Journal of Big Data

RESEARCH Open Access

Social network data analysis to highlight privacy threats in sharing data

Francesca Cerruto, Stefano Cirillo*, Domenico Desiato, Simone Michele Gambardella and Giuseppe Polese

*Correspondence:
scirillo@unisa.it
Department of Computer Science, University of Salerno, via Giovanni Paolo II, n.132, 84084 Fisciano, SA, Italy

Abstract

Social networks are a vast source of information, and they have been increasing impact on people's daily lives. They permit us 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. In this work, we have collected data concerning people who have different social network profiles, aiming to analyse privacy requirements offered by social networks. In particular, we have built a tool exploiting image-recognition techniques to recognise a user from his/her picture, aiming to collect his/her personal data accessible through social networks where s/he has a profile. We have composed a dataset of 5000 users by combining data available from several social networks; we compared social network data mandatory in the registration phases, publicly accessible and those retrieved by our analysis. We aim to analyse the amount of extrapolated data for evaluating privacy threats when users share information on different social networks to help them be aware of these aspects. This work shows how users data on social networks can be retrieved easily by representing a clear privacy violation. Our research aims to improve the user's awareness concerning the spreading and managing of social networks data. To this end, we highlighted all the statistical evaluations made over the gathered data for putting in evidence the privacy issues.

Keywords: Privacy, Social networks, Data analysis

Introduction

Plenty of people are registered over several social networks, where they share a huge amount of information. In particular, most of the social platforms like Facebook, Twitter, and Instagram permit people to share emotions, ways of thinking, points of view, and so on. Social networks play a fundamental role in human interaction. We could say that a virtual life, strictly coupled with the real one, is lived through social networks.

Among all the information shared by users, we are interested in collecting user's data, mostly because, for having a social profile, it is required to insert specific information that characterises a network user. Preserving the privacy of users is a crucial task that social networks have to handle for not jeopardising user's data. Recently, it has been estimated that Facebook has 2.5 billion active users around the world,[1] Twitter has 1.3

[1] https://socialmediamarketing.com/.

billion,[2] and Linkedin has millions of profiles[3]. With such a vast number of users, the need to analyse how they manage privacy over several social networks is becoming vital to detect which information is not privatised, by making users more aware of how to prevent privacy issues.

Privacy issues can also arise during normal user web browsing activities, since they enable network providers to collect a vast amount of information for different business reasons [1–4]. There exist several methodologies for preserving privacy in various applications area, but it is difficult to exploit them in the domain of the social network [5, 6]. The General Data Protection Regulation (GDPR) [7] was defined for specifying specific policies to handle user's data, but it appears to be not always applicable over the social networks domain. When a network user creates a social network profile, often the only advice concerning the management of his/her data is given through management policies hard to interpret. Only a few users are deeply aware of issues related to incorrect usage of information, and many of them spread this information without applying privacy filters.

Users tend to use social networks to share information massively; in most cases, they do not care about privatising 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 issues. In a context in which a social network profile contains detailed information that univocally refers to a specific individual to preserve his/her privacy became a fascinating challenge. With this in mind, the motivations for supporting our study are to make users aware of privacy issues linked to mismanagement of social networks' privacy policies.

In this scenario, several studies have described data privacy issues on social networks [8–10], but only some of them have provided tools capable of improving users' awareness when sharing their data on social networking platforms [1]. In our work, we perform cross social network analysis on 5 platforms to figure out which is the information that is most frequently shared over social networks, and that can jeopardize user's privacy. By exploiting face recognition and data analysis, we have built the social data analyzer (SODA), a tool for extrapolating users' information made available from social networks, aiming to perform an accurate analysis for revealing privacy threats linked to incorrect usage of data sharing in social networks. This has enabled us to evaluate the sensitiveness of information connected to a specific user. Additionally, we have performed an exhaustive analysis to understand how social networks can permit the reconstruction of user's data even if some of them have been protected on other social networks. Furthermore, (SODA) is independent of the privacy settings offered by social networks; it simulates the search of a user retrieving data available in his/her social network profile. In other words, if a user has privatised specific information over a specific social network (SODA) is not able to retrieve that information, but if the same user has the same information not privatised over a different social network (SODA) can retrieve such information. Thus, we could say that (SODA) tests the users' skills in managing privacy settings offered by social network platforms.

[2] https://www.websitehostingrating.com/twitter%20statistics/.

[3] https://kinsta.com/blog/linkedin-statistics/.

In summary, the main contributions of our study are:

- analysing users' data extrapolated from several social networks to evaluate their privacy;
- improving the users' awareness concerning privacy threats in social network platforms.

The paper is organised as follows. In "Related work" section we discuss related works. In "Methodology" section we present our methodology, whereas "Social data analyzer" section presents the SODA tool and illustrates some concepts related to our methodology. In "Experimental evaluation" section we describe the results of our analysis. Finally, conclusions and future research directions are discussed in "Conclusion and future directions" section.

Related work

This section discusses relevant articles in which social networks privacy-preservation is addressed to evaluate risks connected to personal user data.

In the context of privacy preservation for sharing data in social network platforms, several approaches define strategies to make users aware of the privacy issues linked to their posted data. In [11], 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. Besides, 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. The authors have analyzed data concerning 15 users registered over different social networks (Facebook, ResearchGate, LinkedIn, and Google+) to perform their analysis. The main differences concerning our work are the methodology used for collecting data and the analysis made over them, i.e. the number of examined users and the social networks considered. In [12], 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 sector to steal information on the users' interests, to propose his/her targeted advertising, and to capture and reconstruct his/her data. However, this study is limited to consider only common information, i.e., "Likes", without permitting an in-depth investigation of user's data. Instead, in our proposal, we deeply investigate the user profiles to analyze privacy issues concerning his/her data. In [13], the authors explore the privacy-preserving actions regarding information sharing on Facebook. First, they study the information sharing behaviour of the elderly by observing the extent to which they opt-out of sharing information publicly about themselves on their profile pages. Furthermore, the authors observe how much overlap exists between these older Facebook users and their respective friends regarding their public information-sharing habits and explore the differences across gender. In [14], the authors survey the literature concerning privacy in social networks by focusing on online social networks and online affiliation networks. They

formally define the possible privacy breaches and describe the privacy attacks that have been studied. In addition, they present definitions of privacy in the context of anonymization together with existing anonymization techniques.

Social network data represents a rich source of information, mainly when it characterizes users, and malicious users can jeopardize the user's privacy by performing targeted attacks to recover sensitive information. In [15], authors aim to prevent a new type of attack based on the violation of the privacy of "friendly users". In particular, users tend to hide their personal information from people who do not belong to their social network. At the same time, they usually share information with users included in their friend's system, such as "Friendly user". Authors explain that an attacker can violate the user's privacy by studying his/her social network. They present a new algorithm to ensure the anonymity of mutual friends and prevent privacy threats during social networks usage. In [16], aspects linked to track community evolution over time in dynamic social networks are analyzed through a survey. Authors detail a classification of various methods to track community evolution in the dynamic social network. They describe four main approaches by using as a criterion the working principle: (1) based on independent successive static detection and matching; (2) based on dependent successive static detection; (3) based on the simultaneous study of all stages of community evolution, and (4) concerns methods working directly on temporal networks. Authors also provide basic concepts concerning social networks, community structure, and strategies to evaluate community detection methods by describing several approaches with their strengths and weaknesses. In [17], 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 disclose voluntarily or mandatorily. Authors show that voluntary sharing is more likely to be driven by positive factors, such as perceived benefits, social network size, and personalization. Simultaneously, mandatory provision is affected by individual characteristics such as age, privacy policy, and perceived risks. They highlight that perceived risk has less impact on voluntary sharing than previous studies suggested.

Many social networks share information to connect various accounts. However, without suitable protocols capable of ensuring privacy, a user might not be aware that his/her sensitive information could become visible to others. In [18], the authors describe several protocols to create and interact with privacy-preserving collaborative social networks. They implement and validate these protocols and investigate potential improvements in terms of privacy. In [19], an investigation on how to optimize the trade-off between latent-data privacy and customized data utility is described. The authors illustrate a data sanitization strategy that considers the benefits related to social network data and the protection of sensitive latent information. They also aim to preserve both data benefits and social structure while guaranteeing optimal latent-data privacy. In the end, they show their achieved results by highlighting the pros and cons of their sanitization strategy in terms of privacy reached over user data. In [20], an exhaustive study on the quantity of data available in social networks, together with their analysis, is presented, whereas in [21], authors emphasize the benefits of blending methods by integrating qualitative and quantitative approaches for evaluating the strengths and weaknesses

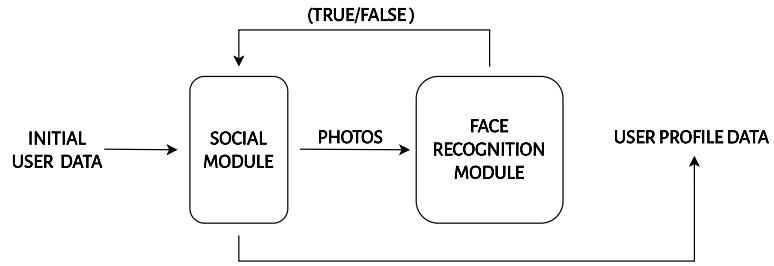

Fig. 1 Single social data extrapolation step

of social networks. In [22], authors describe big data daily produced by social networks and the entire web by showing several techniques to analyze them, whereas in [23], an interesting study concerning the development of social network analysis is presented. It describes the origin of the social network in classical sociology and its more recent formulation in social scientific and mathematical work. Furthermore, the authors illustrate different application areas in which social network analysis has been used. With this in mind, they linked social network analysis with other application areas, such as kinship structure, social mobility, science citations, corporate power, international trade exploitation, class structure, and many other areas. Finally, in [24] properties and characteristics of several social networks are examined to extract company data available over them. Nevertheless, the evaluation results made by the authors are different from ours in terms of analyzed data.

Concerning machine learning applications to preserve privacy in social network contexts, in [25], a comprehensive survey of multiple applications of social network analysis using robust machine learning algorithms is reported. In [26], the authors defined a privacy preservation algorithm that incorporates supervised and unsupervised machine learning anomaly detection techniques with access control models. They evaluated the algorithm over real datasets achieving over 95% accuracy using a Bayesian classifier and 95.53% using deep neural networks. The Authors in [27] perform a depression analysis using machine learning approaches over Facebook data collected from an online public source. They evaluated the efficiency of their method using a set of various psycholinguistic features. The authors put in evidence that their method can significantly improve the accuracy and classification error rate by revealing that the Decision Tree obtains the highest accuracy than other machine learning approaches to discriminate the user's depression.

Finally, a recent study used data from people from social networks to find Multi-SIM subscribers within the same operator or between operators for improving campaigns and churn prediction models of Telecom customers [28].

Methodology

In this section, we describe our methodology by summarising it in two meaningful steps: the single- and the cross-social data extrapolation steps.

In the single social data extrapolation step (Fig. 1), the picture and the name of the user are exploited as the input of the Social Module. The latter performs specific operations only

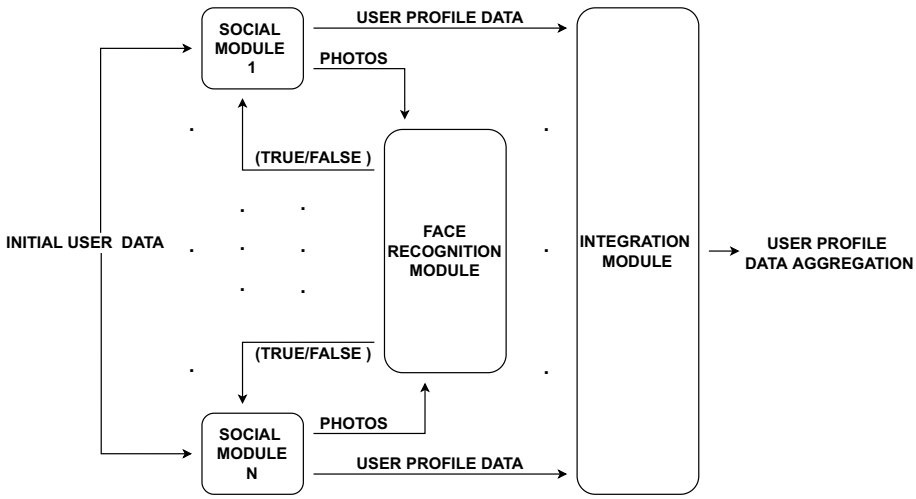

Fig. 2 Cross-social data extrapolation step

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 homonymy of user's names. In Fig. 1, it is possible to see the general process of the single social data extrapolation step.

In the cross-social data extrapolation step (Fig. 2), 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 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 Fig. 2, it is possible to see the general process of the cross social data extrapolation step.

In our methodology, we differentiate a single analysis over a specific social network from a cross-analysis over multiple social networks. In this way, we can 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.

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

Social data analyzer

Extracting user data from multiple social networks is a complex problem. There are several issues related to extraction that yield specific choices for the components of the SODA tool: (1) the number of users involved in the analysis process can be large, (2)

each social network relies on different implementation technologies, and (3) continuous updates of the social network platforms require continuous maintenance of system components. To this end, we have built the tool SODA on top of the existing system Social Mapper,[4] extending several of its components aiming to tackle the issues mentioned above.

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, 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 information: name, surname, city, email, or the company in which the user works. From these, SODA is capable of browsing the web by exploiting Selenium,[5] a framework that is generally used for activities such as testing, browsing, and scraping web content. Thanks to its features, 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 performing 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 [29], Social Mapper relies on the Viola-Jones [30], 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 [31].

With respect to Social Mapper, the proposed tool SODA provides several novel functionalities that allow us 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 of their public profile as an output. The idea of starting with Linkedin for selecting users to analyse is because these users with very high probability were not fake. In fact, it has demonstrated in [32] the amount of faker users registered over Linkedin is very small. Moreover, since we start from the list of people working for companies, the probability of finding a fake user is also very small. In fact, Linkedin provides every company with a tool to monitor the users registered on them [33]. 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 professionalism of

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

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

the company. To this end, exploiting Linkedin for selecting users represents an innovative functionality that permits us 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 we redesigned these modules 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, we 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 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 most closely 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 correspond 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 SODA 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 this 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 times of SODA. To this end, we consider a lower threshold value, i.e., 1, in order to speed up the search time and to 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 [34, 35].

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

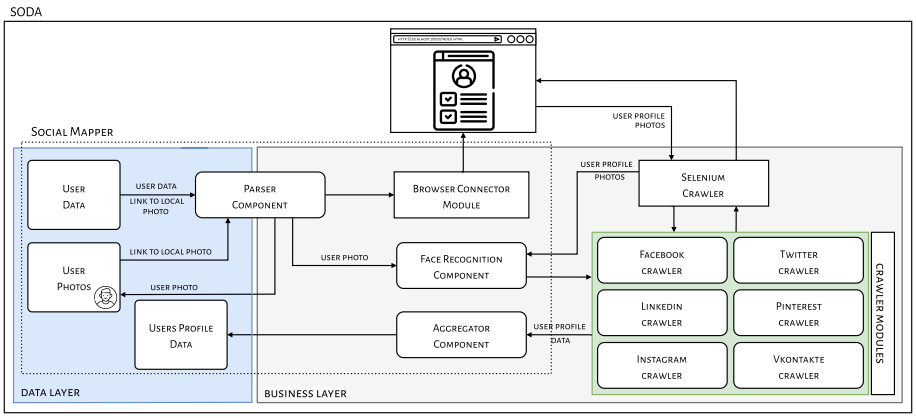

Fig. 3 Architecture of SODA

SODA architecture

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

- 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 our 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 the event of the 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 3 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, we defined six modules, one for each social network on which we can 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.

Experimental evaluation

In this section, we present a single-social and a cross-social evaluation, aiming to investigate the sensitivity of the extrapolated data. In what follows, we describe 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.

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, we implemented an ad-hoc crawler to interact with different web pages and extract only information characterizing the user. To this end, we have 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 our dataset has been to select people from different parts of the world. In particular, we exploited the new feature of SODA enabling search people working for a specific company. To this end, we have randomly selected more than 100 international companies, from which SODA extracted more than 11,000 images of distinct users. The new data have been aggregated into a single structured file and were 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, we only stored a subset of users, i.e., those yielding zero or few null values. Thus, we have selected the data of 7000 users with their information, and these have been considered as initial data for our evaluation. After the execution of SODA, we retrieved data from 5000 users, i.e., 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

As described in the previous section, the experimental evaluation has involved 7000 people randomly selected from the web. Starting from these, we performed the analysis on each social network, also including Linkedin, intending to evaluate the effectiveness of SODA. In particular, among the people involved in our 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 are not registered on any social network were classified as (*TN*), and in our evaluation, we can consider them 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, our experimental evaluation lasted several months, and it is, therefore, understandable that in the meantime, users can 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 [36, 37]. However, although these problems could be affected evaluation results, the effectiveness of SODA is shown with the following metrics:

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

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

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

Evaluation single-social

In this section, we describe statistics obtained by evaluating data extracted by each considered social network. In particular, we highlight the information that are frequently shared by users over every single social network, and analyse how each social network preserves user privacy. For this reason, starting from the 5000 users containing in our dataset, we perform a single social network evaluation. This allows us 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 shows the most frequently shared information on Linkedin extracted by 1570 users registered to it. Among the 5000 initial users, we have considered only the accounts from which we can extract sensitive information useful for our analysis. We can 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 Fig. 4 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.

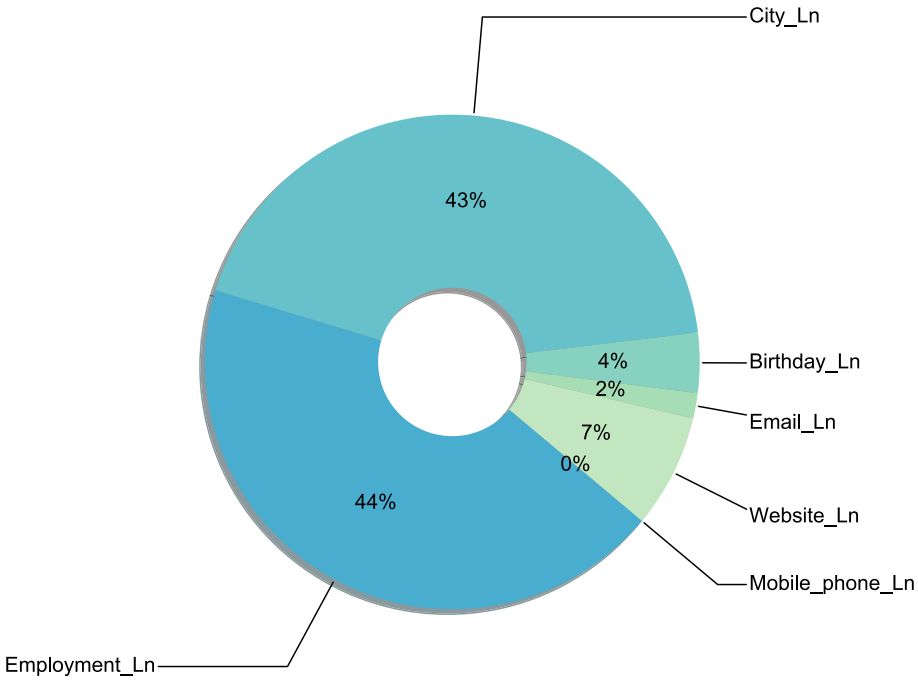

Fig. 4 Analysis of information shared on Linkedin

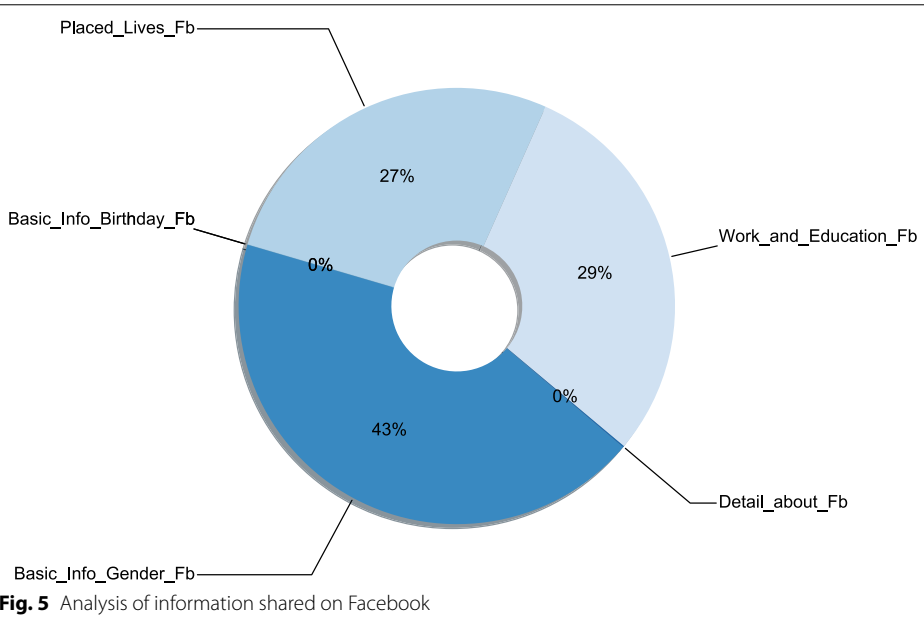

Fig. 5 Analysis of information shared on Facebook

Figure 5 shows the most frequently shared information on Facebook, extracted by 1161 users registered to it. We can 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 Fig. 5, no user has shared his/her details on the date of birth, which combined with the other data could

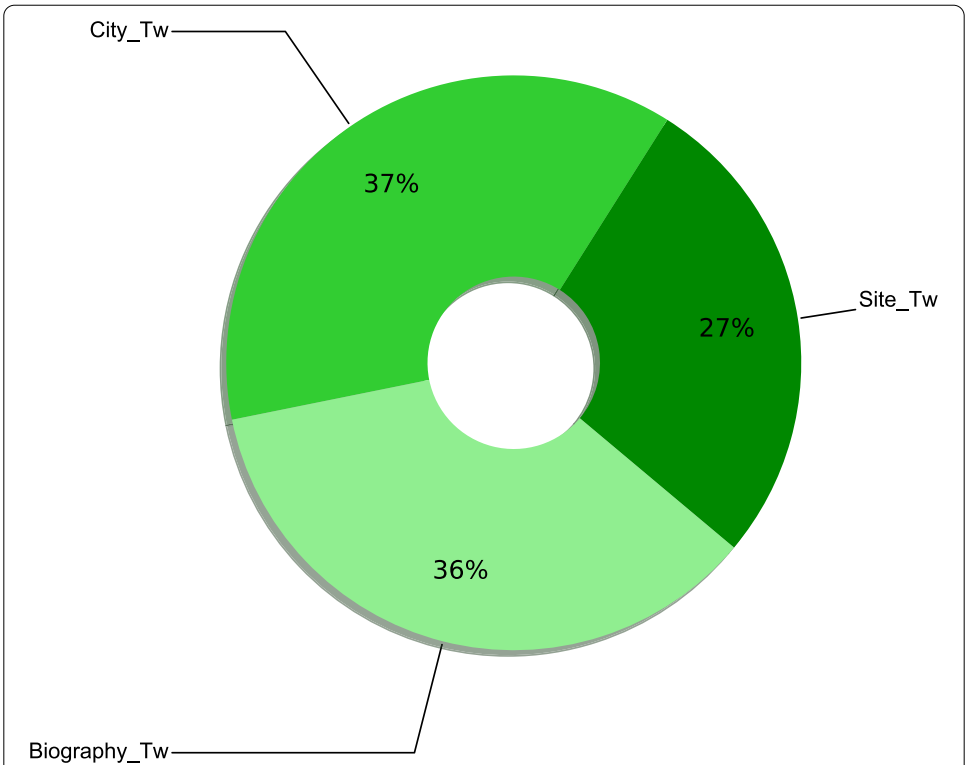

Fig. 6 Analysis of information shared on Twitter

significantly affect privacy. Facebook permits users to hide their date of birth in order to preserve privacy.

Figure 6 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 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 7 shows the most frequently shared information on VKontakte, extracted by 251 users registered to it. We can 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 Fig. 7, 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.

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

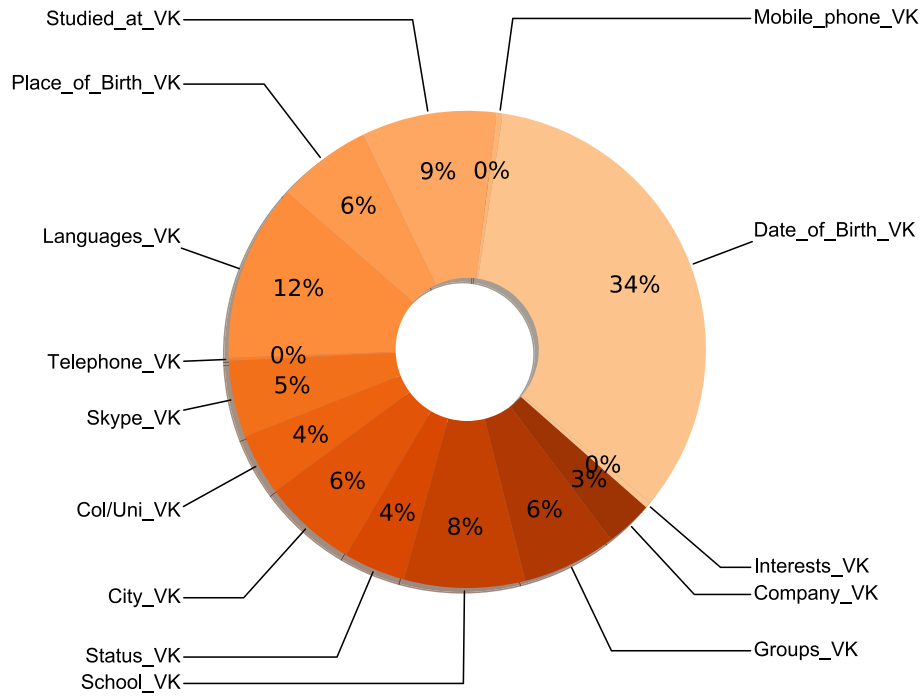

Fig. 7 Analysis of information shared on VKontakte

In Table 1, we summarise, for each specific social network analysed, the information retrieved by it. Yet, we compare "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 our analysis for a specific social network), and "Number of extracted attributes" (i.e., the amount of extracted attributes for each specific social network). As we can notice in Table 1, except for Twitter and Instagram, all other social networks permit us to retrieve different information that starting from the Public attributes can allow us to reconstruct a partial user's profile.

Evaluation cross-social

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

Figure 8 shows the distribution diagram for the users registered over the considered social network platforms. In particular, except for the first bar that highlights the number of users involved in no 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 we can see from Fig. 8, there are no users discovered in more than three social network platforms, and Instagram represents the most used platform from the users involved in our evaluation. However, in Fig. 9, 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

Table 1 Single social features extrapolation

	Required data	Public data	Data extracted
Linkedin	Name and Surname E-mail	Name and Surname City Employment Birthday	Full_Name Mobile_phone_Ln Website_Ln Email_Ln Birthday_Ln City_Ln Employment_Ln
Facebook	Name and Surname Phone Number Birthday Gender	Name and 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 and Surname E-mail Phone Number Birthday	Name and Surname City Biography Website	Full_Name Site_Tw City_Tw Biography_Tw
Instagram	Name and Surname E-mail Phone Number	Name and Surname Biography Website	Full_Name Biography_In
VKontakte	Phone Number Birthday Gender Name and Surname	Name and 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

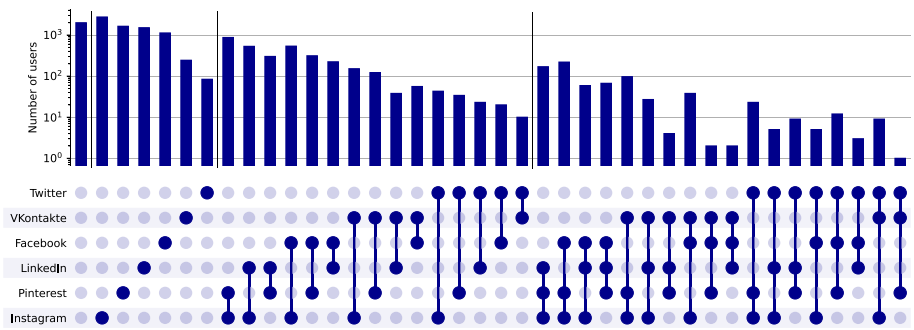

Fig. 8 Distribution diagram of the analyzed users

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.

In Fig. 9, the statistics concerning email sharing over different social networks are shown. By analysing different social networks, we can notice that only Linkedin,

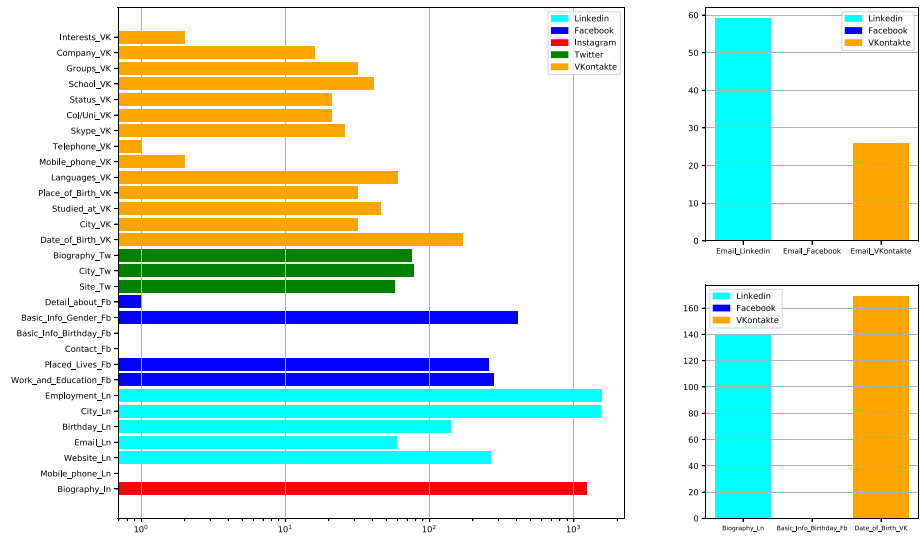

Fig. 9 Attribute statistics of the entire dataset

Facebook, and VKontakte have a special section for inserting this information. Concerning the email histogram in Fig. 9, 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 a high frequency for sharing the attribute *Email*, whereas few are the users shared it on VKontakte, and no one on Facebook.

In Fig. 9 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 Fig. 9, 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, we 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 our analysis, but we could 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 Fig. 10, 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 Fig. 10, 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

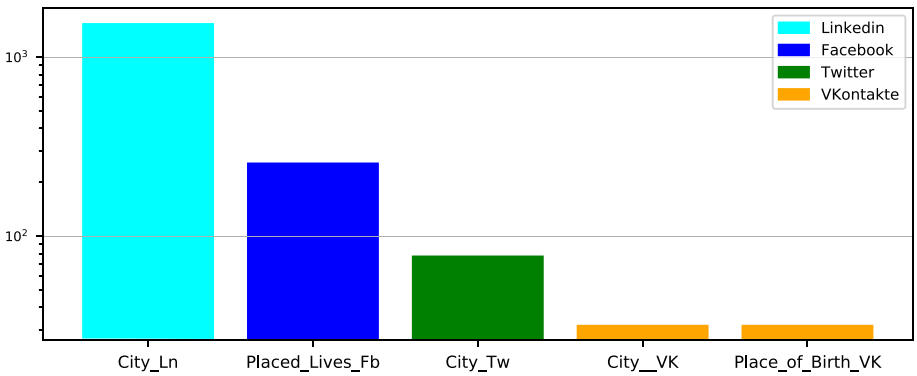

Fig. 10 City attribute shared over analyzed social networks

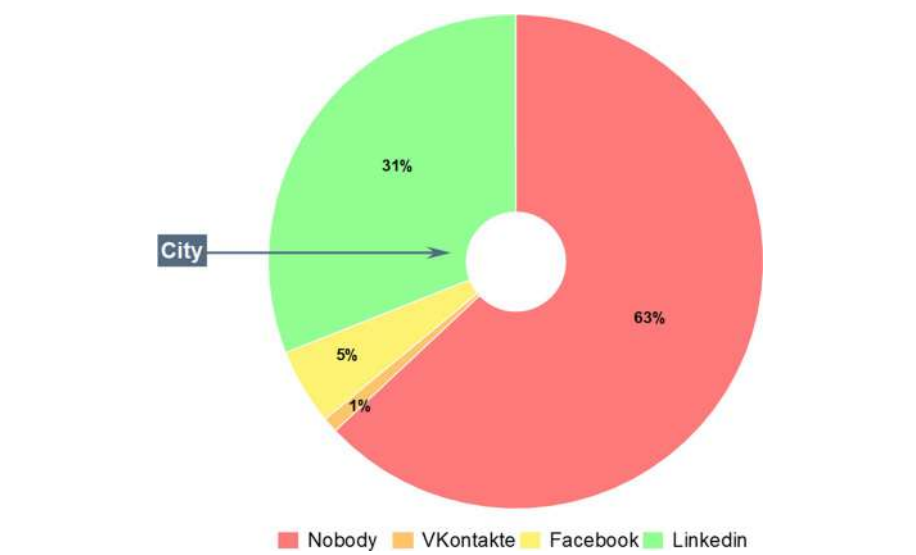

Fig. 11 Job attribute shared over analyzed social networks

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 Fig. 11, 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

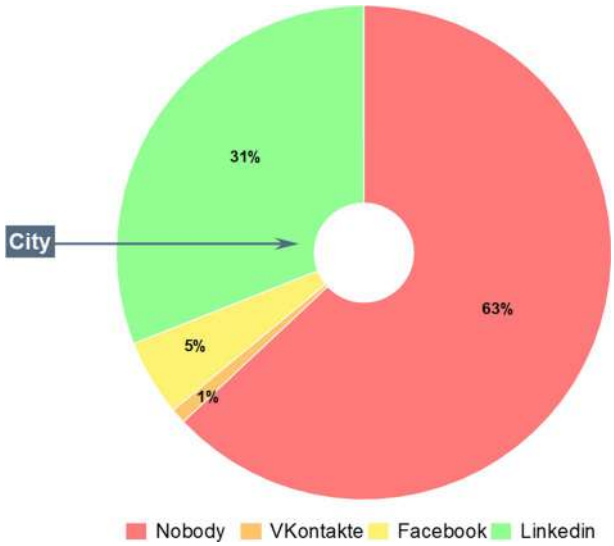

Fig. 12 Attributes reconstruction by exploring all analyzed social except for Twitter

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 our analyses, privatised data, i.e., data that is not publicly available on user profiles, and the data of the users 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 12 shows that 4923 users out of 5000 analyzed, are not registered on Twitter or have privatized this information on it. However, 31% 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 our analysis on Facebook, we have merged the last two attributes. In Figure 13, we show the percentage of information privatised by Facebook users, but published on other social networks:

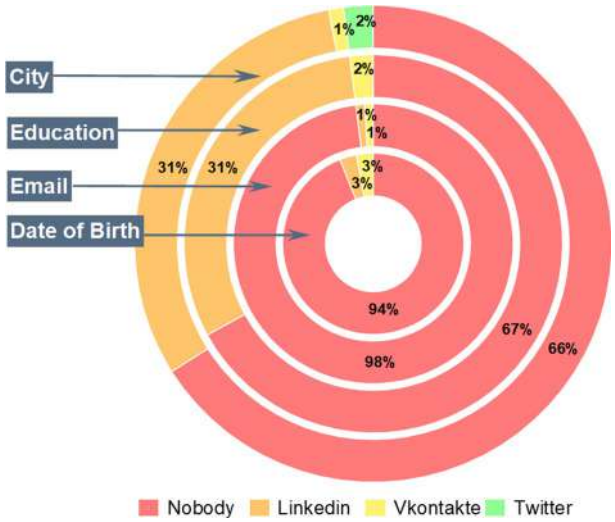

Fig. 13 Attributes reconstruction by exploring all analyzed social except for Facebook

- 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 allows 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 to 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 to other social networks;
- Among the 5000 analyzed users who have privatized the *Email* on Facebook or are not registered to 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 to 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 14, we show the percentage of information privatised on Linkedin, but published on other social networks:

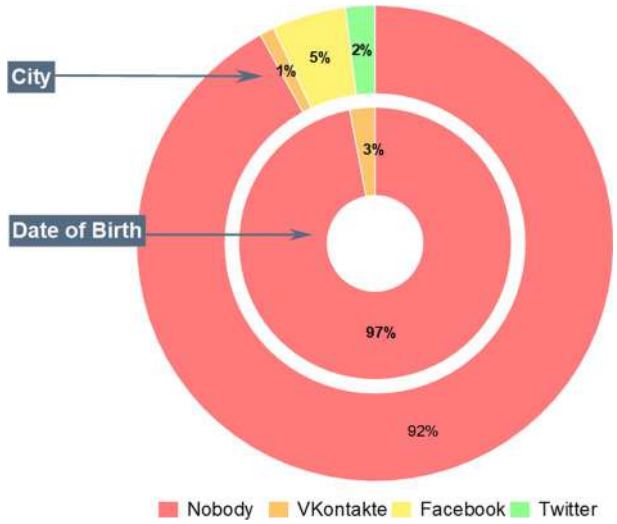

Fig. 14 Attributes reconstruction by exploring all analyzed social except for Linkedin

- 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 are not registered to 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 to 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 to 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 to 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 to 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 Fig. 15,

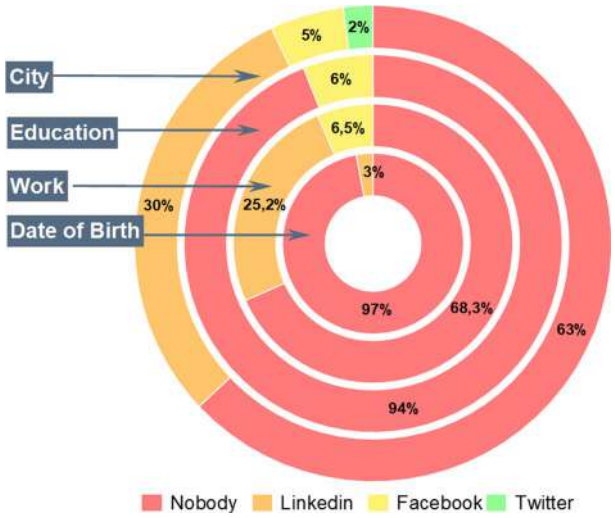

Fig. 15 Attributes reconstruction by exploring all analyzed social except for VKontakte

we show the percentage of information privatised on VKontakte, but published on other social networks:

- 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 are not registered to 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 to 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 to 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 to 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 to 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 to 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.

Table 2 Cross social features extrapolation

	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

Table 3 User's profile information obtained after cross-social analysis

	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.

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 2 summarises the additional information gathered by performing a cross-social analysis over each analysed social network. In particular, for each social networks (rows in Table 2), we have highlighted the additional information retrieved from other ones (columns in Table 2). Of course, the diagonal reports similar information presented in Table 1. As we can notice in Table 2, Facebook, Twitter and Vkontakte permit us to retrieve beneficial information concerning users for creating a more detailed user's profile.

Finally, in Table 3, shows a final overview of the user profile information collected through cross-social analysis. We highlight 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.[6] Data reported in Table 3, singularly are not sensitive for GDPR, but their aggregation permit us to identify a specific user putting at risk his/her privacy.

[6] 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.

Ethical discussion

Social networks represent a vast information source in terms of data. However, processing and analyse data gathered by social networks could raise ethical discussions. In this section, we aim to explain the ethical reasons linked 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 disclosure of personal data in the context of scientific research purposes"* (recital 159). The GDPR defines some other bases so as 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, we have collected social networks users data to perform a specific analysis with the only aim to improve user's awareness concerning privacy threats over different social networks. Our 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, we have not violated user's privacy because this was not our target; we have collected data with the only purpose to emphasise privacy issues related to social network data dissemination. We justify the ethical aspects of our 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

Conclusion and future directions

Guaranteeing privacy, especially over social networks, it is an intrinsic problem of social networks themselves, especially since their goal is to enable users to safely share information. In most cases, social networks users are not familiar with privacy preservation issues. Yet, in a context in which data are becoming a valuable asset, it arises the necessity to develop new methodologies for helping users in understanding issues related to mismanagement of their data.

In our work, we have performed a single-social and a cross-social evaluation concerning users' data, to assess how easily they can be reconstructed from social networks. Our results highlight that it is possible to obtain characterising user's information by analysing the profile of a user over multiple platforms. Moreover, through the cross-social analysis, we also reconstructed other significant users data by exploiting the

combination of several social networks. We want to clarify that our analysis aims not to violate users' privacy over their social network accounts but only emphasise that users do not understand in deeply privacy threats linked to an incorrect sharing/privatisation of their data. Yet, GDPR needs to be extended to define guidelines helping users being aware of "social network privacy threats". We performed this type of analysis to improve the user's perception concerning privacy threats.

In the future, we would like to collect more data concerning users, by integrating information over other social networks. Additionally, we would like to improve our analysis by using semantic analysis to extrapolate information contained in the biography of users, and enhance the face recognition phase in terms of accuracy. Finally, we would also like to investigate the possibility to retrieve information contained within images of users, by exploiting text recognition for gathering data.

Abbreviations
GDPR: General data protection regulation; PDS: Privacy disclosure score; SODA: Social data analyzer; API: Application programming interface.

Acknowledgements
We thank Jacob Wilkin, the author of the Social Mapper tool.

Authors' contributions
GP: methodology, Writing- Original draft preparation; DD: methodology, Writing- Original draft preparation; SC: methodology, Software, Validation, Writing- Original draft preparation; FC: developing and testing of the prototype; SMG: developing and testing of the prototype. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data used to support the findings of this study is available in an anonymized version from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 5 May 2021 Accepted: 26 January 2022
Published online: 14 February 2022

References

1. Breve B, Caruccio L, Cirillo S, Desiato D, Deufemia V, Polese G. Enhancing user awareness during internet browsing. In: ITASEC, 2020;pp. 71–81.
2. Cirillo S, Desiato D, Breve B. Chravat-chronology awareness visual analytic tool. In: 2019 23rd International Conference Information Visualisation (IV), 2019;pp. 255–260. IEEE.
3. García-Sánchez F, Colomo-Palacios R, Valencia-García R. A social-semantic recommender system for advertisements. Inf Proc Manag. 2020;57(2):102153.
4. Choi J, Yoon J, Chung J, Coh B-Y, Lee J-M. Social media analytics and business intelligence research: a systematic review. Inf Proc Manag. 2020;57(6):102279.
5. Desiato D. A methodology for gdpr compliant data processing. In: SEBD 2018.
6. Caruccio L, Desiato D, Polese G, Tortora G. Gdpr compliant information confidentiality preservation in big data processing. IEEE Access. 2020;8:205034–50.
7. European Commission: General Data Protection Regulation—Final version of the Regulation. Released 6 April 2016 2016. http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf.
8. Rao PRM, Krishna SM, Kumar AS. Privacy preservation techniques in big data analytics: a survey. J Big Data. 2018;5(1):1–12.

9. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big Data. 2016;3(1):1–25.
10. Pramanik MI, Lau RY, Hossain MS, Rahoman MM, Debnath SK, Rashed MG, Uddin MZ. Privacy preserving big data analytics: a critical analysis of state-of-the-art. Wiley Interdiscip Rev Data Min Knowl Discov. 2021;11(1):1387.
11. Aghasian E, Garg S, Gao L, Yu S, Montgomery J. Scoring users' privacy disclosure across multiple online social networks. IEEE Access. 2017;5:13118–30.
12. Bhagat S, Saminathan K, Agarwal A, Dowsley R, De Cock M, Nascimento A. Privacy-Preserving User Profiling with Facebook Likes 2018.
13. Chakraborty R, Vishik C, Rao HR. Privacy preserving actions of older adults on social media: exploring the behavior of opting out of information sharing. Decis Support Syst. 2013;55(4):948–56.
14. Zheleva E, Getoor L. Privacy in social networks: a survey. In: Social Network Data Analytics, pp. 277–306. Springer, 2011.
15. Sun C, Philip SY, Kong X, Fu Y. Privacy preserving social network publication against mutual friend attacks 2013.
16. Dakiche N, Tayeb FB-S, Slimani Y, Benatchba K. Tracking community evolution in social networks: a survey. Inf Proc Manag. 2019;56(3):1084–102.
17. Li K, Cheng L, Teng C-I. Voluntary sharing and mandatory provision: private information disclosure on social networking sites. Inf Proc Manag. 2020;57(1):102128.
18. Blosser G, Zhan J. Privacy preserving collaborative social network. 2008, pp. 543–8, IEEE.
19. He Z, Cai Z, Yu J. Latent-data privacy preserving with customized data utility for social network data. IEEE Trans Veh Technol. 2017;67(1):665–73.
20. Aggarwal CC. An introduction to social network data analytics, 2011;1–15.
21. Crossley N. The social world of the network. combining qualitative and quantitative elements in social network analysis. Sociologica 2010;4(1).
22. Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Comput. 2013;17(5):62–9.
23. Scott J. Social network analysis. Sociology. 1988;22(1):109–27.
24. Tichy NM, Tushman ML, Fombrun C. Social network analysis for organizations. Acad Manag Rev. 1979;4(4):507–19.
25. Balaji T, Annavarapu CSR, Bablani A. Machine learning algorithms for social media analysis: a survey. Comput Sci Rev. 2021;40:100395.
26. Aljably R, Tian Y, Al-Rodhaan M. Preserving privacy in multimedia social networks using machine learning anomaly detection. Security and Communication Networks 2020; 2020.
27. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst. 2018;6(1):1–12.
28. Al-Molhem NR, Rahal Y, Dakkak M. Social network analysis in telecom data. J Big Data. 2019;6(1):1–17.
29. Jafri R, Arabnia H. A survey of face recognition techniques. JIPS. 2009;5:41–68.
30. Wang Y-Q. An analysis of the viola-jones face detection algorithm. Image Proc Line. 2014;4:128–48.
31. Sharifara A, Mohd Rahim MS, Anisi Y. A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection, 2014;73–78.
32. Adikari S, Dutta K. Identifying fake profiles in linkedin. 2020, arXiv preprint arXiv:2006.01381.
33. Admin View on LinkedIn Pages. https://www.linkedin.com/help/linkedin/answer/98738/use-your-admin-view-on-linkedin-pages. Accessed: 2021-08-25.
34. Punkamol D, Marukatat R. Detection of account cloning in online social networks. In: 2020 8th International Electrical Engineering Congress (iEECON), 2020;pp. 1–4. IEEE.
35. Bródka P, Sobas M, Johnson H. Profile cloning detection in social networks. In: 2014 European Network Intelligence Conference, 2014;pp. 63–68. IEEE.
36. Krishnapriya K, Albiero V, Vangara K, King MC, Bowyer KW. Issues related to face recognition accuracy varying based on race and skin tone. IEEE Trans Technol Soc. 2020;1(1):8–20.
37. Adjabi I, Ouahabi A, Benzaoui A, Taleb-Ahmed A. Past, present, and future of face recognition: a review. Electronics. 2020;9(8):1188.

Publisher's Note