresholds.

As Tables 4.6 and 4.7 show, the considered datasets share a reduced number of RFDs. Thus, it is possible to test the genuineness of newly generated accounts by checking whether the number of RFDs shared among the considered datasets keeps low also after their creation. Figure 4.29 groups the extracted RFDs according to the cardinality of their LHSs. Obviously, it is simpler to explain the discriminating property related to RFDs with only one attribute on the LHS. To this end, Figure 4.29(a) shows that the number of RFDs with LHS cardinality 1 that hold on fake but not on real accounts drops to 0 for thresholds values above 4. Similar considerations apply for the RFDs with LHS cardinality 1 that hold on Fake but non on Verified accounts (Figure 4.29(b)). On the other hand, taking into consideration the number of RFDs holding on real accounts but not on the union of fake and verified ones (Figure 4.29(c)), it is possible to notice that in most cases, there are no attributes on the LHS. The only two exceptions occurred when the relaxation threshold was set to 2 and 3.

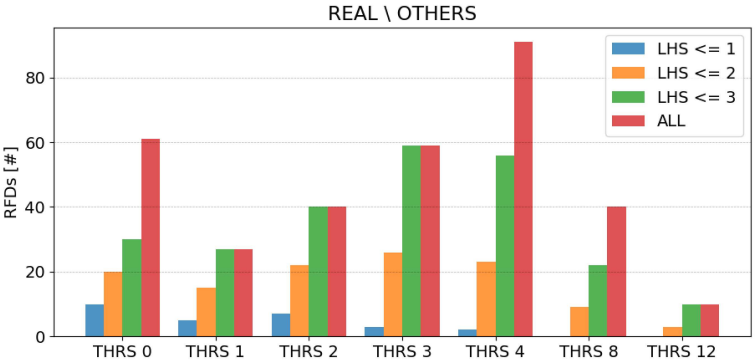
Furthermore, it has been highlighted the impact of each attribute on the LHS (respectively RHS) by counting the number of times an attribute appears on the LHSs (respectively RHSs) of all discovered RFDs. Results in terms of percentages are shown in Figure 4.30. In particular, by considering RFDs of fake accounts



(a) Number of RFDs holding on Fake vs Real accounts.

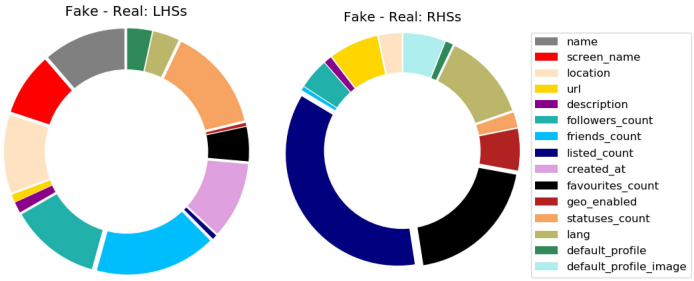


(b) Number of RFDs holding on Fake vs Verified accounts.



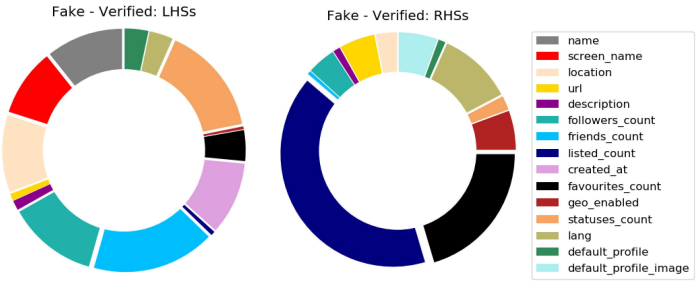
(c) Number of RFDs holding on Real vs Verified and Fake accounts.

Figura 4.29: Experimental results of the proposed technique by varying thresholds.



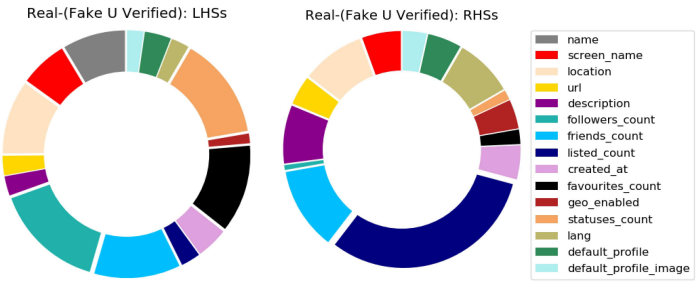
(a) Percentage of LHSs for holding RFDs on Fake vs Real accounts.

(b) Percentage of RHSs for holding RFDs on Fake vs Real accounts.



(c) Percentage of LHSs for holding RFDs on Fake vs Verified accounts.

(d) Percentage of RHSs for holding RFDs on Fake vs Verified accounts.



(e) Percentage of LHSs for holding RFDs on Real vs Fake and Verified accounts.

(f) Percentage of RHSs for holding RFDs on Real vs Fake and Verified accounts.

Figura 4.30: Percentage of incidence of LHSs (left) and RHSs (right) for RFDs holding on the different datasets.

but not of real ones, it is possible to notice that the attributes `statuses_count` and `friends_count` represent the ones that have the greatest impact when they appear on LHSs (Figure 4.30(a)). Instead, as shown in Figure 4.30(b), the most determined attribute (RHSs) is `listed_count`, even if also `favourites_count` appears many time as RHS. In general, by comparing Figures 4.30(a) and 4.30(b), it is possible to notice that the impact on the LHSs is similar for several attributes; instead, a wider difference appears for attributes on the RHSs. Similar considerations can be provided referring to RFDs holding on fake accounts, but not on verified ones, for both LHSs and RHSs, respectively (Figure 4.30(c) and Figure 4.30(d)). Finally, taking into consideration RFDs holding on real accounts, but not on the union of fake and verified ones, it is possible to notice that the attributes `followers_count` and `statuses_count` present the greatest impact w.r.t. other attributes, in the case of LHSs (Figure 4.30(e)). Instead, as shown in Figure 4.30(f), the greatest impact for RHSs is obtained by the attribute `listed_count`. In this case, also `friends_count` is determined many times, whereas all other attributes have little impact on percentage.

As a further evaluation, three statistical measures have been analysed, such as max, min, and average support of RFDs, when thresholds change. This analysis is shown in Figure 4.31, where Max indicates the distribution of maximum values, Min the one of minimum values, and AVG the one of average values by varying thresholds. In detail, referring to the variation of RFDs' support on fake accounts, but not on real ones (Figure 4.31(a)), it is possible to notice that the trend for Max exhibits a linear growth, except for the threshold 1, whereas for Min and AVG it follows a sub-linear growth. Furthermore, referring to the variation of RFDs' support on fake accounts, but not on verified ones (Figure 4.31(b)), it is possible to notice that the trend for Max exhibits a linear growth except for the threshold 1, 8, and 12, but it remains sub-linear for Min and AVG. Finally, taking into consideration the variation of RFDs' support on real accounts, but not on the union of fake and verified ones (Figure 4.31(c)), it is possible to notice

that the trend for Max exhibits a linear growth except for the threshold 1, 8, and 12, whereas for AVG alternates growth and degrowth, and for Min, it follows a sub-linear growth.

In what follows, some meaningful RFDs selected by applying the proposed technique on the fake accounts dataset are shown. In particular, according to the proposed RFD ranking technique, RFDs with higher LHS *support* have been initially considered:

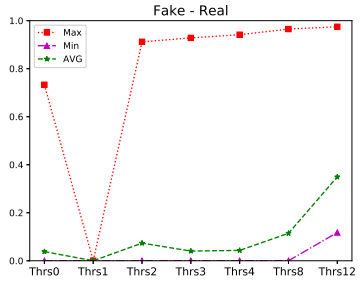
- `followers_count`, `statuses_count`, `default_profile` → `default_profile_image`
- `created_at` → `listed_count`
- `followers_count`, `statuses_count` → `listed_count`

The first RFD highlights the fact that the goal of several bots has only been to spread malicious advertising. Consequently, they do not care about the Twitter profile; instead, they use it without a profile image, and without applying changes to the default profile. By examining the second RFD, it is possible to conclude that often scripts automatically generate and manage fake accounts, and when registering them, they do not randomize the number of groups. Finally, the third RFD emphasizes this aspect further, since it reveals more characteristics of fake accounts based on the "`followers_count`" and "`statuses_count`".

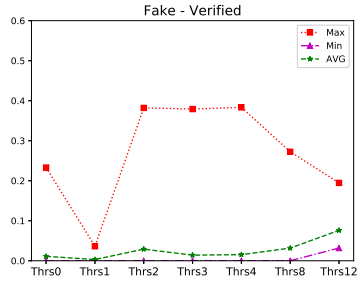
Similar considerations apply for the real account dataset defined in Section 4.3.1. In particular, the RFDs holding on this dataset reveal that automatic procedures cannot emulate human behaviour. The following RFDs are the most relevant ones among those holding on this dataset:

- `name`, `description`, `default_profile` → `lang`
- `name`, `favourites_count` → `listed_count`
- `name`, `followers_count` → `listed_count`

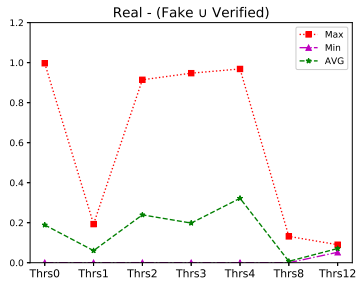
The first RFD is helpful to discriminate humans, since the language is a typical characteristic of a person, implying his/her way



(a) Variation of the support for thresholds on Fake vs Real accounts.



(b) Variation of the support for thresholds on Fake vs Verified accounts.



(c) Variation of the support for thresholds on Real vs Verified and Fake accounts.

Figura 4.31: Experimental results of the support metric by varying thresholds.

of writing. Analogously, by examining the second RFD it is possible to conclude that the number of likes that a person assigns to Twitter posts implies his/her `listed_count`. Finally, the third RFD reveals that attributes `name` and `followers_count` imply the list of friends of a person. Such RFDs only hold on the real account dataset, hence from their analysis it is possible to conclude that automatic procedures can hardly emulate the different human behaviours.

4.3.3 Fake account classification by using RFDs

This section describes how the application of RFDs can represent a useful mean to select features for classifying fake accounts, outperforming machine learning models in terms of accuracy score.

Settings. A dataset in which fake and real accounts are randomly merged has been composed, by also adding an additional feature, i.e., "the class label", labeling fake and real accounts. In particular, the "the class label" is added to the list of features presented in Table 4.5. In this way, it is possible to analyse the dataset in terms of classification purposes.

To use RFDs as a feature selection strategy, it is possible to consider the RFDs automatically extracted from data, by using several attribute comparison thresholds $Thrs$: 0, 1, 2, 3, 4, 8, and 12. In detail, all attributes appearing on the RHS of each RFD have been collected, i.e. all the attributes determined by some other ones at least one time. According to the collected attributes, for a specific threshold it is possible to consider one attribute at a time, delete it from the dataset, and compute the accuracy scores by applying different supervised classification models. In this way, it is possible to analyse how the classification scores vary between the complete dataset (i.e., without removing any attribute, named Baseline) and the dataset with one of the selected attributes filtered out (i.e., by performing the classification task with $n-1$ features, where n is the number of features of the complete dataset).

More specifically, Decision Tree [85], Random Forest [86], Naive Bayes [87], Support-vector Machine (SVM) [88], and Linear Model [89] have been used as supervised classification models, and have been implemented in the well known scikit-learn[5] python library. Finally, the classification scores have been computed by using a 10-fold cross-validation strategy.

Results. Figure 4.32 shows experimental results in terms of accuracy scores per classifier, obtained over the complete dataset (Baseline), or by removing one attribute from it as described abo-

[5]https://scikit-learn.org/stable/#

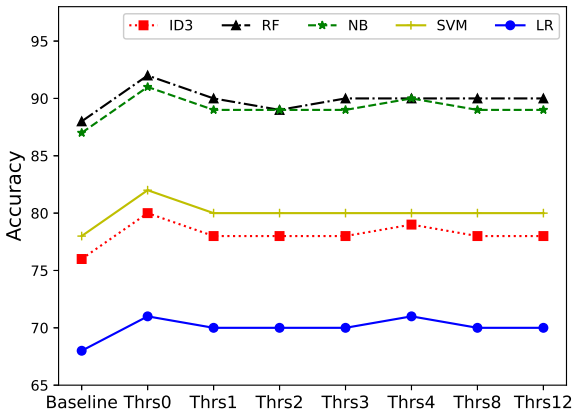


Figura 4.32: Classification score by varying thresholds.

ve, and in accordance with RFDs extracted from data, for each considered *Thrs*. More specifically, for each threshold, the best accuracy reached is reported. Instead, each line in the graph represents a classifier, *ID3* (Decision Tree), *RF* (Random Forest), *NB* (Naive Bayes), *SVM* (Support-vector Machine), and *LR* (Linear Model).

It is possible to notice that threshold *Thrs0* performs better than the other ones. This is probably due to the fact that the selected attributes were almost always boolean, and *Thrs0* can better characterise this type of attributes. In general, it is worth to notice that the usage of RFDs registered an improvement of the accuracy score w.r.t. the Baseline for all thresholds. Thus, it is possible to state that RFDs can offer promising results in fake accounts discrimination and are able to well characterise correlations among data, yielding the possibility to apply them for classification activities as feature selection strategy.

Capitolo 5

A Methodology for GDPR compliant information confidentiality preservation

Nowadays, new laws and regulations, such as the European General Data Protection Regulation (GDPR) [90], require companies to define privacy policies complying with the preferences of their users. The regulation prescribes expensive penalties for those companies causing the disclosure of sensitive data of their users, even if this occurs accidentally. Thus, it is necessary to devise methods supporting companies in the identification of privacy threats during advanced data manipulation activities. Cryptographic and anonymization techniques can be useful to mask the data but in the most cases they are or not applicable in practice (i.e., they are computationally expensive) or they do not permit to expoilt data for business activities, such as analytic ones. To this end, in this chapter, a methodology exploiting relaxed functional dependencies (RFDs) to automatically identify data that could imply the values of sensitive ones is described [9]. With respect to the literature proposals (see Section 2.2), the proposed methodology preserves the information confidentiality by reducing the amount of data to be encrypted, hence increasing data usage. In particular, it permits to partially encrypt data according to the different

privacy preservation requirements that a user could specify. An experimental evaluation demonstrates the effectiveness of the proposed methodology in increasing compliance to GDPR data privacy, while reducing the set of values to be partially masked, hence enhancing data usage.

5.1 General Data Protection Regulation

The General Data Protection Regulation (GDPR) prescribes how companies must process and manage private data [91] of their users, aiming to offer significant improvements to the regulatory environment of companies and institutions. In particular, GDPR establishes a uniform framework for data protection legislation across nations belonging to the European Community, without having to comply with the regulations of the single governments. This represents a significant advantage for companies operating across multiple countries of the European Community. Furthermore, even companies located outside the European Community must abide by the GDPR if they manage data of European users.

GDPR classifies any information related to individuals as *personal data*, without prescribing the usage of specific methodologies/technologies. Even if personal data are obscured and/or partially encrypted, the organization managing them still incurs into violations if it is possible to disclose users' sensitive data during some big data processing activities, such as data integration, entity resolution, and so on. The central concept underlying GDPR concerns the "user agreement", i.e. the specification of how users' data should be processed through an explicit declaration, which should be freely given, specifically informed, and unambiguous.

More specifically, GDPR prescribes the following two activities: (i) the adoption of a privacy preservation methodology, and (ii) the definition of default policies to preserve the privacy of any user. Thus, according to the first activity, organizations need to employ a privacy preservation methodology from the design to

Name	Surname	SSN	Age	Street	Native-Country	Occupation	Sex	City	ZipCode
Katherine	Swavely	029-32-6730	35	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Matthew	Costabile	475-96-3980	58	SARGENT ST	United-States	Unemployed	M	San Francisco City	94132
Jarrett	Albarado	214-20-7035	49	MARNE AVE	England	Worker	M	Newburgh	12550
Rowena	Hemeyer	481-98-9042	79	ESQUINA DR	United-States	Retired	F	San francisco	94134
Corina	Torris	490-03-6515	34	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Carlotta	Bracker	659-05-8786	32	ALOHA AVE	United-States	Unemployed	F	Pearl City	96782
Zane	Bracker	678-14-8279	32	PALOS PL	United-States	Teacher	M	Illinois	60464
Joselyn	Bracker	004-03-6265	32	PALOS PL	United-States	Teacher	M	Illinois	60464
Sherry	Swavely	400-20-9834	80	ALOHA AVE	United-States	Retired	F	Pearl City	96782
Matthew	Costabile	255-73-7429	24	SARGENT ST	England	Unemployed	M	San Francisco City	94132

Tabella 5.1: A database storing customers' information.

the development of their services. Instead, the second activity prescribes the implementation of proper default methodologies/-technologies to guarantee data processing in a trusted way. These prescriptions aim to provide a friendly privacy setting, by also exploiting the possibility to adopt default settings.

Concerning the possibility to share personal data, the GDPR is not limited to the European Economic Area (EEA)[1], since when data are transferred outside the EEA, all privacy preservation policies defined on data are transmitted along with the data themselves. Moreover, GDPR is composed of several recitals addressing the privacy preservation issues to specific activities, such as marketing, user profiling, data integration, and so on. Among all recitals defined in the GDPR, the recital 71 expresses, in a summarised way, the fact that a company performing analytical activities should use appropriate mathematical or statistical procedures, and implement technical and organizational measures to ensure the privacy of data related to a physical person [92].

GDPR has become effective since May 25th, 2018. To this end, by offering the possibility to manage different policy requirements concerning single users, the proposed methodology turns out to be particularly useful in pursuing GDPR compliant privacy preservation.

[1]European Economic Area (EEA), includes all European Community countries, and Island, Liechtenstein, and Norway

5.2 Problem description

According to the GDPR, companies and organizations can use sensitive data only for business application purposes, avoiding their exposure to third parties, or their transfer to commercial activities, such as user profiling. All activities affecting the confidentiality of data have to be considered as data privacy violations. To this end, there is always the need to pay attention to data representations that might refer to users.

Example 2. *Considering the database shown in Table 5.1, which represents a portion of the census income dataset, containing the following data of citizens:* `Name`, `Surname`, `SSN`, `Age`, `Address`, `Native-Country`, `Occupation`, *and* `Sex`. *It represents a sample dataset in which data, especially those referring to individuals, need to be managed by preserving their privacy.*

In a scenario where companies manipulate and store individuals' data, defining new privatization methodologies is important to help them comply with the GDPR.

Data privacy concerns several aspects, among which the focus of this dissertation is on *Information Confidentiality* (IC). The latter is a general privacy preservation concept by which users request to preserve the confidentiality of their specific data, also referred to as *sensitive data*, aiming to protect them against unauthorized accesses [47].

In what follows, the concept of information confidentiality in the context of relational databases is formalized.

Definition 7. (Information confidentiality). *Given a relational database schema $\mathcal{R}$, defined on a set of attributes $attr(\mathcal{R} = \{A_1, \dots A_n\})$, an instance r of it, where each tuple t over r represents a single user, and its projection $t[Y]$ onto $Y \subseteq attr(\mathcal{R})$ the data s/he defines as sensitive, then ensuring the information confidentiality for t requires that i) $t[Y]$ is masked, and ii) no subset of data $t'[Y']$ permits to disclose any value in $t[Y]$.*

Starting from Definition 7, it is possible to derive the concept of *data usage* for the specific context, that is: a data can be used

without jeopardising the privacy of any user if and only if i) it has not been declared as sensitive by its owner, and ii) it cannot be used to disclose any other data declared as sensitive.

In what follows, the *information confidentiality* problem issues are formalized in terms of attribute correlations expressed through RFDs, yielding the concept of *confidentiality-violating attribute set*.

Definition 8. (Confidentiality-violating attribute set). *Given a relational database schema $\mathcal{R}$, an instance r of it, and two attribute sets $X, Y \subseteq attr(\mathcal{R})$, where $Y = \{Y_1, \ldots, Y_h\}$ is the set of data defined as sensitive, then X is a confidentiality-violating attribute set if and only if it is not a key, and there exists $Y_i \in Y$ that is the RHS of an RFD holding on r and having X as LHS.*

According to Definition 8, a relational database schema $\mathcal{R}$ preserves the information confidentiality if and only if: (i) $\mathcal{R}$ contains all the user-specified sensitive attributes in a masked form, and (ii) $\mathcal{R}$ does not contain confidentiality-violating attribute sets. In other words, if the user specifies a set of sensitive attributes, other than obscuring them, there is also the need to prevent the possibility to derive their values from other attribute values. For instance, a sensitive attribute might be derived by the LHS of an RFD φ in which it appears as RHS. In this case, it is possible to say that the LHS of φ determines the RHS. Thus, given a sensitive attribute A, knowing the values of attributes determining A, a third party could infer the values of A with high certainty and accuracy degrees according to the thresholds of φ. As a consequence, there is the need to identify all the confidentiality violating attribute sets, that is, all the attribute sets functionally determining sensitive attributes.

5.3 The Methodology

From the discussion above, it is clear that the GDPR might be a serious burden, especially for big companies managing huge volumes of data concerning their customers. By referring to the

scenario shown in Table 5.1, a solution could be to obscure all data, by means of cryptographic techniques. However, in this way a company could never use such data, even those that are not sensitive, and would have to deal with computationally expensive encryption processes, e.g. not all the data shown in Table 5.1 can be considered as sensitive. Moreover, by manually specifying both sensitive data and those from which they can be derived could require a huge effort when managing big data collections. To this end, a new methodology has been proposed, aiming to reduce the number of attributes to be encrypted while pursuing information confidentiality, hence maximizing data usage. In particular, the methodology exploits attribute correlations expressed in terms of Relaxed Functional Dependencies (RFDs) [70] to identify attribute sets from which sensitive data can be derived.

More specifically, the proposed methodology exploits algorithms to automatically discover RFDs from data [69, 93], together with ranking techniques to decide their application order, aiming to derive a minimal set of attributes to encrypt for pursuing information confidentiality.

Given a relational database schema $\mathcal{R}$, and an instance r of it, there is the need to identify the set $X_\Xi = \{X_{\zeta_1}, \ldots, X_{\zeta_n}\}$ of all confidentiality violating attribute sets X_{ζ_i} within $\mathcal{R}$, and define a way to make each of them not accessible. To this end, the following types of RFDs holding on r have been considered:

$$\left[X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} A_{\Phi_2}\right]_{\mathbb{D}_{\text{TRUE}}} \tag{5.1}$$

where $A \in attr(\mathcal{R})$, aiming to search for the LHSs of RFDs having a sensitive attribute on their RHS.

More formally, in order to preserve the confidentiality of $\mathcal{R}$, there is the need to identify the minimal set of attributes $Z \subseteq attr(\mathcal{R})$ such that there exists no valid RFD $X_{\zeta_i} \backslash Z \rightarrow A$, with A sensitive attribute of $\mathcal{R}$. In other words, it is necessary to invalidate all the RFDs having a user-specified sensitive attribute as RHS. The set of user-specified sensitive attributes is also named *IC-attribute set*.

In order to automatically derive the minimal set Z of attributes to be removed, there is the need to use a heuristic. This is due to the fact that this problem is NP-complete, since the Minimum Feedback Vertex Set [94], which is the problem of finding the smallest set of vertices to be removed from an undirected cyclic graph to make it acyclic, can be reduced to it. In particular, each $X_{\zeta_i} \in X_\Xi$ can be modeled as a cycle in an undirected graph, where the vertices of the cycle are the attributes in X_{ζ_i}. Thus, given an undirected graph G with one or more cycles, the vertices of a cycle can be seen as a confidentiality-violating attribute set X_{ζ_i}. Thus, solving the problem of finding the minimal set Z defined above would also solve the MFVS one.

Heuristics. Three heuristics have been defined: (i) the *counting* heuristic, scoring the number of X_{ζ_i} containing a given attribute; (ii) the *weighted counting*, similar to the counting heuristic, but instead of adding a 1 for each X_{ζ_i} in which an attribute appears, it adds $1/|X_{\zeta_i}|$, which represents the weight of the attribute over X_{ζ_i}; and (iii) the *MFVS* heuristic, derived from an approximate solution for the MFVS problem, which is based on the Depth First Search (DFS) visit to approximately evaluate the number of times a node is involved in a cycle.

In particular, the first two heuristics associate scores to the attributes belonging to the confidentiality-violating attribute sets in X_Ξ, eliminating them in descendant order of their score, until all the RFDs associated to X_Ξ are invalidated. Instead, as mentioned above, by using the third heuristic an undirected graph is produced. In particular, the heuristic scores each node with the number of backward edges encountered during a DFS visit. More specifically, the well-known DFS visit has been adapted to count backward edges. Then, the heuristic removes nodes in descendant order of their score, until no more cycles exist in the graph. For all the three presented heuristics, a basic case is represented by RFDs with one attribute on the LHS, since it is possible to remove it without considering any score.

As an example, given the following two RFDs:

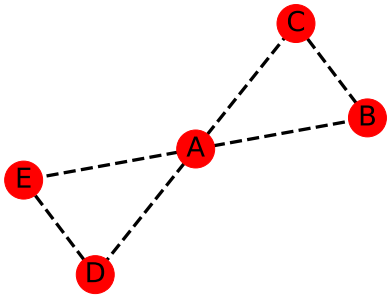


Figura 5.1: MFVS problem associated to (4) and (5).

$$A\ B\ C \rightarrow F \tag{5.2}$$
$$A\ E\ D \rightarrow F \tag{5.3}$$

where F is a confidential attribute. If they are the only RFDs with F on the RHS, there is the need to encrypt some of the attributes on their LHSs together with F, in order to guarantee the information confidentiality of F. By applying the counting heuristic defined above, attribute A has a score of 2, whereas each of the remaining attributes has a score of 1. Thus, it is first removed A, which already invalidates both RFDs (4) and (5). A similar action is decided upon applying the weighted counting heuristic, since attribute A has a score of 2/3, whereas the remaining ones have a score of 1/3 each. Finally, the MFVS heuristic could be used to solve the MFVS problem of the graph in Figure 5.1, yielding the deletion of the vertex A only, since it is the node with the maximum number of backward edges derived from the DFS visit. Moreover, upon removing A, the resulting graph is acyclic. Thus, in all three cases, A will be the only attribute to be encrypted together with F.

Example 3. *Given the database of customers shown in Table 5.1, and supposing that a user wants to "obscure" the* `Occupation` *at-*

tribute in order to preserve his/her privacy. In this case, there are three attribute sets determining `Occupation`*, i.e.* `Name`*,* {`Age`*,* `Sex`}*, and* {`Age`*,* `Street`}*, since they are the LHSs of all the RFDs holding on the considered relation, and having attribute* `Occupation` *as RHS. To guarantee information confidentiality, besides "obscuring" the attribute* `Occupation`*, also attribute* `Name` *should be "obscured", since it determines attributes* `Occupation` *and* `Age`*, based on the defined heuristics.*

Partial encryption. The cryptographic technique used in the proposed methodology is *block cipher* [95]. The latter is a method using a secret key to encrypt text (to produce ciphertext). In particular, it applies encryption to blocks of data (e.g., 64 contiguous bits) rather than to one bit at a time. Formally, a block cipher is a permutation with a key that can be efficiently computed, i.e. $\mathcal{F} : \{0,1\}^n \times \{0,1\}^l \rightarrow \{0,1\}^l$, such that, given a key k and a block of data x to be encrypted

$$F_k(x) \stackrel{\text{def}}{=} F(k,x) \text{ is a permutation} \tag{5.4}$$

where n is the length of k, l the length of x, and $\mathcal{F}_k$, ${\mathcal{F}_k}^{-1}$ must be efficiently computed.

In particular, given the set $X^i = X^i_1, \ldots X^i_m$ of user-specified "sensitive" attributes, together with those derived through RFDs, the block cipher has been applied to all X^i. It is worth to notice that each X^i is encrypted with a different secret parameter k, which is the user's secret parameter that permits to decrypt his/her sensitive data. This implies that it is possible to have a database containing both visible and encrypted data, that is still privacy-preserving. The block cipher guarantees security with respect to the Chosen Plaintext Attacks (CPA-security) [96].

Example 4. *Given the database of customers shown in Table 5.1, and supposing that the last five of them required attribute* `Occupation` *to be confidential. As shown in Table 5.2, by applying the proposed methodology it is possible to obtain partial encryption, where the values denoted as "*****" are encrypted, as explained in the previous examples.*

Name	Surname	SSN	Age	Street	Native-Country	Occupation	Sex	City	ZipCode
Katherine	Swavely	029-32-6730	35	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Matthew	Costabile	475-96-3980	58	SARGENT ST	United-States	Unemployed	M	San Francisco City	94132
Jarrett	Albarado	214-20-7035	49	MARNE AVE	England	Worker	M	Newburgh	12550
Rowena	Hemeyer	481-98-9042	79	ESQUINA DR	United-States	Retired	F	San francisco	94134
Corina	Torris	490-03-6515	34	ALOHA AVE	United-States	Employee	F	Pearl City	96782
*****	Bracker	659-05-8786	*****	ALOHA AVE	United-States	*****	F	Pearl City	96782
*****	Bracker	678-14-8279	*****	PALOS PL	United-States	*****	M	Illinois	60464
*****	Bracker	004-03-6265	*****	PALOS PL	United-States	*****	M	Illinois	60464
*****	Swavely	400-20-9834	*****	ALOHA AVE	United-States	*****	F	Pearl City	96782
*****	Costabile	255-73-7429	****	SARGENT ST	England	*****	M	San Francisco City	94132

Tabella 5.2: A privacy-preserving database of customers' information.

Overview of third parties. In what follows, the robustness of the proposed methodology is analyzed, by considering the power of third parties in disclosing values of attributes specified as confidential. In particular, it is possible to prove how RFDs can help identify confidentiality threats, by also analyzing several critical scenarios.

One of the main properties of RFDs is *minimality* [69], which concerns both the number of attributes on their LHS and the associated similarity thresholds. For the critical scenarios analyzed below, the only concern is on how the minimality property is related to the LHS attributes. Let r be an instance of a relational database schema $\mathcal{R}$, and $\varphi : X \rightarrow Y$ a minimal RFD holding on r, then for each $A \in X$, $\varphi' : X \backslash A \rightarrow Y$ does not correspond to a RFD holding on r. In general, RFD discovery algorithms aim at finding the set of all minimal RFDs holding on a given dataset.

Before detailing the sample scenario, the preliminaries of the third parties considered for the proposed threat model are introduced. Supposing that a third party can access:

- the dataset structure together with metadata concerning the value distribution of each attribute;
- the set of all minimal RFDs holding on the dataset;
- the dataset partially encrypted according to the proposed methodology.

Moreover, assume it is possible that the third party can ask an oracle all the information defined above by simply providing the

name of the dataset. In particular, the value distributions enable the third party to know all possible values that an attribute can assume, whereas the set of minimal RFDs holding on the unencrypted dataset enables the third party to catch the data validating possible RFDs. Notice that, the term dataset has also been used referring to the ones obtained as a result of data integration, data augmentation, or any other big data processing task.

By considering the characteristics of this threat model, it is possible to reduce the likelihood of success for a third party to the safest scenario, i.e. a totally encrypted dataset. Thus, the likelihood of success for a third party in disclosing a target value can be reduced to a random guess on the value distribution it belongs to. In other words, even when a dataset is completely encrypted, the third party can try to disclose the target value by only choosing one of the values of its distribution. To this end, in what follows, it is possible to show that the target is to reduce the likelihood of success of the third party to a random guess on the value distribution of each IC attribute.

Example 5. *Given the sample dataset* 1 *shown in Table 5.2(a), for which it is possible to assume that there is only one IC attribute for the tuple* t_1*, e.g. attribute* D*. This means that the owner of* t_1 *requires confidentiality for the value* $t_1[D]$*. According to the proposed methodology, RFDs implying attribute* D *have been considered. Thus, the only RFD to be considered among those holding on the given dataset is* $\varphi : AB \rightarrow D$*. Then, if only the value* $t_1[D]$ *is obscured, it could still be derived from the correlation expressed by* φ*. In fact, by looking at tuple* t_2*, a third party could infer the value on* $t_1[D]$ *through the similarity between* t_1 *and* t_2 *on the combination of values for attributes* A *and* B*.*

Example 5 shows how minimal RFDs can be used to solve some issues concerning the third parties' derivation process. In fact, since $\varphi : AB \rightarrow D$ is minimal on the dataset shown in Table 5.2(a), then by masking $t_1[A]$ or $t_1[B]$ would guarantee that a third party could not derive the value $t_1[D]$, since both $\varphi' : A \rightarrow D$ and $\varphi'' : B \rightarrow D$ do not hold on the considered dataset. For this

(a) Sample dataset 1

	A	B	C	D
t_1	1	2	7	True
t_2	1	2	8	True
t_3	3	2	6	False
t_4	1	3	9	True

(b) Masked dataset 1

	A	B	C	D
t_1	***	2	7	***
t_2	1	2	8	True
t_3	3	2	6	False
t_4	1	3	9	True

(c) Sample dataset 2

	A	B	C	D
t_1	1	2	7	True
t_2	3	2	8	False
t_3	3	2	6	False
t_4	1	4	9	False

(d) Masked dataset 2

	A	B	C	D
t_1	***	2	7	***
t_2	3	2	8	False
t_3	3	2	6	False
t_4	1	4	9	False

Tabella 5.3: A sample scenario.

reason, by simply observing the values of A (or B), a third party could not derive the value $t_1[D]$, since φ' (φ'') is not valid.

Example 6. *Starting from the scenario described in example 5, and assuming that the methodology prescribes to mask attribute A to break the attribute correlation expressed by $\varphi : AB \rightarrow D$, as shown in Table 5.2(b), the third party can only observe the free values of attribute B and consider the tuples that are similar to $t_1[B] = 2$, which means all tuples. In particular, the dataset contains the value $'True'$ for $t_2[D]$ and the value $'False'$ for $t_3[D]$. Thus, the third party can only try a random guess, which in this case is equivalent to a coin toss. Similar considerations apply when the value $t_1[B]$ is masked and $t_1[A]$ is not.*

In what follows, a borderline case of the aforesaid scenario is analyzed, which is the only one jeopardising the proposed methodology to the risk of value disclosures.

Example 7. *Given the sample dataset shown in Table 5.2(c), and suppose that there is only one IC attribute for the tuple t_1, e.g. attribute D. Consequently, the only RFD to be considered, among*

those holding on the given dataset, is $\varphi : AB \rightarrow D$. Suppose that the methodology prescribes to mask attribute A in order to break the attribute correlation expressed by φ, as shown in Table 5.2(d). If a third party knew that φ is a minimal RFD holding on the dataset, then s/he would be aware that $\varphi' : B \rightarrow D$ did not hold on the dataset. Furthermore, since the value distribution of attribute D is limited to $\{True, False\}$, and the tuples similar to t_1 on B are t_2 and t_3, which have the value $False$, a third party could exactly infer the value of $t_1[D]$, since the only violation invalidating the RFD φ' can be generated from the value $'True'$.

In general, this case can occur only when the RFD violation is caused by the attribute value declared as confidential, which has been obviously masked. However, although this borderline case occurs rarely, there is the need to undertake additional actions in order to guarantee the requested confidentiality.

More formally, let A be an attribute, and t be a tuple for which $t[A]$ is declared as confidential, then a third party can infer the masked value $t[A]$ with higher likelihood than a random guess, if and only if:

1. A third party knows the minimal RFDs holding on the unencrypted dataset, hence s/he can also infer the non-holding RFDs by looking at the partially encrypted dataset;

2. There exists a set of attributes X such that $\varphi' : X \rightarrow A$ does not hold on r, but it holds on $r \backslash t$, and there exists a non-empty set of tuples s whose projection on X is similar to $t[X]$, and all tuples in $r \backslash t$ share the same value of A.

In fact, in this case, the reason why φ' does not hold on r can only be that the masked value $t[A]$ is different from that of the tuples in s, hence the third party can discard that value from his/her guesses. To tackle this borderline case, the value of a further attribute on the LHS of the minimal RFD is encrypted.

Example 8. *By considering the scenario described in the example 7, where for the RFD $\varphi : AB \rightarrow D$ a borderline case is highlighted*

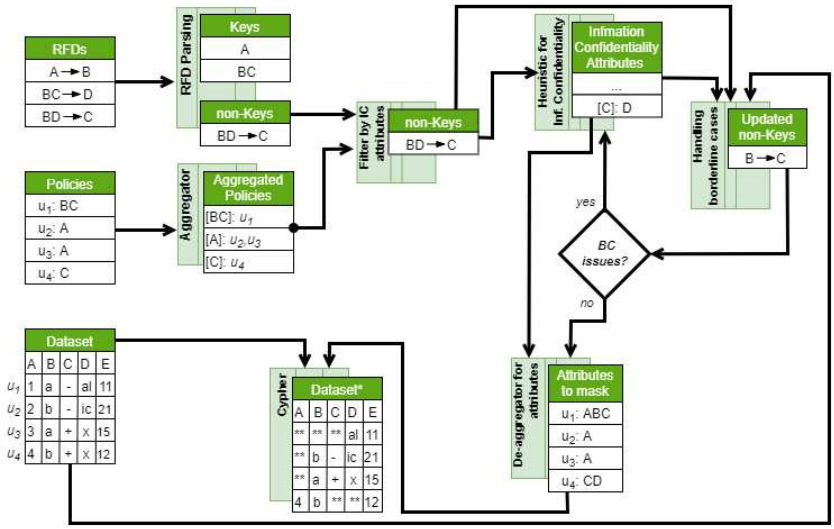


Figura 5.2: The general process for masking data according to users' policies.

(see Table 5.2(d)). According to the proposed methodology, also attribute B is encrypted. In this way, also the violation induced by tuple t_1 is masked, so that a third party could only give a random guess on the value distribution for attribute D.

5.4 The general Process

This section describes the general process of the proposed methodology and it provides a sample scenario.

Figure 5.2 shows how the proposed methodology can be applied to a generic scenario. The process starts by considering a given dataset, the set of RFDs holding on it, and a file containing several IC attribute sets, i.e. users' policies concerning attributes specified as confidential. The first step (RFD parsing) aims to filter out only non-key RFDs from the set of RFDs holding on the given dataset, since key RFDs cannot permit to determine any values, because all tuples differ on the attributes of their LHS, and

hence they are not a threat to confidentiality. Moreover, users are grouped according to the specified policies through an aggregator module. Then, for each specified policy, RFDs are filtered out by selecting those having one of the confidential attributes on their RHS (filter by IC attributes). All of their LHSs will represent the collection of confidentiality-violating attribute sets for the specific policy. Thus, one of the three heuristics defined above can be applied to retrieve the minimal set of attributes to be encrypted. Moreover, to verify whether the borderline case described in Section 5.3 occurs, it is necessary to check whether its two conditions are satisfied. To this end, there is the need to compute the set of non-holding RFDs, which is accomplished by removing attributes to be masked from the LHSs of the RFDs in which they are involved. If a resulting RFD reveals a borderline case, then its LHS will be added as confidentiality-violating attribute set. Once all borderline cases have been detected, the process iterates the application of a heuristic to derive the additional attributes to be encrypted. At the end of this process, a de-aggregator module permits to obtain the attributes to be encrypted for each user, according to his/her specified policy. Finally, the prescribed masking is applied to the entire dataset.

The pseudo-code of the proposed methodology is provided in Algorithm 1. It takes as input the dataset D to be masked, the set Σ of RFDs holding on D, and the set Λ of users specified policies, each defined in terms of IC attribute sets, and it returns as output the masked dataset D^*. At Lines $1-2$, the algorithm invokes the functions `REMOVE_KEY_RFDs` and `POLICY_AGGREGATOR` to remove key RFDs and group equal policies specified by different users. Then, for each specified policy, the algorithm performs the following steps: (i) it computes its associated *confidentiality-violating attribute set* (Line 4), (ii) it applies one of the three proposed heuristics to identify the additional attributes to be encrypted (Line 5); (iii) it verifies whether there exist *borderline cases*, by updating the confidentiality-violating attribute sets (Line 6), and iteratively repeating the application of heuristics until no more borderline cases exist (Lines $7-10$); and (iv) it runs the

Algorithm 1 The main algorithm

INPUT: A dataset D, a set of RFDs Σ, a set of policies Ψ
OUTPUT: A dataset partially encrypted D^*

```
1: Σ' ←REMOVE_KEY_RFDs(Σ)
2: Λ ←POLICY_AGGREGATOR(Σ',Ψ)
3: for each p_i ∈ Λ do
4:     X_ζ ← FILTER_BY_IC_ATTRIBUTES(Σ',p_i)
5:     Z ←GET_IC_ATTRIBUTES(X_ζ)
6:     X_ζ ← UPDATE_X_SET(Z)
7:     while BORDERLINE_CASE(X_ζ,D) do
8:         Z ← ADD_IC_ATTRIBUTES(X_ζ)
9:         X_ζ ← UPDATE_X_SET(Z)
10:    end while
11:    Ψ' ←DEAGGREGATOR_FOR_POLICY(Λ,Z,p_i)
12: end for
13: D* ←DATASET_ENCRYPTION(D,Ψ')
```

DEAGGREGATOR_FOR_POLICY function for mapping attributes to be encrypted to the data of users (Line 11). Finally, the encryption step is performed (Line 13).

Figure 5.3 shows a masked dataset resulting from the application of Algorithm 1 to the *CreditClient* dataset. In particular, to simulate the definition of users' policies, a module that randomly assigns confidential attributes to each user tuple has been implemented. It is possible to notice that at the end of the application of the proposed methodology, only few values are encrypted, whereas many others remain free. Moreover, it is worth to notice that many differences among the encrypted values are obtained. More specifically, the number of values encrypted for each tuple depends on (i) its associated policy, and (ii) the RFDs holding on the considered dataset. In this way, data declared as confidential can never be derived from *free* data. Thus, this strategy permits to limit the number of values to be encrypted in order to preserve information confidentiality, by increasing the possibilities to perform

data analytic processes.

Proof of correctness. In the following, the correctness of the proposed methodology is proved.

Theorem 5.4.1. *Each attribute value $t[A_i]$ defined as sensitive by user t is preserved after the application of the proposed methodology.*

Dimostrazione. It is possible to proceed by contradiction. Assume that the user t defines a sensitive value on the attribute A_i and a third party is able to disclose $t[A_i]$ after the application of the proposed methodology. To this end, according to the threat model described in Section 5.3.*c*, the third party can access i) the value distribution of each attribute $d(A_i)$, ii) the set Σ of all minimal RFDs holding on the dataset D, and iii) the partially encrypted dataset D^* resulting upon the application of the proposed methodology. The latter says that $t[A_i]$ is encrypted, together with other values on t. Thus, since $t[A_i]$ is encrypted on D^*, the third party has been able to disclose $t[A_i]$ by only using some free values on

A	B	C	D	E	F	G	H	I	J
LastName	FirstName	StreetAddress	CredentialType	Status	BirthYear	CEDueDate	FirstIssueDate	LastIssueDate	ExpirationDate
*****	Vasanti	39111 Janiya Harbors	*****	SUPERSEDED	1969	*****	20110927	20110927	20130927
*****	Kinda	93379 Rocky Knolls	*****	EXPIRED	1968	?	*****	20161109	*****
*****	Wassan	7661 Wolff Motorway	Nursing Assistant Registration	EXPIRED	1984	?	20091008	*****	20100826
*****	India	29040 Champlin Cape	Counselor Registration	EXPIRED	1957	?	20090224	20090224	*****
*****	*****	05672 Tyrese Turnpike	Nursing Assistant Registration	*****	1990	?	*****	20081110	20090413
*****	Stacey	723 Magali Ways	Licensed Practical Nurse	*****	*****	?	19931013	19961202	19961202
Roy	Casey	05663 Jarred Pine	Medical Assistant Registration	ACTIVE	1981	?	20180501	20190410	20210524
EVERETT	*****	7160 Tyreek Stream	*****	EXPIRED	1970	?	19930323	19930901	19930901
*****	Meaghan	46333 Brice Village	Registered Nurse License	*****	*****	*****	?	*****	*****
*****	Lauren	790 Billy Terrace	Social Worker Independent Clinical License	PENDING	1981	?	?	*****	?
*****	Melissa	*****	Counselor Agency Affiliated Registration	EXPIRED	*****	*****	20170913	20180608	20190706
*****	ROSALIND	023 Heaney Mission	Nursing Assistant Registration	EXPIRED	1947	?	*****	*****	20010317
*****	Debra	51018 Blaze Ways	Nursing Assistant Certification	ACTIVE	1953	?	20100209	*****	20200521
OLSON	JEAN	856 Rudy Knolls	Nursing Assistant Registration	EXPIRED	1951	?	20070327	20100127	20110127
BORST	AMY	*****	Health Care Assistant Certification	EXPIRED	1980	?	20021010	20041010	20041010
*****	Miriam	3281 Bart Creek	*****	EXPIRED	*****	?	20140818	20140818	20150910
Mathew	*****	*****	*****	EXPIRED	1965	?	19900227	19920514	19920514
*****	LILLIE	8421 Cayla Summit	Nursing Assistant Registration	EXPIRED	*****	*****	19911002	19950406	19950406
*****	JUDY	97813 Garret Pine	Health Care Assistant Certification	EXPIRED	1935	*****	19920501	19990304	*****
*****	Kimberly	337 Effertz Mountain	Nursing Assistant Registration	EXPIRED	1990	*****	20160823	20170912	*****
*****	Alyssa	61764 Shawn Spur	Medical Assistant Registration	ACTIVE	1997	*****	20171201	20181227	*****
Patel	Ankit	89136 Baron Parks	Pharmacist Intern Registration	EXPIRED	1991	?	*****	20180531	20190706
*****	Irina	*****	*****	ACTIVE	*****	?	*****	20190706	*****
*****	Kathryn	93355 Sylvester Ville	Counselor Certified Certification	ACTIVE	1977	*****	*****	*****	*****
Numata	Jaclyn	4762 Bogan Alley	Emergency Medical Technician Certification	INOPERABLE	1984	?	20120820	20151210	20191130
*****	ELLEN	*****	Emergency Medical Technician Certification	EXPIRED	1978	?	*****	20040202	20071031
Hatfield	*****	53117 Wiza Rue	Registered Nurse Temporary Practice Permit	EXPIRED	1974	?	20090731	20090731	20100131
*****	Patricia	*****	*****	EXPIRED	*****	*****	*****	19890810	19890810
*****	*****	1416 Crona Locks	Medical Assistant Certification	EXPIRED	*****	?	20130701	20140326	20160228
*****	Flavia	6368 Rhett Prairie	Registered Nurse Temporary Practice Permit	*****	*****	*****	20110729	20110729	20120129
STALLARD	IDABELLE	452 Brown Inlet	Registered Nurse License	EXPIRED	1926	?	19500925	19931031	19931031
*****	*****	*****	*****	EXPIRED	1959	*****	*****	19890328	19890328
WAHYUNI	RETNO	*****	Nursing Assistant Registration	*****	1958	?	20011029	20031002	20031002
*****	JUDY	427 Williamson Dam	Nursing Assistant Certification	EXPIRED	1969	?	19940307	20000824	*****

Figura 5.3: Masked dataset after the application of the proposed methodologies.

D^* and some of the RFDs in Σ. This could occur if and only if at least one of the two following cases occurs:

1. there exists a tuple t' containing a combination of free values, such that $t'[X]$ is similar to $t[X]$, and there exists an RFD $\varphi : X \rightarrow A_i$. This means that whenever two tuples (t, t') are similar on X, then almost always they are similar on A_i, yielding the possibility of determining $t[A_i]$ by looking at $t'[A_i]$;

2. there exists a combination of free values on all the tuples of D^*, such that $t[Z]$ is similar to any $t'[Z]$ on D^*, then all tuples in $D^* \setminus t$ are free and have a similar value on A_i, and it does not exist an RFD $\varphi' : Z \rightarrow A_i$ in Σ, but there exists at least one RFD $\varphi : X \rightarrow A_i$ in Σ such that Z is a direct subset of X, i.e. $ZB = X$ for at least an attribute $B \notin Z$ and $B \neq A_i$. This means that, since all tuples of D^* are similar on Z, and all tuples of $D^* \setminus t$ are similar on A_i, then only the tuple t represents a violation making $\varphi' : Z \rightarrow A_i$ not holding on D, yielding the possibility of determining $t[A_i]$ with a higher likelihood than a random guess, by looking at the value distribution of $d(A_i)$ and by excluding all the values similar to at least one $t'[A_i]$ on $D^* \setminus t$. This becomes a certainty for $|A_i| = 2$.

However, the third party is unable to exploit case 1), since the proposed methodology considers the LHS of each RFD in Σ that determines A_i as a confidentiality-violating attribute set, and it encrypts at least one attribute for each of them. Thus, no combination of free values can satisfy the LHS of any RFD in Σ that determines A_i. Moreover, the third party is unable to exploit case 2), since the proposed methodology considers it as *borderline case* and it forces the encryption of at least another attribute on the confidentiality-violating attribute set representing the LHS of an RFD revealing such borderline case. This implies that neither case 1) nor case 2) occurs, and a third party cannot exploit attribute

correlations and free values to disclose values declared as sensitive, contradicting the original assumption.

■

5.5 Experimental Evaluation

This section presents the experiments performed for evaluating the proposed methodology on several public datasets. The goal is to evaluate the performances of the three defined heuristics on different real-world datasets. This is due to the fact that they represent approximate solutions to the problem of finding the minimum number of attributes to be encrypted. For this reason, it is reasonable to expect that heuristics produce different results in terms of the number of attributes.

5.5.1 Experimental Settings

Implementation details. Several tools have been implemented in Java language to support the proposed methodology. In particular, to discover the RFDs holding on a given dataset the discovery algorithm defined in [69] has been used, and have been analyzed through the algorithm described in Section 5.4. The latter also implements the three heuristics *counting*, *weighted counting*, and *MFVS*, in order to select values to be partially encrypted. In particular, the values of selected attributes are encrypted with AES in Cipher-Block-Chaining (CBC) mode [96].

Datasets. Three public datasets have been considered [97], and have been augmented by artificially introducing some confidential data. In particular, attributes `Name`, `Surname`, and `StreetAddress` have been added by randomly selecting their values for all tuples. Statistics on the characteristics of the considered datasets are reported in Table 5.4.

Datasets	# Columns	# Rows	# FD	Size [KB]
CreditClient	10	30000	152	1730
Health	10	45250	72	4650
London	13	17414	514	1500

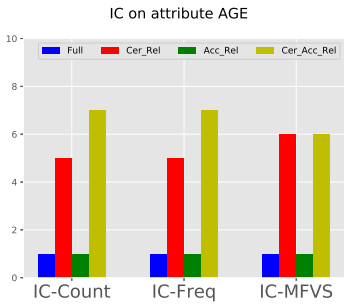
Tabella 5.4: Statistics on the datasets used in the evaluation.

5.5.2 Results

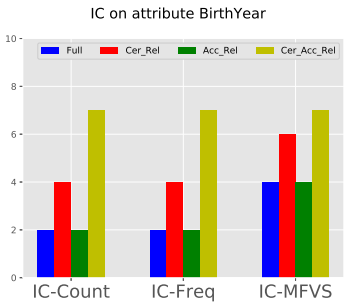
Evaluation session on individual datasets. A privacy preservation scenario has been defined, supposing that each user specified one attribute to be confidential. Moreover, for each of them, the proposed methodology has been used to derive the minimum *number of attributes* to be encrypted for guaranteeing users' privacy. This scenario has been evaluated through four experimental sessions, in which different sets of RFDs have been considered, according to several threshold settings. In particular, canonical FDs, RFDs relaxing on the extent only (total accuracy degree), on the attribute comparison method only (total certainty degree), and on both relaxation criteria have been considered.

In the first session, total certainty and total accuracy degree have been considered. In the second session, the certainty degree has been reduced by also considering RFDs relaxing on the extent only, i.e., by admitting a $g3$-error of 10%. In the third session, the accuracy degree has been reduced, by considering RFDs relaxing on the attribute comparison method only, and by setting a distance threshold equal to 1 for each attribute in the dataset. Finally, in the last session, RFDs relaxing on both criteria have been considered.

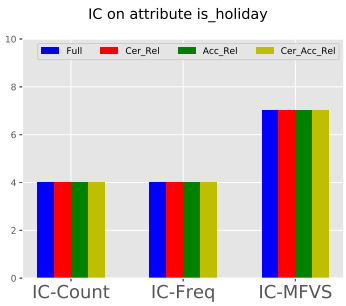
Figure 5.4 shows evaluation results for each considered dataset, grouping bars according to the used heuristics: counting (IC-Count), weighted counting (IC-Feq), and MFVS (IC-MFVS). More specifically, it shows the number of attributes to be encrypted for each used heuristic, and each of the sessions specified above. In detail, the following labels have been used: *Full* to denote no relaxation, *Cer_Rel* to denote relaxation on the extent only, *Acc_Rel* to denote relaxation on the attribute comparison method only, and



(a) CreditClient



(b) Health



(c) London

Figura 5.4: Evaluation results of the proposed methodology for Information Confidentiality.

Cer_Acc_Rel to denote relaxation on both.

In Figure 5.4(a) it is possible to notice that although the number of attributes to be encrypted for the *CreditClient* dataset is quite different across several configurations, it is quite similar across the three heuristics on the same configuration. Among the three heuristics, IC-MFVS is the best-performing one with the *Cer_Acc_Rel* configuration. On the contrary, IC-Count and IC-Freq heuristics achieve better performances than IC-MFVS with the *Cer_Rel* configuration. In Figure 5.4(b) it is possible notice that on the *Health* dataset IC-Count and IC-Freq achieve better performances than IC-MFVS in all configuration settings, except for the *Cer_Acc_Rel* configuration, where the number of attributes

to be encrypted is the same as the other two heuristics. Similar considerations apply for the *London* dataset (Figure 5.4(c)), where IC-MFVS results are worse than those of IC-Count and IC-Freq. Moreover, for this dataset, it is possible to notice that no variability is encountered across several configuration settings.

It is possible to conclude that relaxation settings can affect the number of attributes to be encrypted. As expected, the RFD relaxation usually increases the number of attributes to be encrypted, since more attribute correlations are generated, but this also yields stronger confidentiality preservation. In particular, results highlight the trade-off between the amount of encryption and the degree of confidentiality preservation that could be achieved with the proposed methodology.

General evaluation session on integrated datasets. Since the proposed methodology aims to highlight the information confidentiality risks arising during several big data processing activities, like for instance data integration, a further evaluation session has been performed, by considering datasets derived through data integration processes. In particular, the general evaluation sessions defined above have been repeated for each of the following integrated datasets:

(i) *CreditHealth*, integrating *CreditClient* and *Health*;

(ii) *CreditLondon*, integrating *CreditClient* and *London*;

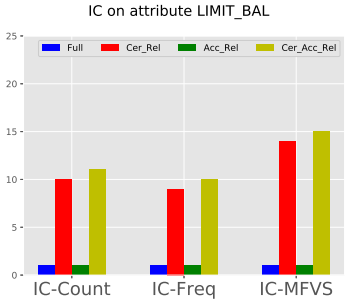
(iii) *HealthLondon*, integrating *Health* and *London*;

Table 5.5 reports the statistics on the characteristics of the integrated datasets. The data integration process has been accomplished based on the following attributes that are shared among the three datasets: `Name`, `Surname`, and `StreetAddress`.

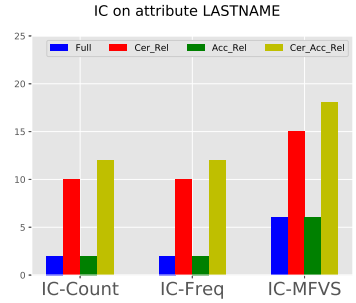
Figure 5.5 shows the obtained results. In detail, in Figure 5.5(a) it is possible to notice that for the *CreditHealth* dataset the IC-Freq heuristic performs better than the other two. In particular, this arose in both *Cer_Rel* and *Cer_Acc_Rel* settings. Figure 5.5(b) shows that the IC-Count and IC-Freq heuristics perform better than the IC-MFVS on the *CreditLondon* dataset, in

Datasets	# Columns	# Rows	# FD	Size [KB]
CreditHealth	16	30000	1518	368000
CreditLondon	20	17414	11364	193000
HealthLondon	20	17414	6816	272000

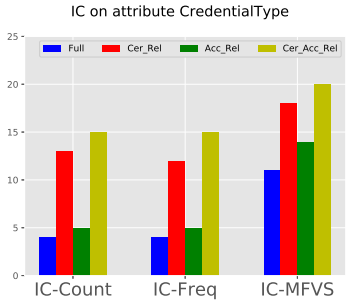
Tabella 5.5: Statistics of the integrated datasets.



(a) CreditHealth



(b) CreditLondon



(c) HealthLondon

Figura 5.5: Evaluation results of the proposed methodology for Information Confidentiality on the integrated datasets.

all configuration settings. A similar behaviour occurred for the *HealthLondon* dataset (Figure 5.5(c)).

Although it is expected that the number of attributes to be encrypted would increase when integrating datasets with respect to the single datasets, by comparing results in Figure 5.4 and Figure 5.5, it is possible to notice that this happens only when

Attributes	Length	Description
AGE	1	User's age.
PAY_0	2	User's amount paid added.
FIRSTNAME	3	User's first name added.
BILL_AMT1	4	User's bill in September 2005 added.
STREETADDRESS	5	User's street address added.

Tabella 5.6: Attributes selected for evaluating IC variability on the CreditClient dataset.

Attributes	Length	Description
BirthYear	1	User's year of birth.
CEDueDate	2	User's CE due date posted added.
FirstName	3	User's first name added.
Status	4	User's status added.
CredentialType	5	User's credential type added.

Tabella 5.7: Attributes selected for evaluating IC variability on the Health dataset.

RFD relaxation is introduced. This might be due to the fact that RFD relaxation potentially increases the possibility to catch inter-schema relationships between attributes.

IC variability evaluation session on individual datasets. As a further experiment, a IC variability evaluation has been performed, in which the number of attributes to be encrypted as the number of IC attributes grows is monitored.

Table 5.6 shows the attributes used for the IC variability evaluation on the dataset *CreditClient*, Table 5.7 attributes used for the *Health* dataset, and Table 5.8 attributes used for the *London* dataset. The IC attributes have been varied in the range $[1, 5]$, by adding an attribute concerning personal or non-personal user's data at a time. Furthermore, in order to compare results concerning all three defined heuristics, for each of them an analysis has been performed to verify how the number of attributes to be encrypted changes as the number of IC attributes increases.

Figure 5.6 shows the results achieved on the *CreditClient* dataset. In particular, the x-axis represents the number of IC attributes, whereas the y-axis represents the number of attributes to be encrypted in order to guarantee requirements on IC attributes for

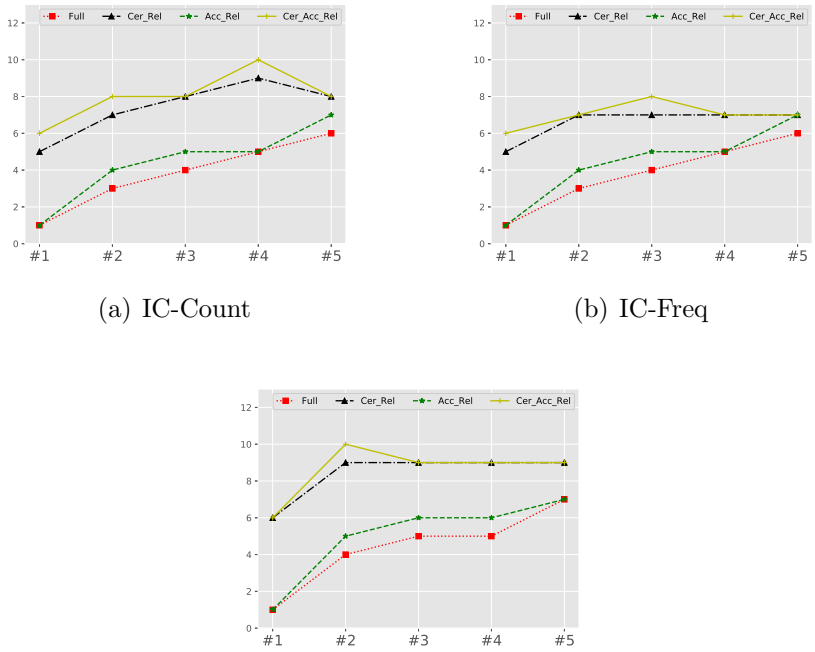
Attributes	Length	Description
is_holiday	1	User's boolean value.
is_weekend	2	User's boolean value added.
firstname	3	User's first name added.
lastname	4	User's last name added.
streetAddress	5	User's street address added.

Tabella 5.8: Attributes selected for evaluating IC variability on the London dataset.

each defined heuristic. More specifically, for the *CreditClient* dataset, all the defined heuristics show a linear growth of the number of attributes to be encrypted w.r.t the number of IC attributes, for both *Full* and *Acc_Rel* configurations, and a sub-linear growth for both *Cer_Rel* and *Cer_Acc_Rel* configurations. However, although it is possible to notice an increasing trend for all the three heuristics, sometimes the number of attributes to be encrypted decreases as the number of IC attributes increases (*Cerr_Acc_Rel* configuration).

Figure 5.7 shows the results achieved on the *Health* dataset. In particular, IC-Count in Figure 5.7(a) and IC-Freq in Figure 5.7(b) exhibit a linear growth for all the configurations. Instead, IC-MFVS in Figure 5.7(c) shows more variability in the growing trend for all configurations. In particular, for the *Cer_Rel* configuration, results exhibit a strong growth in the range $[1-2]$, and a constant trend in the range $[2-4]$. Similarly, for the *Acc_Rel* configuration, a strong growth is registered in the range $[1-3]$ and a constant trend in the range $[3-5]$.

Figure 5.8 shows results achieved on the *London* dataset. In particular, IC-Count and IC-Freq heuristics (Figure 5.8(a)-Figure 5.8(b)) follow a similar trend for each considered configuration, i.e. a linear growth for both *Full* and *Acc_Rel* configurations, and a sub-linear growth for both *Cer_Rel* and *Cer_Acc_Rel* configurations. However, the number of attributes to be encrypted is greater for IC-Count than for IC-Freq. For IC-MFVS (Figure 5.8(c)) the trends are similar to those described above for *Full*, *Cer_Rel*, and *Acc_Rel* configurations, but not for *Cer_Acc_Rel*, due to the strong growth registered in the range $[2-3]$. Moreover, it



(a) IC-Count

(b) IC-Freq

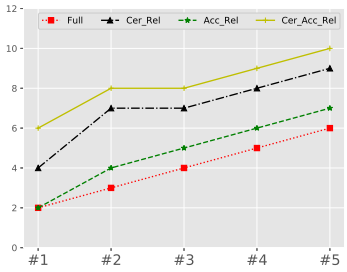
(c) IC-MFVS

Figura 5.6: Evaluation results of the proposed methodology on IC variation for the CreditClient dataset.

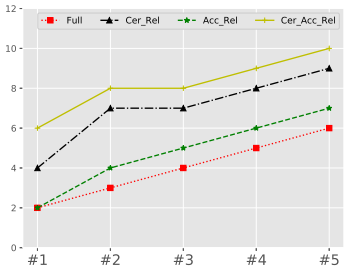
is also registered a decrease in the range $[3-4]$. Generally, results of IC-MFVS are worse in terms of the number of attributes to be encrypted w.r.t. the other two heuristics.

In general, it is not obvious that the number of attributes to be encrypted decreases when the number of confidential attributes increases. However, when a new confidential attribute is added, the process typically considers many more RFDs. Thus, the incidence of each attribute w.r.t. the selection criteria of a heuristic could change. Consequently, a heuristic could converge towards a more optimal solution, i.e. fewer attributes to be encrypted.

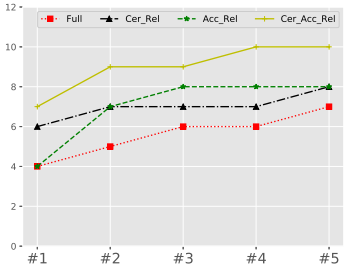
IC variability evaluation session on integrated datasets. A further IC variability evaluation session has been accompli-



(a) IC-Count



(b) IC-Freq

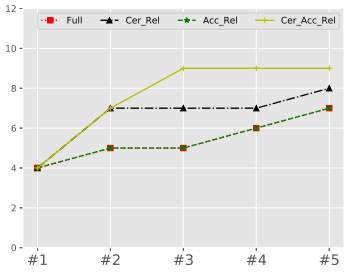


(c) IC-MFVS

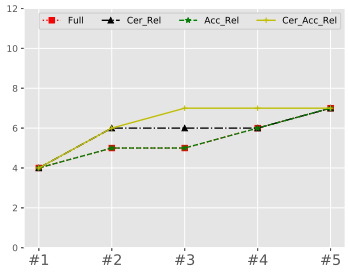
Figura 5.7: Evaluation results of the proposed methodology on IC variation for the Health dataset.

shed by considering the previously described integrated datasets (Table 5.5). In particular, also in this case the evaluation started by specifying one confidential attribute, and adding new ones up to 5. More precisely, Table 5.9 shows the attributes used for the *CreditHealth* dataset, Table 5.10 those used for the *CreditLondon* dataset, and Table 5.11 those used for the *HealthLondon* dataset.

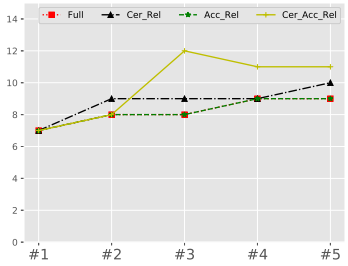
Figure 5.9 shows the results achieved on the *CreditHealth* dataset. In general, all three heuristics mainly show an increasing trend for all considered configurations. More specifically, the trend is exactly the same for *Full* and *Acc_rel* configurations with IC-Freq. Moreover, in these two configurations the number of attributes to be encrypted remains sufficiently low. Instead, a remarkable gro-



(a) IC-Count



(b) IC-Freq



(c) IC-MFVS

Figura 5.8: Evaluation results of the proposed methodology on IC variation for the London dataset.

Attributes	**Length**	**Description**
LIMIT_BAL	1	User's balance.
ExpirationDate	2	User's expiration date added.
AGE	3	User's age added.
Status	4	User's status added.
EDUCATION	5	User's education added.

Tabella 5.9: Attributes selected for evaluating IC variability on the CreditHealth dataset.

wth occurs with IC-MFVS for *Full* and *Acc_Rel* configurations in the variability range $[1-2]$.

Figure 5.10 shows the results achieved on the *CreditLondon* dataset. In particular, also for this dataset IC-MFVS exhibited

Attributes	Length	Description
LASTNAME	1	User's last name.
t1	2	User's temperature added.
AGE	3	User's age added.
is_holiday	4	User's boolean value added.
MARRIAGE	5	User's boolean added.

Tabella 5.10: Attributes selected for evaluating IC variability on the CreditLondon dataset.

Attributes	Length	Description
CredentialType	1	User's credential type.
hum	2	User's humidity added.
BirthYear	3	User's birth year added.
weather_code	4	User's weather code added.
StreetAddress	5	User's street address added.

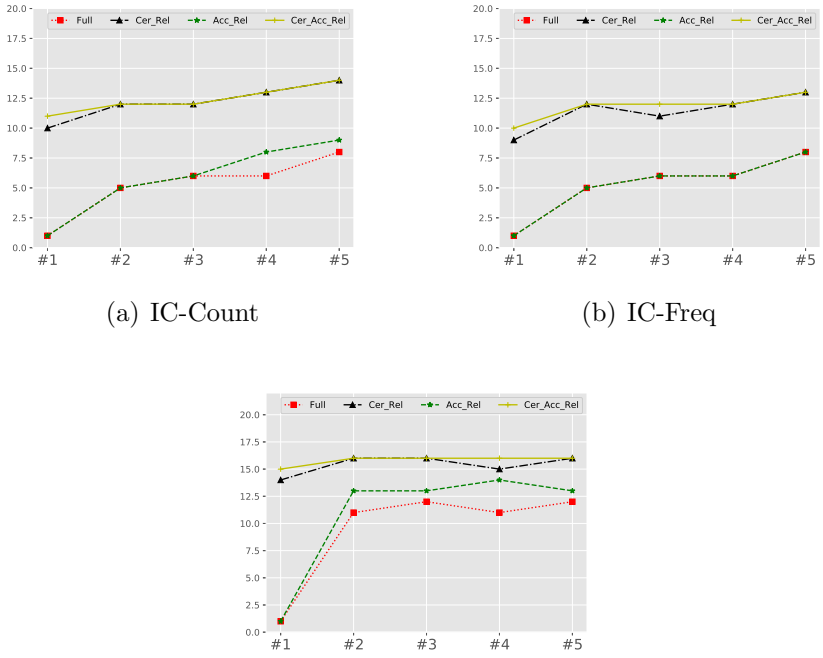
Tabella 5.11: Attributes selected for evaluating IC variability on the HealthLondon dataset.

a similar behaviour for *Full* and *Acc_Rel* configurations, and for *Cer_Rel* and *Cer_Acc_Rel* configurations. Better performances are achieved with IC-Count and IC-Freq, where the latter follows a non-monotonic trend for the *Cerr_Acc_Rel* configuration.

Figure 5.11 shows the results obtained on the *HealthLondon* dataset. In particular, IC-Count and IC-Freq heuristics (Figure 5.11(a)-Figure 5.11(b)) did not require to encrypt many attributes for *Full* and *Acc_Rel* configurations. This does not occur with IC-MFVS for the same configurations (Figure 5.11(c)). Moreover, it is possible to notice that although the *Cerr_Acc_Rel* configuration requires the maximum number of attributes to be encrypted, it follows a quasi-constant trend with all three defined heuristics.

By comparing results achieved in this evaluation w.r.t. the previous one, it is possible to notice that there are no relationships between the trends on the integrated datasets and those on the single datasets from which they are derived. Often, *Full* and *Acc_Rel* configurations required less attributes to be encrypted than *Cer_Rel* and *Cer_Acc_Rel* configurations.

Information gain evaluation session. This section describes another evaluation session that aim to analyze the effectiveness



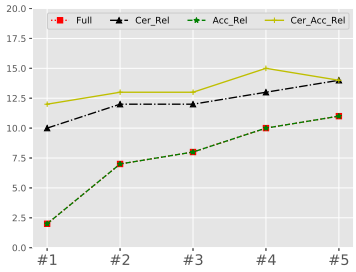
(a) IC-Count

(b) IC-Freq

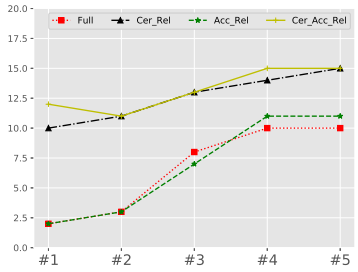
(c) IC-MFVS

Figura 5.9: Evaluation results of the proposed methodology on IC variation for the CreditHealth dataset.

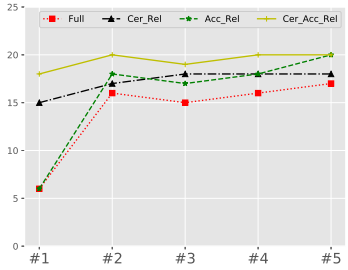
of the proposed methodology in preserving the quality of data after the privatization process. In particular, it has been considered a classification scenario in which it is important to guarantee the quality of data even if the privacy preservation must be ensured. Thus, the data quality in terms of information gain has been measured on the unencrypted dataset, and compared it to the partially encrypted dataset obtained by applying of the proposed methodology. More specifically, the aim is to understand the dispersion of the data in terms of information gain [98], which exploits the concept of entropy. The latter is defined in equation (5.5), and characterizes the purity of an arbitrary collection of examples.



(a) IC-Count



(b) IC-Freq



(c) IC-MFVS

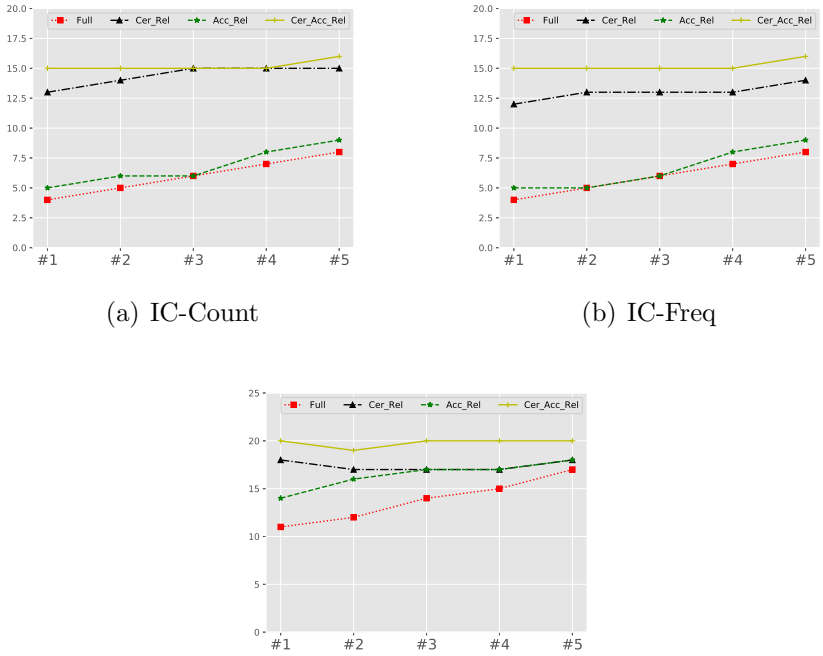
Figura 5.10: Evaluation results of the proposed methodology on IC variation for the CreditLondon dataset.

$$\text{Entropy} \;=\; H(X) = -\sum p(X)\log p(X) \tag{5.5}$$

where

- X is an attribute of the dataset;
- $H(X)$ is the entropy of X;
- $p(X)$ is the probability of getting a value of X when randomly selecting one from the set.

Instead, the Information Gain is the expected reduction in the entropy caused by partitioning the examples according to a gi-



(a) IC-Count

(b) IC-Freq

(c) IC-MFVS

Figura 5.11: Evaluation results of the proposed methodology on IC variation for the HealthLondon dataset.

ven attribute. The formal definition of the information gain is expressed in (5.6).

$$\text{Information Gain} = \ I(X,Y) = H(X) - H(X|Y) \tag{5.6}$$

where

- X and Y are attributes of the dataset;
- $I(X,Y)$ is the information gain on the attribute Y;
- $H(X)$ is the entropy on X;
- $H(X|Y)$ is the entropy of X given Y.

The analysis evaluated the variation of information gain for each attribute in the Health dataset, and used `Status` as the target attribute of the classification scenario. In other words, the information gain of each attribute has been evaluated w.r.t. the `Status` attribute. Moreover, the information gain has been computed by considering every encrypted value belonging to the same class, like in the case of null values.

Figure 5.12 shows the obtained results, where the blue bar is related to the information gain computed on the attributes without encryption (e.g. exposed to privacy threats), and the red one is related to the information gain computed after the application of the proposed methodology, i.e. the partially encrypted dataset. According to Figure 5.12, it is possible to notice that the variation of information gain is almost always small. This highlights the fact that the proposed methodology is a useful means to guarantee privacy preservation without heavily affecting the quality of data. More specifically, some exceptions have been encountered. A slightly worse behavior is obtained for *IG-Attr1*, i.e. `LastName`, due to the many encryptions on an attribute whose distribution contains many values. Instead, for *IG-Attr3* the information gain remains unchanged.

This evaluation represents a specific analysis scenario, which permitted to verify how in a real-world scenario it is possible to work with partially encrypted data, aiming to ensure both privacy preservation and data usage.

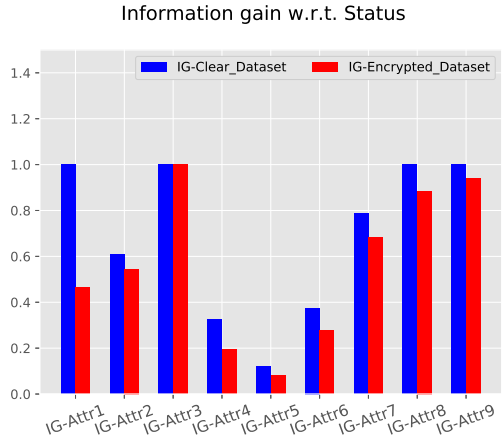


Figura 5.12: Evaluation results of the information gain on the Health dataset.

Capitolo 6

A Methodology for Privacy Preserving Machine Learning

Machine learning (ML) is being increasingly exploited in various application domains, yielding to the proliferation of systems ranging from intrusion detection to recommendation systems.

The input data of an ML algorithm are usually represented as a set of samples. Each sample contains a set of feature values. An ML algorithm uses a training set formed of multiple feature vectors and their associated labels. The analysis of such data by a ML algorithm is called the training or learning phase, thorugh which a predictive model is derived. Thus, when a new test sample is tested, this model should predict the label (person's name or identification data in face-recognition applications). The ability of such a model to accurately predict the label is a measure of how well it generalizes in order to predict unseen data. This is empirically measured through the test error (generalization error), and it can depend on the quality and quantity of the data used for training the model, the setting of hyperparameters (e.g., using cross-validation), and even the feature extraction method used (if any required).

Supervised learning utilizes labelled data, where each feature

vector is associated with an output value, which might be a class label (classification) or a continuous value (regression). For example, with classification, the samples (feature vectors) belong to two or more classes, and the objective of a ML algorithm is to determine the class to which the new sample belongs. Instead, in unsupervised learning, the target is to find the underlying structure or distribution of the data in order to learn more about them.

Although it is widely recognized that the data and their quality can affect ML model performances, when ML applications require to base their learning phase over private individuals' data, the latter are uploaded to centralized locations in clear text, enabling ML algorithms to extract patterns and build models from them. In this context, the privacy-preserving problem is not limited to the threats associated to the possible exposure of such data to hacking attacks, but it is possible to glean extra information about the private datasets by analysing the results provided by the application of ML models, even though the data have previously been anonymized [99]. In fact, the application of machine learning techniques to large and distributed data archives might yield the disclosure of sensitive and confidential data, due to the excessive processing capabilities reached by these techniques. With this in mind, in the next section, a novel solution to anonymize data is presented, which guarantees the possibility to use ML techniques for classification tasks. In particular, it exploits RFDs to directly evaluate combinations of attributes together with possible generalizations over the data, and it defines suitable generalization configurations for data anonymization. This also permits to overcome the main limitation of related proposals (see Section 2.3), which typically perform generalization steps over a single attribute, hence neglecting possible correlations in the data. Moreover, the proposed approach also leverages the Pareto principles to filter such configurations by means of both privacy and quality measures (e.g., information gain, accuracy and privacy).

6.1 Data Anonymization

The exponential growth of data volumes and the need to analyze them using various techniques, including machine learning, have led the research community to address anonymization issues in data sharing. However, data can univocally refer to persons whose anonymity must be guaranteed when information is shared for different reasons, such as for data analytics activities. In the following, a methodology for identifying data anonymization strategies is presented; it exploits data correlations expressed in terms of relaxed functional dependencies (RFDs) automatically extracted from data. To this end, generalization rules to anonymize data are identified and validated by using the well-known k-anonymity model. Furthermore, a decision tree classifier has been used in order to compute several data utility measures, such as accuracy and information gain over anonymized data. Experimental results over real-world datasets show that the proposed approach achieves promising results in data utility, while maintaining a high anonymization level for data sharing activities.

6.1.1 Problem statement

Classification models capture correlations between the attributes of individuals and a class value, and are often used to predict the class value for any unseen new observation. Classification models are built from a training dataset, which might contain sensitive information. This information could be inferred from the classification model, by exploiting the correlations encoded in it [100]. To this end, training data are usually anonymized by removing identifiable information before the classifier is trained. However, data can be still re-identified by using quasi-identifiers [101]. Quasi-identifiers are sets of attributes not identifiable when considered singularly, but their combination could yield a unique identifier. For instance, it has been shown that the combination of zip code, gender, and date of birth permits to uniquely identify around 87 percent of the US population [102].

	age	workclass	fnlwgt	education	maritial-status	occupation	relationship	sex	capital-gain	classes
t_1	39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	Male	2174	>50K
t_2	50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	0	>50K
t_3	38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	Male	0	<=50K
t_4	53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	<=50K
t_5	37	Private	159449	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Female	0	>50K
t_6	37	Private	284582	Masters	Married-civ-spouse	Exec-managerial	Wife	Female	0	<=50K
t_7	49	Private	160187	9th	Married-spouse-absent	Other-service	Not-in-family	Female	0	>50K
t_8	52	Self-emp-not-inc	209642	HS-grad	Married-civ-spouse	Exec-managerial	Husband	Male	0	<=50K
t_9	38	Private	45781	Masters	Never-married	Prof-specialty	Not-in-family	Female	14084	>50K
t_{10}	49	Private	159449	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	5178	>50K

Tabella 6.1: Example dataset containing users' information.

Example 9. *Considering the dataset in Table 6.1, extracted from the Adult dataset*[1]*, in which each tuple describes an individual, where* `age`*,* `workclass`*,* `fnlwgt`*,* `education`*,* `maritial-status`*,* `occupation`*,* `relationship`*,* `sex`*, and* `capital gain` *are attributes characterizing the individual, and the attribute* `classes` *indicates whether his/her annual income is greater than* $50K$ *or not. From this dataset is possible to narrow down tuple* t_1 *to a specific individual by looking, for instance, at the* `age` *attribute, as this is the only tuple for which* `age` *is equal to* 39*.*

This simple example shows that only removing identifiable information from a dataset is not sufficient to guarantee anonymization. Anonymized data can be re-identified by linking the data by means of other data sources [103]. Therefore, before disclosing a dataset containing highly sensitive information, data owners must often transform it to reduce the risk that its records can be re-identified. An anonymization model largely used for this is k-anonymity, which requires that at least k individuals in the dataset share the same set of attribute values (cf. Section 3.1.2.1 for details).

A common way to achieve k-anonymity is through generalization [10]. Intuitively, generalization is used to replace the values in a dataset with more general ones. For example, numerical data can be replaced by intervals, whereas categorical attributes can be generalized into a set of distinct values. Therefore, the application of generalization strategies aims at grouping different tuples, in order to make them indistinguishable, especially for quasi-

[1] `https://www.openml.org/d/179`

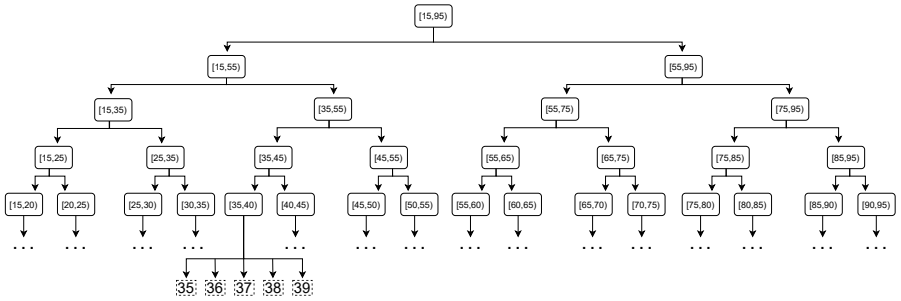


Figura 6.1: Taxonomy of the `age` attribute for the dataset in Table 6.1.

identifier attributes, thus contributing to achieve the desired level of k-anonymity.

The values of an attribute can be generalized at a different granularity, providing different levels of generalization for the attribute, yielding different levels of k-anonymity. Generalization levels can be organized in a hierarchical structure (hereafter called *attribute taxonomy*), which can be used to regulate the level of generalization to be applied for an attribute. In this work, every attribute in the dataset is assumed to be associated with an attribute taxonomy representing all generalization levels defined for the attribute.

Example 10. *Figure 6.1 shows the taxonomy of the* `age` *attribute for the example dataset in Table 6.1. As shown in the figure, the leaf nodes (level* 0*) represent the values in Table 6.1, which can be generalized at different levels. For instance, value* 39 *can be replaced with interval* $[35, 40)$ *at level* 1*, with interval* $[35, 45)$ *at level* 2*, and so on. Based on the taxonomy for the attribute* `age` *in Figure 6.1, it is possible to observe that by applying generalization at level* 1 *for the* `age` *attribute on (a projection of) the dataset in Table 6.1, k-anonymity with* $k = 2$ *is obtained (cf. Table 6.1(a)), whereas k-anonymity with* $k = 5$ *is reached by using the generalization at level* 2 *(cf. Table 6.1(b)).*

This example shows that by increasing the generalization level of an attribute, it is possible to achieve a higher anonymity

(a) Level 1

	age
t_1	[35,40)
t_2	[50,55)
t_3	[35,40)
t_4	[50,55)
t_5	[35,40)
t_6	[35,40)
t_7	[45,50)
t_8	[50,55)
t_9	[35,40)
t_10	[45,50)

(b) Level 2

	age
t_1	[35,45)
t_2	[45,55)
t_3	[35,45)
t_4	[45,55)
t_5	[35,45)
t_6	[35,45)
t_7	[45,55)
t_8	[45,55)
t_9	[35,45)
t_10	[45,55)

Tabella 6.2: generalization of (a projection of) the dataset in Table 6.1 over attribute `age` by considering two generalization levels defined in Figure 6.1.

level (represented by the value of k). Nonetheless, the application of generalization can have a negative impact on data utility. For example, generalization can decrease the performance of a classifier when trained on a generalized dataset, as generalization might weaken the correlations in the data [104, 105, 106]. Finding suitable generalization strategies that preserve anonymity while not affecting (too much) data utility is not trivial and it requires finding a trade-off between anonymity and data utility. This trade-off boils down to determine suitable levels of generalization that guarantee data anonymization while maintaining as much data utility as possible.

In this work, a novel anonymization technique has been proposed. It uses generalization and k-anonymity validation to anonymize a dataset while minimizing the loss of data utility. To this end, data correlations in the dataset, expressed in terms of relaxed functional dependencies (RFDs), have been used as a guideline to define suitable generalization strategies.

Starting from a dataset and the attribute taxonomies as input, the methodology presented in the next section shows how to extract RFDs suggesting generalization levels that ensure a given level of data ananymization, while maintaining as much data

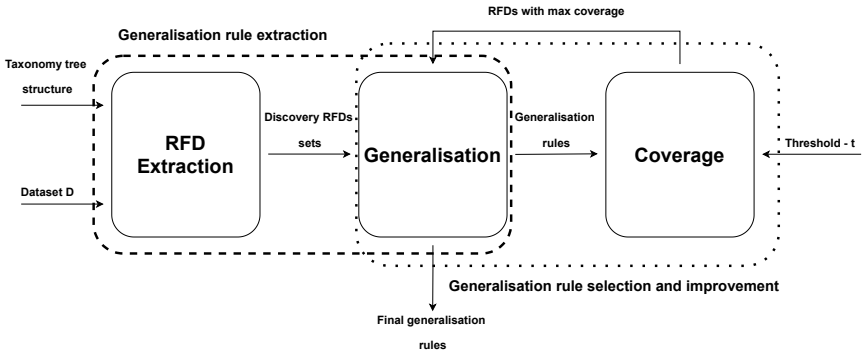


Figura 6.2: Overview of the proposed methodology.

utility as possible.

6.1.2 The Proposed Methodology

This section presents the methodology and shows how data correlations, expressed in terms of relaxed functional dependencies (RFDs), can be used to define strategies for guaranteeing data anonymization in classification activities. Intuitively, RFDs can be used as guidelines to determine which subsets of attributes, together with their generalization levels, are suitable for classification tasks while guaranteeing, at the same time, a given level of data anonymization.

6.1.2.1 Overview

Figure 6.2 shows an overview of the proposed methodology. Given an input dataset and the taxonomy for the attributes occurring in the dataset, generalization rules expressed in terms of RFDs (*RFD Extraction*) can be extracted by means of the first step.

These generalization rules are used to determine which attributes should be generalized and at which level. To assess the quality of generalization rules in terms of data anonymization and data utility, the attribute values in the input dataset are replaced according to the generalization rules, whereas the k-anonymity level,

the classification accuracy, and the information gain are computed on the generalized dataset (*Generalization*). In a second step, the coverage of the RFDs that satisfy a given level of anonymization are extended by joining generalization rules to increase data utility (*Coverage*). The data anonymization and utility provided by the obtained extended RFDs are then assessed as in the previous step (*Generalization*). The obtained generalization rules provide data owners with a view of which generalization rules can be used to anonymize their datasets, together with their effects in terms of data utility and anonymization. The next section presents the detailed steps of the proposed methodology.

6.1.2.2 Generalization rule extraction

The first step of the proposed methodology (represented by the two blocks within the dashed line in Figure 6.2) aims to extract generalization rules in terms of RFDs, along with measures assessing their data anonymization and utility.

RFDs are extracted along with the generalization levels (defined with respect to the given attribute taxonomies) from the input dataset by using roll-up dependencies.

Definition 9. (Roll-up dependency) *Let G be a genschema of a relation schema R, $X, Y \subset attr(R)$, a roll-up dependency RUD) $X_{\Phi_1} \rightarrow Y_{\Phi_2}$ holds on an instance r of R, if and only if for each tuple pair (t_1, t_2) of r, if t_1 and t_2 are α-equivalent on the attributes in X, then they must also be α-equivalent on the attributes in Y.*

In particular, this type of dependencies permits to retrieve not only attribute correlations, but also the generalization level of the attributes, according to a given attribute taxonomy.

During RFD extraction, only the RFDs involving the classification attribute on the Right-Hand-Side (RHS) with a generalization level equals to 0 have been considered (i.e., attribute `classes` in the example dataset of Table 6.1). This is because this work focuses on the generation of anonymized datasets that can be used

to train a classification model. Accordingly, the focus is on correlations involving the classification attribute and preserving its original values.

Example 11. *The classification attribute* `classes` *of the dataset in Table 6.1 can take two values, namely ">50K" and "<=50K". If this attribute is generalized to a single value, for example,* [Any classes]*, all tuples in the dataset will have the same value for it, making the dataset ill-suited to train a classification model.*

The obtained RFDs identify which attributes along with their generalization level can allow performing classification activities based on the data correlations within the dataset. Accordingly, each RFD can be used to produce an anonymized version of the dataset, in which only the attributes involved in the RFD are selected and generalized at the level specified by the RFD. This is done by replacing the value in the original dataset according to the levels specified on the Left-Hand-Side (LHS) attributes of the RFD, and the corresponding generalization levels defined in the attribute taxonomy. All attributes that do not occur in the RFD are mapped to the highest level defined in the attribute taxonomy, since they are not involved in the correlation defined by the RFD.

Example 12. *Suppose the following RFD is extracted from the dataset of Table 6.1:*

$$\texttt{age}_{\leq 3},\ \texttt{fnlwgt}_{\leq 2} \rightarrow \texttt{classes}_{\leq 0}$$

Its RHS contains the classification attribute `classes`*, whereas its LHS contains the subset of attributes* `age` *and* `fnlwgt` *to be generalized. The generalization level is defined by the values after the tag "≤", representing the required generalization levels.*

Table 6.3 shows the dataset resulting from the application of the RFD to the dataset in Table 6.1. It can be observed that the values of attributes `age` *and* `fnlwgt` *have been generalized by replacing their original values with those defined by the generalization levels indicated by the RFD (as an example, the taxonomy for attribute* `age` *is reported in Figure 6.1). The values of other attributes are*

	age	fnlwgt	Classes
t_1	[35,55)	[0,100000)	>50K
t_2	[35,55)	[0,100000)	>50K
t_3	[35,55)	[200000,300000)	<= 50K
t_4	[35,55)	[200000,300000)	<= 50K
t_5	[35,55)	[100000,200000)	>50K
t_6	[35,55)	[200000,300000)	<= 50K
t_7	[35,55)	[100000,200000)	>50K
t_8	[35,55)	[200000,300000)	<= 50K
t_9	[35,55)	[0,100000)	>50K
t_{10}	[35,55)	[100000,200000)	>50K

Tabella 6.3: A sample application scenario of a single RFD.

generalized to the highest level. For the sake of clarity, they are not shown in Table 6.3.

The extracted RFDs can provide different levels of data anonymization and data utility. Therefore, the level of data anonymization and data utility offered by each RFD should be assessed to determine which RFD(s) should be used for the generation of the generalized dataset. Then, the anonymization level of the generalization strategy driven by an RFD is measured through the k-anonymity model proposed in [55]. Accordingly, the anonymization level represents the minimum number of tuples in the dataset (obtained by applying the generalization) with identical quasi-identifiers, i.e., the value of k in the k-anonymity model. From Table 6.3 it is possible to observe that the application of the generalization strategy driven by the RFD presented in Example 12 achieves k-anonymity level with $k = 3$.

On the other hand, the data utility of the generalized dataset has been measured by computing the accuracy and the information gain that a Decision Tree classifier can achieve on the dataset generalized by using the RFD. In particular, in order to compute the accuracy and the information gain for all generalizations derived by the RFDs, a Decision Tree classifier has been used, i.e., the ID3 algorithm [107], since it is one of the most widely used machine learning models, due to the fact that it works well with noisy or missing data.

In summary, this step of the methodology returns a list of RFDs, which permits to define different generalization rules, along with their anonymization level (measured in terms of k-anonymity) and data utility (measured in terms of accuracy and information gain).

Example 13. *In what follows, several RFDs extracted from the dataset in Table 6.1 are presented, together with the corresponding data anonymization and data utility measures, computed for each generalization rule derived from them:*

r_1: $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 65; IG : 0.011657;$

r_2: $[\texttt{age}_{\leq 3}, \texttt{gender}_{\leq 1} \rightarrow \texttt{Classes}_{\leq 0}]; k : 4; A : 66; IG : 0.043581;$

r_3: $[\texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 67; IG : 0.072174;$

r_4: $[\texttt{workclass}_{\leq 2}, \texttt{age}_{\leq 4}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 5; A : 61; IG : 0.007948;$

r_5: $[\texttt{relationship}_{\leq 1}, \texttt{education}_{\leq 2}, \texttt{capital-gain}_{\leq 3} \rightarrow \texttt{Classes}_{\leq 0}]; k : 2; A : 68; IG : 0.079399;$

where, k, A, and IG represent the anonymization level, the accuracy, and the information gain, respectively. Accordingly, it is possible to observe that r_4 achieves the best anonymization level ($k = 5$), but the lowest accuracy ($A = 61$). On the other hand, r_5 achieves the best accuracy ($A = 68$), but the lowest anonymization level ($k = 2$).

As shown in the previous example, data owners are left with the task to determine which generalization rules should be used for the anonymization of their datasets. This can be a complex task, since a large number of RFDs can be extracted from a dataset [70, 69], and not all of them might satisfy the desired level of anonymization. In addition, RFDs usually capture basic correlations in the data, which involve a limited number of attributes, limiting the data utility that can be achieved from their application. By

extending the coverage of an RFD to capturing multiple correlations will make it possible to consider more attributes for the anonymization of the dataset, and hence increase its data utility [105]. However, the use of more attributes could reduce the level of anonymization guaranteed by the generalization rules. Therefore, data utility can be improved only where, and to the extent that, the minimum level of anonymization required by the data owner is satisfied.

The next section presents the proposed approach to identify those generalization rules that satisfy a given level of anonymization while maximizing data utility. To this end, it is necessary to devise an RFD join strategy to increase the coverage of RFDs, in an attempt to increase the data utility provided by the baseline generalization rules obtained from the RFDs.

6.1.3 Generalization rule selection and improvement

This step of the methodology (represented by the two blocks within the dotted line in Figure 6.2) aims to generate a set of candidate generalization rules that satisfy a minimum level of anonymization, while increasing the data utility from the RFDs obtained in the previous step.

Some RFDs identified in the previous step might not guarantee a level of anonymization that is acceptable for the data owner. In particular, the data owner might define minimum anonymization requirements for a dataset to be shared with other parties. According to the k-anonymity model, these requirements have been represented as a threshold t, indicating the minimum value of k that a generalization rule should satisfy. Thus, from the RFDs obtained in the previous step, those for which the corresponding generalized dataset does not meet the requirement for k-anonymity, i.e. $k < t$, have been filtered out.

The RFDs obtained in the previous step only capture basic correlations in the data, hence limiting the data utility that can be achieved through their application. To this end, the attribu-

tes involved in all the RFDs have been analyzed, and a coverage strategy has been defined to increase the number of attributes to be used for the anonymization of the dataset. The strategy compares the RFDs and determines which ones can be combined to improve data utility. The intuition is that joining dependencies allows to account for multiple data correlations simultaneously, hence increasing the number of attributes that can be used. Since combined RFDs need to hold on the considered dataset, not all RFDs can be combined with each other.

Before presenting the procedure for generating the candidate generalization rules, it is necessary to introduce the notion of *compatible RFDs*, which specifies when two RFDs can be joined. Intuitively, two RFDs are compatible if and only if their LHS attributes are disjoint or occur with the same generalization level. Formally:

Definition 10 (RFD Compatibility)**.** *Let $X_\Phi \rightarrow C_{\leq 0}$ be an RFD, $X = \{A_1, \ldots, A_n\}$, $X' = \{B_1, \ldots, B_m\}$, and each attribute A_i (B_j) be associated with a generalization level ϕ_i (ϕ'_j) in Φ (Φ'). The two RFDs are said to be compatible if and only if:*

- *$X \cap X' = \emptyset$, or*
- *$\forall A_i \in X$ and $B_j \in X'$, such that $A_i = B_j \in X \cap X'$, then $\phi_i = \phi'_j$.*

Algorithm 2 presents the procedure used to generate the candidate generalization rules. The algorithm takes as input the list of RFDs Z obtained in the previous step, the dataset D with the corresponding attribute taxonomies T, and a threshold t representing the minimum level of anonymization to be satisfied, and it returns a list of candidate generalization rules R that satisfy the required level of anonymization. The algorithm starts by filtering out those RFDs in Z that do not satisfy the minimum level of anonymization t, by using the function `FILTER_BY_T` (line 1). The algorithm is iterative and it uses a list W to keep track of which RFDs should be considered at each iteration to create new RFDs. List W is initialized to the set Z', which comprises the

Algorithm 2 Join procedure

INPUT: Dataset D, taxonomy T, list of RFDs Z, threshold t
OUTPUT: List of generalization rules R

1: $Z' \leftarrow$`FILTER_BY_T(`Z`,`t`)`
2: $W := Z'$
3: $R := Z'$
4: **while** $W \neq \emptyset$ **do**
5: $L := \emptyset$
6: **for each** $(x_i, y_i \in W)$ **do**
7: Let $x_i = X_\Phi \rightarrow C_{\leq 0}$
8: Let $y_i = Y_{\Phi'} \rightarrow C_{\leq 0}$
9: **if** $(X \cap Y = \emptyset) \vee (\forall a \in X \cap Y \text{ level}(Y[a]) = \text{level}(X[a]))$ **then**
10: $c_i = X_\Phi, Y_{\Phi'} \rightarrow C_{\leq 0}$
11: $d_i \leftarrow$ `COMPUTE_generalization(`c_i`,`D`,`P`)`
12: $L \leftarrow L \cup$ `FILTER_BY_T(`$\{d_i\}$`,`t`)`
13: **end if**
14: **end for**
15: $W \leftarrow L$
16: $R \leftarrow R \cup W$
17: **end while**
18: **return** R

RFDs in Z having anonymization level at least t (line 2). The RFDs in W are considered pairwise (lines 6-14). If two RFDs x_i and y_i are compatible (cf. Definition 10), a new RFD c_i is created by joining x_i and y_i (lines 9-10). The anonymization level and the data utility of the new RFD are then computed by using the function `COMPUTE_GENERALIZATION`, which generalizes the dataset D by using the RFD, and it assesses its level of k-anonymity, accuracy, and information gain based on the generalized dataset, as described in Section 6.1.2.2 (line 11). The function `FILTER_BY_T` is then used to determine whether the RFD satisfies the minimum level of anonymization t, in which case the RFD is added to L (line 12). After that, all rules in W have been considered, L contains the generalization rules obtained by combining the RFDs in W and that satisfy the minimum level of anonymization. These rules are used in the next iteration. The algorithm terminates when no ge-

neralization rules satisfying the minimum level of anonymization can be created, returning the new set of RFDs (which contains at least the RFDs in Z') together with their anonymization and data utility level. It is worth noting that the RFDs considered at the beginning of an iteration are exactly the ones resulting from the previous iteration (line 15). In this way, no candidate RFDs are missed.

It is easy to observe that: *(i)* a set of RFDs can be combined into a new RFD if and only if every RFD is compatible with the others; and *(ii)* if the combination of two RFDs does not meet the minimum anonymization level, any combination of RFDs that includes those RFDs will also not satisfy the minimum anonymization level property, hence it will be discarded.

Example 14. *Consider the RFDs presented in Example 13 and a minimum anonymization level* $t = 3$*. Algorithm 2 filters out the RFDs that do not meet the minimum anonymization level, hence it discards* r_4*. Then, the remaining RFDs are analyzed pairwise, and compatible ones are combined, obtaining:*

r_6*:* $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{gender}_{\leq 1} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 67;$ $IG : 0.074251;$

r_7*:* $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3},$ $\texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 70; IG : 0.098579;$

r_8*:* $[\texttt{age}_{\leq 3}, \texttt{gender}_{\leq 1}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3},$ $\texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 71; IG : 0.099719;$

r_9*:* $[\texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2},$ $\texttt{age}_{\leq 4} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 69; IG : 0.096718$

It can be observed that r_6 *is obtained by joining rules* r_1 *and* r_2*,* r_7 *by joining rules* r_1 *and* r_3*,* r_8 *by joining rules* r_2 *and* r_3*, and finally,* r_9 *by joining rules* r_3 *and* r_4*. Notice that* r_4 *is not combined with* r_1 *and* r_2 *because they are incompatible: attribute* `age` *occurs at generalization level* 4 *in* r_4 *and at generalization level* 3 *in* r_1 *and* r_2*. Also, all rules satisfy the minimum anonymization*

level $k = 3$. Thus, the set of generalization rules resulting from this process are used in the second iteration. In particular, by combining rules r_6 and r_7 (but also r_6 and r_8, or r_7 and r_8) the following rule is obtained:

r_{10}: $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{gender}_{\leq 1}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 4}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 72; IG : 0.109829;$

On the other hand, rule r_9 cannot be merged with any other rule, due to its incompatibility on attribute `age`*. As no new RFD can be created, the procedure returns the set of candidate generalization rules $\{r_1, r_2, r_3, r_4, r_6, r_7, r_8, r_9, r_{10}\}$, which represents the complete set of generalization rules meeting the minimum anonymization requirement to be satisfied.*

Algorithm 2 returns a list of candidate generalization rules satisfying the given minimum level of anonymization. These rules provide a different anonymization and data utility levels, allowing the data owner to control the trade-off between these dimensions. However, the large number of rules that can be potentially returned might hamper the selection of the generalization rule to be used. Identifying the optimal candidate rules can be seen as a multi-objective optimization problem. Thus, the notion of Pareto-optimality and Pareto frontier [108] have been used in order to support the selection of optimal generalization rules.

In Pareto-optimality, the objective function comprises multiple criteria, and the multi-objective optimization problem can be formulated as follows:

$$\max F(X), \quad F = f_1(X), f_2(X), \ldots, f_m(X) \tag{6.1}$$

A solution X is said to dominate a solution Y if $\forall i = \{1, 2, \ldots, m\}, f_i(X) \geq f_i(Y)$, and there exists $j \in \{1, 2, \ldots, m\}$ such that $f_j(X) > f_j(Y)$. Solution X is called Pareto optimal if it is not dominated by any other solution. More than one Pareto-optimal solution exists when no solution is optimal with respect to every criterion. The curve or surface composed of the Pareto-optimal solutions is known as the Pareto frontier [109].

Rule	Privacy	Accuracy	Information Gain
r_1	3	65	0.011657
r_2	4	66	0.043581
r_3	3	67	0.072174
r_4	5	61	0.007948
r_6	3	67	0.074251
r_7	3	70	0.098579
r_8	3	71	0.099719
r_9	3	69	0.096718
r_{10}	3	72	0.109829

Tabella 6.4: Generalization rules of the Example 14 together with anonymization and data utility levels.

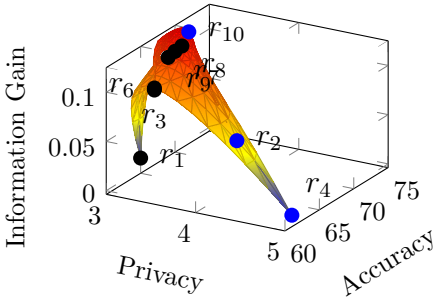


Figura 6.3: Visual representation of the Pareto Frontier for the generalization rules in Table 6.4.

The Pareto frontier is used to identify which generalization rules extracted from Algorithm 2 are (Pareto) optimal with respect to anonymization and data utility. In this light, the objective functions are represented by the k-anonymity level, accuracy, and information gain, whereas the goal is to find the solutions for which there is no other solution improving one criterion without reducing any other criterion. Thus, the generalization rules on the Pareto frontier represent the rules that provide the data analyst with the best trade-off between anonymization and data utility requirements.

Example 15. *Consider the generalization rules returned in Example 14, which are summarized in Table 6.4. The generalization rules are highlighted in gray from the Pareto Frontier. A visual representation of the Pareto frontier is shown in Figure 6.3, where the x-axis represents the accuracy level A, the y-axis represents the anonymization level k, and the z-axis represents the information gain IG. The blue points represent the Pareto Frontiers, i.e., the generalization rules that are not dominated by any other rule (r_2, r_4 and r_{10}).*

6.1.4 Experimental Evaluation

The methodology proposed in Section 6.1.2 has been evaluated to understand the trade-off between anonymization and data utility that can be achieved by using generalization rules, and to devise strategies to select the generalization rules to be used for data anonymization. More specifically, the goal here is to answer the following research questions:

RQ1: What is the impact of combining generalization rules on data utility?

RQ2: Which trade-off between anonymization and data utility can be achieved using generalization rules?

RQ3: How much effort is required by a data owner to identify the generalization rule to apply?

The first research question (**RQ1**) aims to provide insights into the impact of combined generalization rules on the data utility. In fact, an assumption underlying this work is that by combining generalization rules allows achieving a higher information gain, as it allows exploiting multiple data correlations simultaneously (cf. Section 6.1.3). **RQ2** aims to assess the trade-off between anonymization and data utility that can be achieved by using generalization rules. In particular, it is necessary to understand how the enforcement of a given anonymization level impacts data utility. A large number of generalization rules could potentially satisfy both anonymization and data utility requirements. This could affect the data owner, who has to decide which generalization rule to apply on his/her dataset. To this end, **RQ3** aims to evaluate the effort required to a data owner to determine the generalization rule to apply for the anonymization of his/her dataset, in terms of the number of rules returned by the proposed methodology. The remainder of this section presents the settings and the results of the experiments.

Datasets	#Rows	#Attributes	Attribute types
Electricity	45312	8	Numeric
Adult	48842	14	Nominal, Numeric

Tabella 6.5: Statistics on the datasets used in the evaluation.

6.1.4.1 Experiment settings

Datasets. To evaluate the proposed approach, two real-world datasets have been considered. An overview of the selected datasets is are reported in Table 6.5.

Electricity:[2] This dataset comprises records from the Australian New South Wales Electricity Market from May 1996 to December 1998. Each record refers to a period of 30 minutes, and is characterized by 8 numerical attributes, including the day of the week, the timestamp, the South Wales electricity demand, and the Victoria electricity demand. The class label identifies the price change (UP or DOWN) in New South Wales relative to a moving average of the last 24 hours.

Adult:[3] This dataset describes 48842 individuals using a mix of numeric and categorical attributes (14 attributes in total), such as age, occupation, and education. The class attribute represents individuals income, which has two possible values: '$> 50K$' and '$< 50K$'.

Attribute Taxonomies. The methodology requires the attribute taxonomies for the attributes in the given datasets to enable data generalization. The taxonomy for numerical attributes has been computed by using a bottom-up approach, whereas for categorical attributes a top-down approach based on k-Means clustering is used [110]. More specifically, the generalization levels for numeric attributes were created by ordering the attribute values (i.e., the leaf nodes) in descending order and by grouping them in sets of size five. Then, at each level, pairs of contiguous sets

[2] `https://datahub.io/machine-learning/electricity`

[3] `https://archive.ics.uci.edu/ml/datasets/Adult?ref=datanews.io`

Tabella 6.6: Overview of the attribute taxonomies for the considered datasets.

(a) Electricity

Attribute	Type	Domain size	#Taxonomy levels
date	numeric	934	5
day	numeric	8	3
period	numeric	47	4
nswprice	numeric	4088	7
nswdemand	numeric	5275	6
vicprice	numeric	6203	10
vicdemand	numeric	2845	6
transfer	numeric	1877	10

(b) Adult

Attribute	Type	Domain size	#Taxonomy levels
age	numeric	73	6
workclass	nominal	7	2
fnlwgt	numeric	26740	6
education	nominal	16	3
education-num	numeric	16	4
marital-status	nominal	7	4
occupation	nominal	14	3
relationship	nominal	6	2
race	nominal	5	3
sex	nominal	2	2
capitalgain	numeric	120	6
capitalloss	numeric	96	4
hoursperweek	numeric	95	4
native-country	nominal	40	4

were grouped to create a new level, until a single set representing the taxonomy's root was created. On the other hand, k-Means was applied on categorical attributes to ensure that similar tuples were grouped together to minimize accuracy loss. In particular, k-Means was used to partition the set of all attribute values (the taxonomy's root) into two clusters, and then applied recursively to each cluster until no further split was obtained. The last level of the taxonomy (leaf nodes) was generated by creating a node for each attribute value, which was connected to the node representing the cluster containing that value. The final taxonomy was obtained by ensuring that each increase in the generalization level corresponds to an increase in the anonymization level. To this end, generalization levels that produced no improvement in terms of k-anonymity were removed from the taxonomy. Table 6.6 presents an overview of the number of taxonomy levels and size for each attribute in the Electricity and the Adult datasets.

RFD Extraction. To extract the RFDs used to generate generalization rules, the *DOMINO RFD discovery algorithm* has been employed [93]. The advantage of using this algorithm is that it automatically infers not only RFDs from data, but also their associated thresholds. *DOMINO* extracts RFDs that hold on the

entire dataset, i.e., every tuple pair in the dataset should satisfy the RFD similarity constraint in order to be returned by *DOMINO*. This can be too restrictive when discovering roll-up RFDs over generalization taxonomies. Thus, an RFD discovery algorithm tolerating exceptions would be needed. However, the only discovery algorithm for hybrid RFDs existing in the literature is not capable of automatically discovering similarity and coverage thresholds, requesting the user to specify them in input [69]. Given that in the analyzed context the automatic derivation of thresholds is a fundamental requirement, since they represent the generalization levels to be used, the proposed methodology adopts a dataset sampling strategy, so that by using *DOMINO* on the sampled dataset it discovers roll-up RFDs that do not hold on the entire original dataset, hence increasing the set of discovered roll-up RFDs. In particular,*DOMINO* has been adapted to create a generalization map in which keys represent distance patterns and values represent the number of tuple pairs complying with each pattern has been created. Then, for each attribute in the considered dataset, a distance pattern (computed between each pair of tuples) maps the number of generalizations to use for including two attribute values in the same taxonomy level. In the experiments, the most frequent distance patterns yielding the coverage of an x-percentage of tuple pairs, with $x \in \{5, 10, 20, 50\}$, have been considered.

Anonymization & Utility Measures. To compute the anonymization level, the k-anonymity model proposed in [55] has been implemented, whereas the data utility measures used in the experiments, i.e., the accuracy and information gain, have been computed using the $J48$ decision tree implementation of Weka[4]. To guarantee the effectiveness of the obtained predictive model, the 10-fold cross-validation to compute data utility measures has been used.

[4]`https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html`

6.1.4.2 Results

This section presents the results of the experiments and answers the research questions.

RQ1: What is the impact of combining generalization rules on data utility? This research question aims to evaluate the benefits of combining RFDs to generate strategies for data anonymization, which maximize data utility while guaranteeing a desired level of privacy. It was expected that the combination of RFDs would provide generalization rules with higher data utility compared to those directly extracted from the data. To measure these effects, the number of generalized rules obtained by combining RFDs has been assessed, and the data utility, in terms of information gain and accuracy, has been compared. As a primary requirement for the evaluation, generalization rules that achieve k-anonymity with $k \geq 2$ have been considered.

Figures 6.4 and 6.5 show the accuracy that can be achieved using the generalization rules directly extracted from the data (red boxes) and using the combined rules (blue boxes) at the varying of sampling percentage for the Electricity and Adult datasets, respectively.

From Figure 6.4, it is possible to observe that, for the Electricity dataset, combining generalization rules improves information gain for all sampling percentages, except for the 50% sampling percentage. This is because many generalization rules extracted for this sampling percentage contain the same attributes with different generalization levels and, thus, they are incompatible (for more details see Section 6.1.3), or their combination violated the privacy requirement over k (i.e., $k < 2$). Similarly, Figure 6.5 shows that combining generalization rules improves accuracy also for the Adult dataset, although the improvement is less prominent for this dataset. It is worth noting that the accuracy achieved for the Adult dataset, when it is anonymized using generalization rules directly extracted from the data, is already relatively high (over 75% vs. 60% for the Electricity dataset), given that the accuracy achieved on the original data is 85% (vs. 75% for the

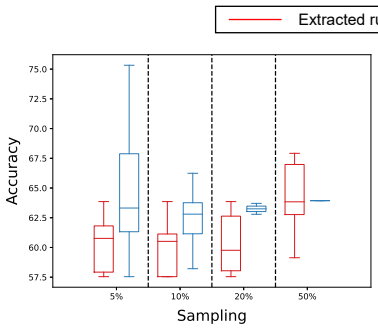


Figura 6.4: Accuracy achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Electricity dataset

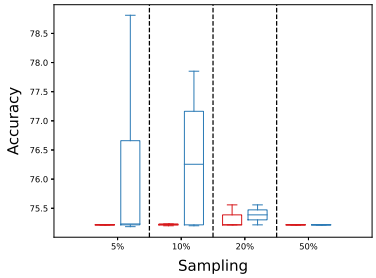


Figura 6.5: Accuracy achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Adult dataset

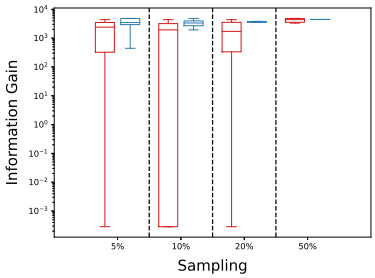


Figura 6.6: Information gain achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Electricity dataset

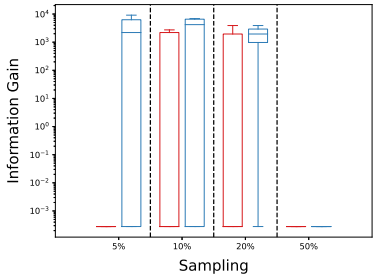


Figura 6.7: Information gain achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Adult dataset

Electricity dataset).

The experiments show that combining generalization rules also improves information gain for both the Electricity and the Adult dataset, as illustrated in Figures 6.6 and 6.7, respectively. Overall, consider more correlations in the data simultaneously and, thus, account for more attributes in the anonymization process, allows generating anonymized datasets holding a higher data utility.

RQ2: Which trade-off between anonymization and data utility can be achieved using generalization rules?

It was expected that the data utility would decrease when the anonymization level increased. This is because achieving a higher level of anonymization requires higher generalization levels, leading to the less specificity of data. These effects are quantified by showing how accuracy and information gain vary when the anonymization level increases.

Figures 6.8 and 6.9 show the trade-off between anonymization and accuracy, for the Electricity and the Adult datasets, respectively. The x-axis reports the anonymization levels (in log scale), whereas the y-axis reports the best accuracy that can be achieved by applying the generalization rules satisfying a given anonymization level. The baseline accuracy is obtained over the non-anonymized version of the datasets ($k = 1$). Each line in the plots represents the results for a given sampling percentage (5%, 10%, 20%, and 50%).

As expected, the accuracy decreases when the anonymization level increases for both datasets. Moreover, the highest anonymization level (represented by the vertical dashed lines) is achieved for the 5% sampling with both the Electricity and the Adult datasets. This could be justified by the fact that the 5% sampling not only generates a larger number of generalization rules, but also these rules typically encompass attributes with a higher generalization level. Nevertheless, differences can be noticed in the maximum anonymization level that can be achieved using different sampling percentages for the two datasets. For the Adult dataset, the maximum anonymization level that can be achieved ranges from 78 for the 50% sampling to 469 for the 5% sampling (cf. Fi-

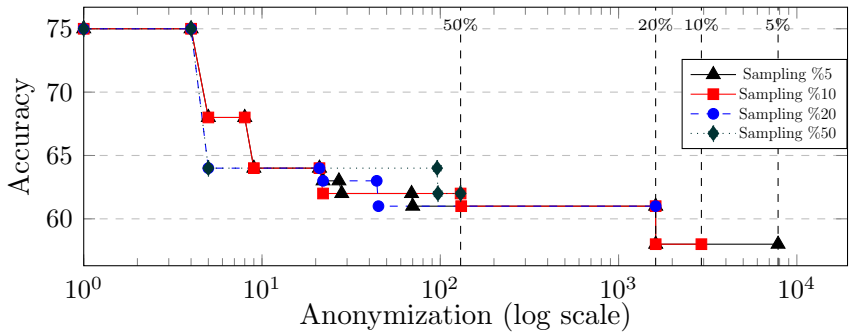


Figura 6.8: Trade-off between anonymization and accuracy for the Electricity dataset.

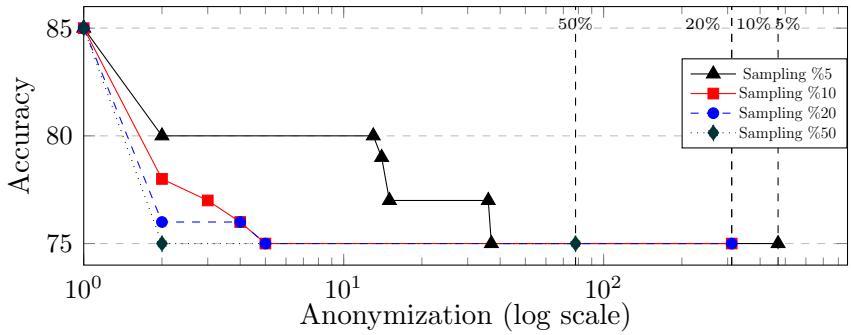


Figura 6.9: Trade-off between anonymization and accuracy for the Adult dataset.

gure 6.9). These differences are more evident for the Electricity dataset, where the maximum anonymization level ranges from 130 for the 50% sampling to 7846 for the 5% sampling (cf. Figure 6.8).

It is worth noting that for the Electricity dataset, all samplings preserve the baseline accuracy for $k \leq 4$. On the other hand, although none of the generalization rules extracted from the Adult dataset guarantees the baseline accuracy, the accuracy loss is limited between 5% and 10%. The smaller accuracy loss for the Adult dataset could be due to the defined attribute taxonomies, which generally have a higher depth for the Electricity dataset (cf. Table 6.6). This difference in the attribute taxonomies for the two

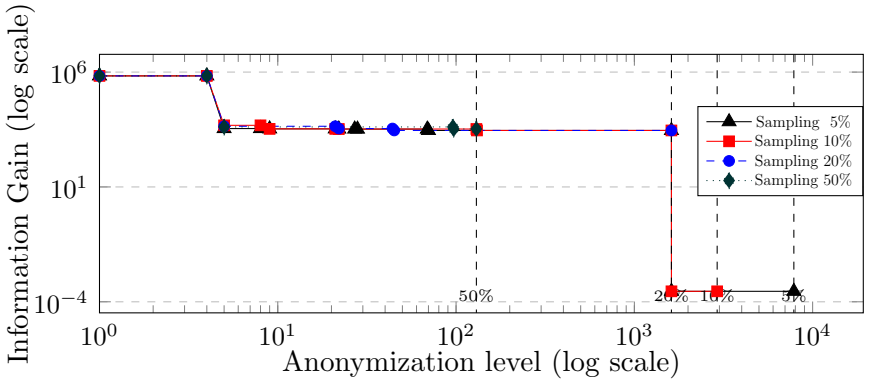


Figura 6.10: Trade-off between privacy and information gain for the Electricity dataset.

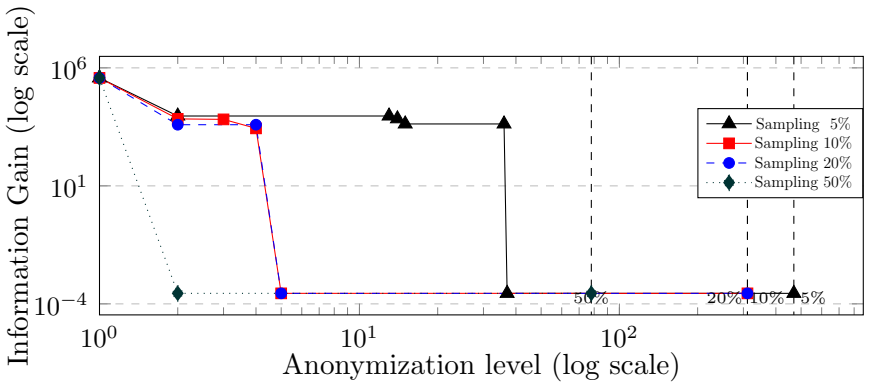


Figura 6.11: Trade-off between privacy and information gain for the Adult dataset.

datasets also affects the number of cut-off points, which is smaller for the Adult dataset.

Figures 6.10 and 6.11 show the trade-off between information gain and anonymization level.

Similarly to results obtained for the accuracy, information gain decreases when the anonymization level increases, for both the Electricity and the Adult datasets, and the highest anonymization level is achieved for the 5% sampling over both datasets.

Moreover, for the Electricity dataset, all samplings preserve the baseline information gain for $k \leq 4$.

However, it is possible to observe that, compared to accuracy, information gain decreases significantly faster, tending to zero at the increase of the anonymization level. In particular, for the Adult dataset, information gain is close to zero already with an anonymization level of 2 ($k = 2$) for the 50% sampling, and with an anonymization level of 5 ($k = 5$) for 10% and 20% samplings. For the 5% sampling, high information gain degrades to a value close to zero for higher anonymization levels ($k \geq 27$). On the contrary, this effect is less prominent for the Electricity dataset, where a high information gain can be achieved for extremely high anonymization levels ($k \geq 1614$). This is mainly because the Electricity dataset is characterized by several numerical attributes for which many generalization levels were included in their taxonomy.

RQ3: How much effort is required by a data owner to identify the generalization rule to apply? A large number of generalization rules can be returned by the proposed methodology, leaving the data owner with the burden to identify which generalization rule should be applied. To assist the data owner in this task, an approach based on Pareto-optimality to identify those rules providing a suitable trade-off between privacy and data utility is used (cf. Section 6.1.3). In what follows, an evaluation of such approach is proposed. More precisely, it has been evaluated the effort required to a data owner in determining the generalization rule to apply, in terms of the number of rules returned by the proposed methodology.

Tables 6.7 and 6.8 provide descriptive statistics on the number of generalization rules achieving k-anonyminity with $k \geq 2$, obtained using the proposed approach on the Electricity and Adult datasets, respectively. For each sampling percentage (*%sampling*), the tables report the number of rules directly extracted from the data (*Extracted*), the number of rules obtained by combining RFDs (*Combined*), the total number of obtained rules (*Total*), before (RFDs) and after (*Pareto*) the application of Pareto-optimality.

%sampling	RFDs			**Pareto**		
	Extracted	Combined	Total	Extracted	Combined	Total
5%	8	33	41	5	10	15
10%	9	16	25	5	4	9
20%	8	2	10	3	1	4
50%	10	1	11	4	0	4

Tabella 6.7: Number of generalization rules for the Electricity dataset.

%sampling	RFDs			**Pareto**		
	Extracted	Combined	Total	Extracted	Combined	Total
5%	25	53	78	4	6	10
10%	22	17	39	6	7	13
20%	7	2	9	3	2	5
50%	5	4	9	2	1	3

Tabella 6.8: Number of generalization rules for the Adult dataset.

From the results obtained before the application of Pareto-optimality (the three columns under header RFDs), it is possible to observe that combining RFDs always leads to the definition of new generalization rules. The sampling percentage has a large impact on the number of such rules: for both datasets, lower sampling percentages typically provide a larger number of combined generalization rules. This is mainly due to the fact that generalization rules obtained for lower sampling percentages typically involve few attributes, yielding many possibilities to combine them with each other.

The experiments show that the application of Pareto-optimality reduces the number of generalization rules to be considered by data owners when anonymizing their datasets (the three columns under header *Pareto*). In particular, it is possible to observe that the use of Pareto-optimality for filtering yields a reduction of the total number of generalization rules between 36% and 40% for the Electricity dataset, and between 12% and 56% for the Adult dataset, where the largest reduction is obtained for the

5% sampling. When deriving rules using low sampling percentages, Pareto-optimality tends to preserve more combined rules than rules directly extracted from the data, whereas this consideration is reversed when the sampling percentage increases. An inspection of the generalization rules extracted over low sampling percentages showed that these rules typically involve fewer attributes, yielding a larger set of combined rules, among which it is possible to find rules that maintain the same anonymization level while providing higher data utility. On the contrary, since rules extracted for high sampling percentages typically contain many attributes, their combination tends to decrease the anonymization level.

The results discussed so far show the capability of Pareto-optimality to significantly reduce the space of candidate generalization rules, with respect to the use of RFDs only. Overall, the candidate generalization rules returned by the proposed approach are in the orders of tens for both the Electricity and the Adult datasets. Nevertheless, exploring these solutions to determine which generalization rule should be applied is, at this point, up to the data owner. Visualizing the Pareto frontier can assist data owners in obtaining an overview of the space of the rules providing a suitable trade-off between data utility and anonymization level and, thus, in effectively carrying out their analysis with respect to their privacy and data utility requirements. As an example, Figures 6.12 and 6.13 show the Pareto frontier, represented by the red dots, for both the Electricity and the Adult datasets, when a 5% sampling is used for extracting RFDs. Based on the Pareto frontier, the data owner can determine the expected accuracy and information gain for a given anonymization level and, possibly, ensuring stronger privacy guarantees at the cost of decreasing one of these data utility metrics.

6.1.5 Discussion

The proposed methodology exploits the notion of RFD to support data owners in the anonymization of their dataset, aiming to let them achieve a given level of privacy while reducing the loss of da-

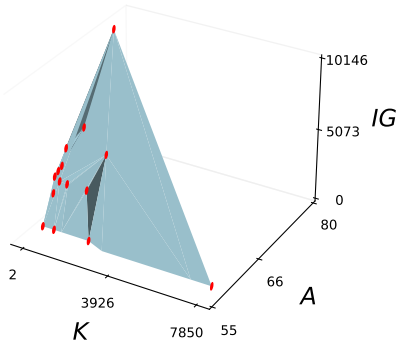


Figura 6.12: Pareto frontier for Electricity 5%.

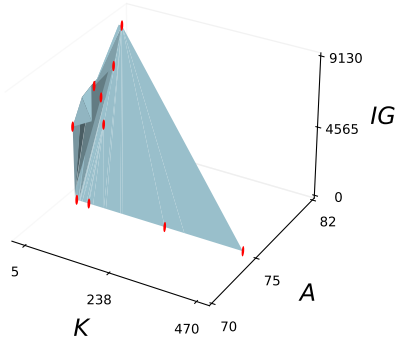


Figura 6.13: Pareto frontier for Adult 5%.

ta utility due to anonymization. In particular, the methodology uses RFDs automatically extracted from the data to define possible generalization rules and combines them to achieve a higher data utility. Then, Pareto-optimality is employed to identify those generalization rules that provide a suitable trade-off between privacy and data utility.

The effectiveness of the proposed methodology has been evaluated considering i) the impact of combining RFDs on data utility (RQ1), ii) the trade-off between anonymization and data utility (RQ2), and iii) the effort required to a data owner to identify the generalization rule to apply (RQ3).

Next, a summary of the lessons learned obtained by applying the proposed methodology over the considered real-world datasets is presented.

Using RFDs for defining generalization rules. Exploiting attribute correlations in terms of RFDs with capability to map possible generalization levels resulted in a novel and effective privacy preservation approach. In particular, the use of roll-up dependencies, i.e., the type of RFDs considered in this work, allows accounting for the generalization levels in the extraction of RFDs, thus directly considering their impact on the attribute to be classified. In the experiments, the *DOMINO* algorithm has been used for the discovery of this type of RFDs from the data (cf. Section 6.1.4.1). However, although this algorithm is capable

of automatically extracting both the RFDs and their associated similarity thresholds, it only extracts RFDs holding on the entire dataset, which can be too restrictive and yield the extraction of few or no RFDs when applied to real-world datasets. On the other hand, algorithms from the literature capable of tolerating exceptions require the data owner to specify thresholds in input, nullifying the benefits of the proposed methodology. Thus, to let *DOMINO* tolerate exceptions, a sampling strategy is used on input data and *DOMINO* has been applied only on the sampled portion of the dataset, yielding a higher number of generalization rules, as results show, hence achieving a higher privacy level. In general, it is possible to observe that exploiting data correlations, expressed in terms of roll-up dependencies, for the definition of anonymization strategies helps preserving the data utility of anonymized datasets.

Construction of attribute taxonomy. The results of the experiments show that the effectiveness of extracted generalization rules depends on the quality of the attribute taxonomies defining the generalization levels. In particular, it is possible to observed that the use of an attribute taxonomy comprising several generalization levels typically leads to a higher number of generalization rules (e.g., leading to the potential of finding more suitable trade-offs between privacy and data utility), from which the data owners can choose for the anonymization of their datasets. In this work, a generalization strategy based on VGH is used, as this approach better preserves data correlations compared to DGH. In particular, a clustering approach to build the taxonomies of categorical attributes is used. Although the overall results of our approach are promising, the obtained taxonomies for categorical attributes often contain a limited number of generalization levels.

Combining generalization rules to improve data quality. Combining generalization rules helps reducing data utility loss in the anonymized dataset, as this approach has the potential of accounting for a larger number of attributes over which the dataset is anonymized (Recall that the attributes which do not occur in the applied generalization rule are removed from the

dataset). Experiments showed that combined generalization rules always provide a higher data utility than the rules directly extracted from the data (cf. the results for **RQ1** in Section 6.1.4.2), thus offering an effective way to minimize data utility loss.

Privacy and data utility metrics. To assess the trade-off between privacy and data utility offered by anonymization strategies a number of metrics to measure data utility and privacy level of an (anonymized) dataset has been employed. In particular, the proposed methodology uses the k-anonymity model to measure the privacy level guaranteed by datasets, together with accuracy and information gain as data utility measures. While providing an effective measure for data anonymization, k-anonymity is susceptible to several attacks (see Section 3.1.2.1). This drawback has drawn the attention of the research community, and a large body of the literature has concerned metrics to evaluate the privacy guarantees of anonymized datasets, e.g. ℓ-diversity [53], t-closeness [50]. At the same time, several metrics have been proposed to measure the data utility of anonymized datasets (e.g., precision, recall, F-score, entropy, Gini index), where the choice of the data utility measure to be used depends on the purpose of the data publishing activities.

Selecting anonymization strategies. The experiments show that the number of obtained generalization rules remains manageable for being analyzed by a human (cf. the results for **RQ3** in Section 6.1.4.2). This suggests that the proposed approach can be effective in practice to obtain usable indications of the strategies that can be applied for the anonymization of a dataset. Additionally, the Pareto frontier provides a useful aid to data owners to visualize the achievable trade-off between privacy and data utility that generalization rules produce, letting them select the one that better fits their privacy and data utility requirements.

Capitolo 7

Conclusion and future work

This thesis presented methodologies to preserve information confidentiality in the GDPR context, and anonymization during data analysis, together with tools to improve user awareness concerning privacy issues in Web browsing and social network data sharing. In particular, in the majority of the proposals, data profiling strategies have been exploited, by adapting them to work in the context of data privacy. In detail, concerning the preservation of information confidentiality in the context of GDPR, data correlations have been used in terms of relaxed functional dependencies (RFDs) for identifying sensitive information and privatize them for data management purposes. Similarly, RFDs have also been considered to define an anonymization strategy to preserve individuals' privacy when data have to be analyzed through machine learning processes. Finally, data correlations have also been exploited to identify significant data patterns for discriminating malicious accounts in social network contexts.

As mentioned before, tools to enhance the users' awareness concerning privacy issues in Web browsing and data sharing over social networks have also been proposed. In particular, the study firstly aimed at understanding of how user data are shared among different network providers, by also investigating which informa-

tion are recoverable from the Web. Moreover, concerning privacy preservation in sharing data over social networks, the second aim was to understand the ethical aspects derived by the availability of data, and how to perform statistical analysis over the collected data in order to improve the users' awareness.

Results and insights obtained by each mentioned proposal, after their application and/or evaluation in their specific contexts, are presented in the following.

Improving user privacy awareness in Web browsing. In most cases, web service providers collect user's data without clearly describing which kind of data they collect, and how they exploit these data for their business analysis. The visual analytics tool VIPAT proposed in this context (see Section 4.2), enables users to observe changes in network environments through several interactive graphs, which visualize the providers that share user information, and the frequency rate of exchanged packets over the network. In this way, as demonstrated by a user study, the majority of users, especially those without computer science backgrounds, gained consciousness on how their data were exchanged, so enhancing their awareness about possible privacy risks.

Improving user privacy awareness in data sharing over social networks. Social network users tend to share a vast amount of information, among which sensitive ones, without taking care of how to manage privacy policies offered by social network platforms correctly. The tool SODA proposed in this dissertation (see Section 4.3) enabled to assess how easily data can be reconstructed from multiple social networks by analyzing the user profiles. Moreover, through the cross-social analysis, other significant user data have been also reconstructed by exploiting the combination of several social networks.

Fake account discrimination. One of the most critical problems in the social network domain is the discrimination of malicious accounts, since the latter can compromise the trustability of several network activities. To this end, the proposed technique (see Section 4.1) applies data profiling strategies in the social network context, and defines a new heuristic to derive the application

order of RFDs extracted from social network data for discriminating fake accounts. Results have shown that the defined heuristic prioritizes the application of RFDs that are more effective in discriminating fake accounts from those having human holders. It has also been shown how to use such results as a feature selection strategy, to simplify the learning phase, aiming at improving the classification results to discriminate fake accounts.

Information confidentiality preservation in the GDPR context. GDPR defines how companies must process and manage users' private data. To this end, it is necessary to devise methods supporting companies in the identification of privacy threats during advanced data manipulation activities. Thus, as demonstrated by several experimental sessions, the methodology proposed in such context (see Chapter 2) can help detect many confidentiality threats while encrypting a reduced number of attributes to prevent them. It represents one of the fist proposals capable to preserve unobscured data useful for data analytics processes, without risking to jeopardize users' privacy. Such risk can be avoided, since all possible implications of sensitive attributes are caught by RFDs, so enabling its proper management in the proposed strategy.

Anonymity preservation for analytics analysis. The application of analytic processes, such as machine learning, on personal data, could expose users to privacy risks. To this end, the methodology proposed in such context (see Section 6.1) exploits RFDs to identify suitable generalization rules to anonymize data before the execution of classification tasks. Results have demonstrated that this methodology permits to obtain a good trade-off between privacy and data quality requirements, so guiding the data owner in the selection of the optimal anonymization strategy to apply.

In general, the proposed methodologies and tools represent effective solutions to support users and companies in the management of personal data. In particular, there are several lessons learnt by the application of them into their specific application scenarios. Firstly, it would be useful to release (visual) tools to

improve the user awareness concerning privacy issues, which, as demonstrated, are particularly appreciated by non-expert users. Moreover, concerning scenarios entailing a proper analysis and management of personal data, it has been found that data profiling metadata, and in particular, relaxed functional dependencies (RFDs), resulted in a useful mean to guide privacy-preserving methodologies in different application scenarios.

Possible future work aims at extending the proposed methodologies and tools in several directions. First, it would be possible to exploit other metadata and/or other kinds of RFDs to further extend the proposed methodologies, given the availability of tools to efficiently extract them from data [8, 111, 69]. They can potentially detect additional useful properties for both improving the methodology effectiveness, enlarging the set of managed threats, and/or enforcing ranking and filtering strategies on the RFD discovery results. Moreover, one of the future goals is to embed the proposed methodologies within self-service data preparation tools, especially those targeted to end-users and data stewards. Moreover, another goal is to extend the experimental evaluations of the proposed tools, by enlarging datasets and data types. It would be also possible to design a comprehensive visual tool targeted at end-users, to enable them to properly define and manage access control and data sharing requirements (exploiting RFDs) over all network services they interact with.

Bibliografia

[1] J. Bonneau and S. Preibusch, "The privacy jungle: On the market for data protection in social networks," in *Economics of information security and privacy*. Springer, 2010, pp. 121–167.

[2] L. Caruccio, D. Desiato, and G. Polese, "Fake account identification in social networks," in *2018 IEEE international conference on big data (big data)*. IEEE, 2018, pp. 5078–5085.

[3] X. Lin, R. Lu, and X. S. Shen, "MDPA: multidimensional privacy-preserving aggregation scheme for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 10, no. 6, pp. 843–856, 2010.

[4] F. Li, B. Luo, and P. Liu, "Secure and privacy-preserving information aggregation for smart grids," *International Journal of Security and Networks, IJSN*, vol. 6, no. 1, pp. 28–39, 2011.

[5] A. C. Squicciarini, M. Shehab, and F. Paci, "Collective privacy management in social networks," in *Proceedings of the 18th International Conference on World Wide Web, WWW*, 2009, pp. 521–530.

[6] V. M. Shelake and N. Shekokar, "A survey of privacy preserving data integration," in *Proceedings of the 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques, ICEECCOT*, 2017, pp. 59–70.

[7] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-preserving record linkage for big data: Current approaches and research challenges," in *Handbook of Big Data Technologies*, 2017, pp. 851–895.

[8] B. Breve, L. Caruccio, S. Cirillo, D. Desiato, V. Deufemia, and G. Polese, "Enhancing user awareness during internet browsing." in *ITASEC*, 2020, pp. 71–81.

[9] L. Caruccio, D. Desiato, G. Polese, and G. Tortora, "Gdpr compliant information confidentiality preservation in big data processing," *IEEE Access*, vol. 8, pp. 205 034–205 050, 2020.

[10] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10 562–10 582, 2017.

[11] D. Banisar and S. Davies, "Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments," *J. Marshall J. Computer & Info. L.*, vol. 18, p. 1, 1999.

[12] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE access*, vol. 4, pp. 2751–2763, 2016.

[13] A. E. Attipoe, J. Yan, C. Turner, and D. Richards, "Visualization tools for network security," in *Proceedings of Visualization and Data Analysis*, 2016, pp. 1–8.

[14] D. W. Bachmann, M. E. Segal, M. M. Srinivasan, and T. J. Teorey, "NetMod: A design tool for large-scale heterogeneous campus networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 1, pp. 15–24, 1991.

[15] Y. Z. et al., "MVSec: Multi-perspective and deductive visual analytics on heterogeneous network security data," *Journal of Visualization*, vol. 17, no. 3, pp. 181–196, 2014.

[16] Z. Constantinescu, M. Vlădoiu, and G. Moise, "VizNet - Dynamic visualization of networks and internet of things," in *RoEduNet Conf.: Networking in Education & Research*, 2016, pp. 1–6.

[17] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery, "Scoring users' privacy disclosure across multiple online social networks," *IEEE access*, vol. 5, pp. 13 118–13 130, 2017.

[18] S. Bhagat, K. Saminathan, A. Agarwal, R. Dowsley, M. De Cock, and A. Nascimento, "Privacy-preserving user profiling with facebook likes," pp. 5298–5299, 2018.

[19] N. Dakiche, F. B.-S. Tayeb, Y. Slimani, and K. Benatchba, "Tracking community evolution in social networks: A survey," *Information Processing & Management*, vol. 56, no. 3, pp. 1084–1102, 2019.

[20] K. Li, L. Cheng, and C.-I. Teng, "Voluntary sharing and mandatory provision: Private information disclosure on social networking sites," *Information Processing & Management*, vol. 57, no. 1, p. 102128, 2020.

[21] N. R. Al-Molhem, Y. Rahal, and M. Dakkak, "Social network analysis in telecom data," *Journal of Big Data*, vol. 6, no. 1, pp. 1–17, 2019.

[22] N. Kumar and R. N. Reddy, "Automatic detection of fake profiles in online social networks," Ph.D. dissertation, 2012.

[23] I. Sen, A. Aggarwal, S. Mian, S. Singh, P. Kumaraguru, and A. Datta, "Worth its weight in likes: Towards detecting fake likes on instagram," in *Proceedings of the 10th ACM Conference on Web Science*. ACM, 2018, pp. 205–209.

[24] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," *Computers & Electrical Engineering*, vol. 65, pp. 165–177, 2018.

[25] R. Raturi, "Machine learning implementation for identifying fake accounts in social network," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 4785–4797, 2018.

[26] R. Kaur, S. Singh, and H. Kumar, "Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches," *Journal of Network and Computer Applications*, vol. 112, pp. 53–88, 2018.

[27] P. K. Roy and S. Chahar, "Fake profile detection on social networking websites: A comprehensive review," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 3, pp. 271–285, 2020.

[28] J. Miracle and M. Cheatham, "Semantic web enabled record linkage attacks on anonymized data." in *Privacy and the Semantic Web - Policy and Technology (PrivOn2016)*. CEUR-WS.org, 2016.

[29] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-preserving record linkage for big data: Current approaches and research challenges," in *Handbook of Big Data Technologies*. Springer, 2017, pp. 851–895.

[30] A. Cuzzocrea and L. Puglisi, "Record linkage in data warehousing: State-of-the-art analysis and research perspectives," in *2011 22nd International Workshop on Database and Expert Systems Applications*. IEEE, 2011, pp. 121–125.

[31] D. Vatsalan, P. Christen, and V. S. Verykios, "A taxonomy of privacy-preserving record linkage techniques," *Information Systems*, vol. 38, no. 6, pp. 946–969, 2013.

[32] D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, "Distance-aware encoding of numerical values for privacy-preserving record linkage," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 2017, pp. 135–138.

[33] S. M. Randall, A. P. Brown, A. M. Ferrante, J. H. Boyd, and J. B. Semmens, "Privacy preserving record linkage using homomorphic encryption," *Population Informatics for Big Data, Sydney, Australia*, vol. 10, 2015.

[34] C. Clifton, M. Kantarcioglu, A. Doan, G. Schadow, J. Vaidya, A. K. Elmagarmid, and D. Suciu, "Privacy-preserving data integration and sharing," in *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2004, pp. 19–26.

[35] D. Vatsalan, P. Christen, C. M. O'Keefe, and V. S. Verykios, "An evaluation framework for privacy-preserving record linkage," *Journal of Privacy and Confidentiality*, vol. 6, no. 1, 2014.

[36] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using bloom filters," *BMC medical informatics and decision making*, vol. 9, p. 41, 2009.

[37] T. Churches and P. Christen, "Blind data linkage using n-gram similarity comparisons," in *Proceedings of the 8th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD*, 2004, pp. 121–126.

[38] D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, "Distance-aware encoding of numerical values for privacy-preserving record linkage," in *Proceedings of the 33rd IEEE International Conference on Data Engineering, ICDE*, 2017, pp. 135–138.

[39] K. Schmidlin, K. M. Clough-Gorr, and A. Spoerri, "Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality," *BMC medical research methodology*, vol. 15, no. 1, p. 46, 2015.

[40] B. C. Fung, K. Wang, and S. Y. Philip, "Anonymizing classification data for privacy preservation," *IEEE transactions on knowledge and data engineering*, vol. 19, no. 5, pp. 711–725, 2007.

[41] B. C. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *21st international conference on data engineering (ICDE'05)*. IEEE, 2005, pp. 205–216.

[42] S. Kisilevich, Y. Elovici, B. Shapira, and L. Rokach, "kactus 2: Privacy preserving in classification tasks using k-anonymity," in *Annual Workshop on Information Privacy and National Security*. Springer, 2008, pp. 63–81.

[43] A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, 2008.

[44] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[45] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "Combining top-down and bottom-up: scalable sub-tree anonymization over big data using mapreduce on cloud," in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2013, pp. 501–508.

[46] O. Tene and J. Polonetsky, "Big data for all: Privacy and user control in the age of analytics," *Northwestern Journal of Technology and Intellectual Property*, vol. 11, p. xxvii, 2012.

[47] R. W. Proctor, E. E. Schultz, and K.-P. L. Vu, "Human factors in information security and privacy," in *Handbook of research on information security and assurance*. IGI Global, 2009, pp. 402–414.

[48] S. Agrawal, J. R. Haritsa, and B. A. Prakash, "FRAPP: a framework for high-accuracy privacy-preserving mining," *Data Mining Knowledge Discovery*, vol. 18, no. 1, pp. 101–139, 2009.

[49] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *Journal of official statistics*, vol. 2, no. 3, p. 329, 1986.

[50] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[51] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, pp. 1–36, 2015.

[52] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model." *Journal of Information Science and Engineering*, vol. 32, no. 1, pp. 63–78, 2016.

[53] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[54] A. Pawar, S. Ahirrao, and P. P. Churi, "Anonymization techniques for protecting privacy: a survey," in *2018 IEEE Punecon*. IEEE, 2018, pp. 1–6.

[55] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[56] M. Keyvanpour and S. S. Moradi, "Classification and evaluation the privacy preserving data mining techniques by using a data modification-based framework," *Computing Research Repository*, vol. abs/1105.1945, 2011.

[57] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE access*, vol. 4, pp. 1821–1834, 2016.

[58] R. P. Priyadarsini, S. Sivakumari, and P. Amudha, "Enhanced l–diversity algorithm for privacy preserving data mining," in *Annual Convention of the Computer Society of India*. Springer, 2016, pp. 14–23.

[59] Z. Abedjan, L. Golab, and F. Naumann, "Data profiling: A tutorial," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 1747–1751.

[60] P. Pawluk, J. Gryz, S. Hazlewood, and P. van Run, "Trusted data in ibm's master data management," *DBKDA 2011*, p. 1, 2011.

[61] B. Knight, E. Veerman, J. M. Moss, M. Davis, and C. Rock, *Professional Microsoft SQL Server 2012 Integration Services*. John Wiley & Sons, 2012.

[62] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[63] L. S. Meyers, G. C. Gamst, and A. Guarino, *Performing data analysis using IBM SPSS*. John Wiley & Sons, 2013.

[64] I. H. Witten, E. Frank, L. E. Trigg, M. A. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," 1999.

[65] D. Pyle, *Data preparation for data mining*. Morgan Kaufmann, 1999.

[66] Y. Sismanis, P. Brown, P. J. Haas, and B. Reinwald, "Gordian: efficient and scalable discovery of composite keys," in *Proceedings of the 32nd international conference on Very large data bases*, 2006, pp. 691–702.

[67] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: an efficient algorithm for discovering functional and approximate dependencies," *The computer journal*, vol. 42, no. 2, pp. 100–111, 1999.

[68] V. M. Markowitz and J. A. Makowsky, "Identifying extended entity-relationship object structures in relational schemas," *IEEE transactions on Software Engineering*, vol. 16, no. 8, pp. 777–790, 1990.

[69] L. Caruccio, V. Deufemia, and G. Polese, "Mining relaxed functional dependencies from data," *Data Mining and Knowledge Discovery*, vol. 34, no. 2, pp. 443–477, 2020.

[70] ——, "On the discovery of relaxed functional dependencies," in *Proceedings of the 20th International Database Engineering & Applications Symposium, IDEAS*, 2016, pp. 53–61.

[71] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of IJCAI-03 Workshop on Information Integration on the Web, IIWeb-03*, 2003, pp. 73–78.

[72] P. Biondi, "Scapy documentation (!)," https://scapy.readthedocs.io/en/latest/, 2019.

[73] E. den Heijer and A. Eiben, "Using scalable vector graphics to evolve art," *IJART*, vol. 9, no. 1, pp. 59–85, 2016.

[74] R. Jafri and H. Arabnia, "A survey of face recognition techniques," *JIPS*, vol. 5, pp. 41–68, 06 2009.

[75] Y.-Q. Wang, "An analysis of the viola-jones face detection algorithm," *Image Processing On Line*, vol. 4, pp. 128–148, 06 2014.

[76] A. Sharifara, M. S. M. Rahim, and Y. Anisi, "A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection," *IEEE*, pp. 73–78, 2014.

[77] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin," *Computing Research Repository*, vol. abs/2006.01381, 2020.

[78] "Admin view on linkedin pages," https://www.linkedin.com/help/linkedin/answer/98738/use-your-admin-view-on-linkedin-pages, accessed: 2021-08-25.

[79] D. Punkamol and R. Marukatat, "Detection of account cloning in online social networks," in *2020 8th International Electrical Engineering Congress (iEECON)*. IEEE, 2020, pp. 1–4.

[80] P. Bródka, M. Sobas, and H. Johnson, "Profile cloning detection in social networks," in *2014 European Network Intelligence Conference*. IEEE, 2014, pp. 63–68.

[81] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[82] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, 2020.

[83] Q. Zhao and S. S. Bhowmick, "Association rule mining: A survey," Nanyang Technological University, Singapore, Tech. Rep., 2003.

[84] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[85] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[86] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[87] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands*, pp. 403–412, 2018.

[88] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design," *IEEE transactions on neural networks*, vol. 11, no. 1, pp. 124–136, 2000.

[89] F. O. Redelico, F. Traversaro, M. d. C. García, W. Silva, O. A. Rosso, and M. Risk, "Classification of normal and pre-ictal eeg signals using permutation entropies and a generalized linear model as a classifier," *Entropy*, vol. 19, no. 2, p. 72, 2017.

[90] D. Desiato, "A methodology for GDPR compliant data processing," in *Proceedings of the 26th Italian Symposium on Advanced Database Systems, Castellaneta Marina (Taranto), Italy, June 24-27, 2018*, ser. CEUR Workshop Proceedings, S. Bergamaschi, T. D. Noia, and A. Maurino, Eds., vol. 2161. CEUR-WS.org, 2018.

[91] European Commission, "General data protection regulation - final version of the regulation," 2016, released 6 April 2016. [Online]. Available: http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

[92] ——, "Recital 71 of reform of EU data protection rules 2018," 2018. [Online]. Available: https://www.privacy-regulation.eu/en/r71.htm

[93] L. Caruccio, V. Deufemia, F. Naumann, and G. Polese, "Discovering relaxed functional dependencies based on multi-attribute dominance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 9, pp. 3212–3228, 2021.

[94] F. V. Fomin, S. Gaspers, A. V. Pyatkin, and I. Razgon, "On the minimum feedback vertex set problem: Exact and enumeration algorithms," *Algorithmica*, vol. 52, no. 2, pp. 293–307, 2008.

[95] W. Stallings, "The offset codebook (OCB) block cipher mode of operation for authenticated encryption," *Cryptologia*, vol. 42, no. 2, pp. 135–145, 2018.

[96] O. Reparaz and B. Gierlichs, "A first-order chosen-plaintext DPA attack on the third round of DES," in *Proceedings of the 16th International Conference on Smart Card Research and Advanced Applications, CARDIS*, 2017, pp. 42–50.

[97] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," http://archive.ics.uci.edu/ml/index.php, 1998, last accessed: May 11rd, 2018.

[98] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Machine learning*, vol. 3, pp. 319–342, 1989.

[99] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49–58, 2019.

[100] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE access*, vol. 6, pp. 12 103–12 117, 2018.

[101] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 70–78.

[102] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[103] H. Goldstein and N. Shlomo, "A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets," *Journal of Official Statistics*, vol. 36, no. 1, pp. 89–115, 2020.

[104] T. Šarčević, D. Molnar, and R. Mayer, "An analysis of different notions of effectiveness in k-anonymity," in *International Conference on Privacy in Statistical Databases*. Springer, 2020, pp. 121–135.

[105] K. Mohammed, A. Ayesh, and E. Boiten, "Utility promises of self-organising maps in privacy preserving data mining," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 2020, pp. 55–72.

[106] G. Loukides and J. Shao, "Data utility and privacy protection trade-off in k-anonymisation," in *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, 2008, pp. 36–45.

[107] C. Jin, L. De-Lin, and M. Fen-Xiang, "An improved id3 decision tree algorithm," in *2009 4th International Conference on Computer Science & Education*. IEEE, 2009, pp. 127–130.

[108] Y. Jin and B. Sendhoff, "Pareto-based multiobjective machine learning: An overview and case studies," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 3, pp. 397–415, 2008.

[109] A. V. Lotov and K. Miettinen, "Visualizing the pareto frontier," in *Multiobjective optimization*. Springer, 2008, pp. 213–243.

[110] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[111] L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese, "Real-time visualization of profiling metadata upon data insertions." in *EDBT/ICDT Workshops*, 2021.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo