Cross-Social Network Investigation to Highlight Privacy Violations in Data Sharing Activities

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

University of Salerno, Department of Computer Science, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Abstract

Social networks represent a vast source of information and have an increasing impact on people's daily lives. In fact, they permit to exhibit users' lives, share emotions, passions, and interactions with other users around the world. These data need to be monitored because they could produce privacy violations, especially when they involve sensitive information. In this scenario, the definitions of privacy policies for safeguarding users' data represent a difficult challenge that social networks have to deal with. In fact, although social network platforms offer privacy settings to protect data, often, users are unable to properly manage them to safeguard their privacy. To this end, in this work, we present a statistical investigation concerning privacy policies offered by social network platforms. In particular, we have defined a tool relying on image-recognition techniques capable of exploring social network platforms and identifying user profiles starting from their pictures. Moreover, we have composed a dataset of 5000 users by retrieving their data available over different social network platforms in order to compare publicly accessible data provided in the registration phases, and those retrieved by our analysis. The proposed work underlines privacy violations over social network platforms when privacy policies are not managed correctly, and is targeted to improve the users' awareness concerning the spreading and managing of their data. We have highlighted all the statistical evaluations made over the gathered data for putting in evidence the privacy issues.

Keywords

Privacy, Social Networks, Data Analysis

1. Introduction

Plenty of people are registered over several social networks, sharing a vast amount of information. Moreover, social networks play a fundamental role in human interactions since they permit people to share emotions, ways of thinking, points of view, and so on. In this scenario, preserving users' privacy is crucial for social network platforms since they cannot permit the jeopardization of users' data.

Users tend to use social networks to share information massively, and in most cases, they do not care about privatizing data and are unaware of the threats they can be exposed to. Moreover, the growing number of users signing up on these platforms yields the necessity of analyzing how they manage their privacy, mainly when using multiple social networks.

Several studies have discussed data privacy issues on social networks [1, 2, 3, 4], but only

some of them have provided tools capable of improving users' awareness when sharing their data on social networking platforms [5, 6, 7]. In our work, we perform cross social network analysis on five social network platforms to figure out which is the information that is most frequently shared over social networks, and that can jeopardize user's privacy [8, 9].

In this discussion paper, we describe the tool SOcial Data Analyzer (SODA) proposed in [10], which is able to find and extract available information of users on different platforms considering only their photos. SODA allowed us to perform an accurate analysis for revealing privacy threats linked to incorrect usage of data sharing in social networks. Moreover, the tool allowed us to evaluate the sensitiveness of information shared by users and perform an exhaustive analysis to understand how social networks can reconstruct users' data even if some of them are protected on other platforms.

SODA is independent of the privacy settings offered by social networks since it simulates the search of a real user and retrieves data publicly available in social network profiles. In other words, if a user has privatized specific information over a specific social network, SODA cannot retrieve that information. However, if the user has the same information not privatized over a different social network, the SODA retrieves such information. Thus, we could say that (SODA) tests the users' skills in managing privacy settings offered by social network platforms.

In summary, the main contributions of our study are $i)$ a new tool capable of finding users and extracting their data from different social network platforms, and $ii)$ a detailed analysis of users' data extracted from different social networks aiming to evaluate their privacy and improve their awareness concerning privacy threats in social network platforms.

The paper is organized as follows, in Section 2 we present our methodology, whereas Section 3 presents the architecture of SODA tool. In Sections 4 and 5 we describe the results of our analysis. Finally, conclusions and future research directions are discussed in Section 6.

2. 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, the picture and the name of the users are exploited as the input of the Social Module. Then, for each social network platform, the module performs specific operations for scraping the content of the pages and extracting photos and names associated with each profile. This information is given as input to a face recognition module that tries to match the discovered photos and the initial user's picture. If a match is found, the social network module extracts all data available on a user profile.

Similar to the single social step, the cross-social data extrapolation step starts by considering the picture and the name of a user for searching a user from multiple platforms. However, the main difference is the exploitation of multiple social network modules for extrapolating several user profile data. In particular, each module extracts user profile data from each social network platform according to its representation of the data. 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, which is responsible for aggregating all the collected data.

The differentiation of the single- and cross-social analysis allows us to estimate the minimum amount of user data that is possible to extract from each social network and evaluate the maximum number of data that can be aggregated from different platforms. In the following section, we first provide an overview of the new SODA system, and then we describe the architecture underlying it.

3. Social Data Analyzer

Extracting user data from multiple social networks is a complex problem. There are several issues related to extraction yielding specific choices for the components of the 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 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[1] , extending several of its components aiming to tackle the issues mentioned above.

In particular, Social Mapper is capable to search user profiles on multiple social network platforms such as Facebook, Linkedin, Instagram, VKontakte, Twitter, Pinterest, Weibo, and Douban. Because SODA is an extension of Social Mapper, it can search people by only considering an image and at least one information including name, surname, city, email, or the company in which the user works. Starting from these, SODA exploits the Selenium framework for creating a bot capable of automatically browsing web pages, by simulating the behaviors of a real user during a web browsing session. In this way, SODA can exploit the search engines behind each social network platform, by performing searching operations by means of the search bars provided by each platform.

With respect to Social Mapper, SODA provides several novel functionalities enabling to perform an in-depth analysis of the data shared by users, and extend the search on a large scale. The first new functionality allows SODA to find people working for a specific company, by exploiting the search mechanism of Linkedin for selecting users that work in a given company. As demonstrated in [11] the amount of fake users registered on Linkedin is very small, which allowed us to create a dataset with a large number of real users. Most of the remaining extensions concern the crawling components. In fact, Social Mapper is limited to only extracting the URLs of the different user profiles. Thus, in SODA, we have re-designed the crawler modules aiming to add several new navigation features capable of adapting to the different structures of the web pages. The combination of these strategies with a powerful recognition algorithm, i.e., Viola-Jones [12], allows SODA to achieve accurate results on multiple platforms. It is important to notice that, the face recognition algorithm return as output a user profile 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. Figure 1 shows the architecture of SODA. 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

---

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

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 analyzed. 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.

Figure 1: Architecture of SODA.

4. Evaluation single-social

The single social evaluation allowed us to highlight the information that is frequently shared by users over every single social network and analyze how each social network preserves user privacy. To perform this type of analysis, we have created a dataset of 5000 users, by considering people working for a specific company that has been extracted by means of a new feature of SODA enabling search people working for a specific company. Starting from the 5000 users contained in our dataset, we perform a single social network evaluation with the aim to independently analyze the results obtained by each social network, avoiding considering whether a user is present on multiple platforms, which will be discussed in Section 5. It is important to note that the number of people evaluated for each social network corresponds to those who were actually found by SODA after exploring the different platforms.

Concerning Linkedin, 1570 user profiles have been evaluated. In particular, *Employment* and the *City* are the most frequently shared information on Linkedin. More specifically, the attribute city can refer to the place of residence or the place of birth, but in most cases, these are equal. Results 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.

Concerning Facebook, 1161 user profiles have been evaluated. In particular, the 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. However, 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 date of birth in order to preserve privacy.

Concerning Twitter, 86 user profiles have been evaluated. Despite not many users involved in the analysis, 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.

Concerning VKontakte, 251 user profiles have been evaluated. In particular, the *Date of birth*, the *Spoken languages*, and the *Education* information are the most frequently shared data on this social network. More specifically, not 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. 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 helpful in 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.

5. Evaluation cross-social

In this section, we describe the statistics derived by performing a cross-social analysis of the publicly accessible information extracted from available social networks, and we investigate the possibility of aggregating them aiming to perform a more detailed analysis.

Figure 2 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.
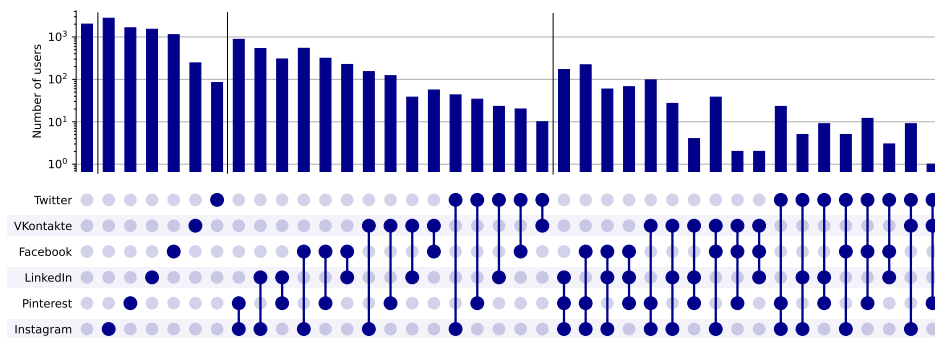
Figure 2: Distribution diagram of the analyzed users.

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 privatize some information on a specific social network, where s/he can choose to unmask the same information

over other social networks. It means that it is possible to obtain more detailed information by analyzing a specific user over different social networks.

The most frequently accessed information on Twitter is the city since it can be reconstructed through other social networks. In particular, 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. We detail the percentage of information privatized by Facebook users but published on other social networks:

- 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 privatized 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 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 on 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.

Finally, most of the analyzed users who have privatized a given data on Facebook have also privatized it on other social networks. Among all considered social networks, Linkedin has proved to be helpful for the reconstruction of users' information.

The information that are most frequently accessible on Linkedin are *Mobile phone*, *City*, *Date of birth*, *Email*, and *Employment*. We detail the percentage of information privatised on Linkedin, but published on other social networks:

- Similarly to Facebook, 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 privatized 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.

Finally, most of the analyzed users who have privatized a given data on Linkedin have also privatized it on other social networks. Furthermore, among all considered social networks, Facebook has proven to be helpful for the reconstruction of users' information.

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

- Similarly to the previous analysis, among the 5000 analyzed users who have privatized their *Mobile phone* 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 analyzed users shared it on other social networks, so 97% have completely privatized 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 analyzed users shared it on other social networks, so 99% have completely privatized 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 analyzed users shared it on other social networks, so 94% have completely privatized 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 analyzed users shared it on other social networks, so 68.3% have completely privatized it.

Finally, most of the analyzed users who have privatized a given data on VKontakte have also privatized 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 helpful for the reconstruction of users' information.

6. Conclusion and Future directions

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 characterizing user's information by analyzing his/her profile over multiple platforms. Moreover, through the cross-social analysis, we also reconstructed other significant users' data by exploiting the combination of several social networks.

In the future, we would like to collect more data concerning users by integrating information over other social networks. Finally, we would also like to investigate the possibility of retrieving information contained within users' images by exploiting text recognition for gathering data.

References

[1] P. R. M. Rao, S. M. Krishna, A. S. Kumar, Privacy preservation techniques in big data analytics: a survey, Journal of Big Data 5 (2018) 1–12.

[2] P. Jain, M. Gyanchandani, N. Khare, Big data privacy: a technological perspective and review, Journal of Big Data 3 (2016) 1–25.

[3] M. I. Pramanik, R. Y. Lau, M. S. Hossain, M. M. Rahoman, S. K. Debnath, M. G. Rashed, M. Z. Uddin, Privacy preserving big data analytics: A critical analysis of state-of-the-art, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1387.

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

[5] S. Cirillo, D. Desiato, B. Breve, Chravat-chronology awareness visual analytic tool, in: 2019 23rd International Conference Information Visualisation (IV), IEEE, 2019, pp. 255–260.

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

[7] G. Bonifazi, E. Corradini, D. Ursino, L. Virgili, A social network analysis–based approach to investigate user behaviour during a cryptocurrency speculative bubble, Journal of Information Science (2021) 01655515211047428.

[8] D. Desiato, G. Tortora, A methodology for gdpr compliant data processing., in: SEBD, volume 2161, 2018.

[9] L. Caruccio, D. Desiato, G. Polese, G. Tortora, Gdpr compliant information confidentiality preservation in big data processing, IEEE Access 8 (2020) 205034–205050.

[10] F. Cerruto, S. Cirillo, D. Desiato, S. M. Gambardella, G. Polese, Social network data analysis to highlight privacy threats in sharing data, Journal of Big Data 9 (2022) 1–26.

[11] S. Adikari, K. Dutta, Identifying fake profiles in linkedin, arXiv preprint arXiv:2006.01381 (2020).

[12] Y.-Q. Wang, An analysis of the viola-jones face detection algorithm, Image Processing On Line 4 (2014) 128–148.