HAL
open science

Automated Risk Analysis on Privacy in Social Networks

Younes Abid

To cite this version:

Younes Abid. Automated Risk Analysis on Privacy in Social Networks. Social and Information Networks [cs.SI]. Université de Lorraine, 2018. English. NNT : 2018LORR0088 . tel-01863354

HAL Id: tel-01863354

https://theses.hal.science/tel-01863354v1

Submitted on 28 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE LORRAINE

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

École doctorale IAEM Lorraine

Analyse automatisée des risques sur la vie privée dans les réseaux sociaux

THÈSE

présentée et soutenue publiquement le 05/07/2018

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Younes ABID

Composition du jury

Rapporteurs :	Mme BENTAYEB Fadila	Maître de conférences - Université Lumière Lyon 2
	Mme BERRUT Catherine	Professeur - Université Grenoble Alpes
Examinateurs :	M. COQUERY Emmanuel	Maître de conférences - Université Claude Bernard Lyon 1
	M. IMINE Abdessamad	Maître de conférences - Université de Lorraine
	M. RUSINOWITCH Michaël	DR - INRIA - Directeur de thèse
	M. SMAILI Kamel	Professeur - Université de Lorraine
Invité :	M. RIGOLOT Marc	Directeur de la fondation MAIF

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

> Lorsque vous résolvez un problème, vous devriez remercier Dieu et passer au problème suivant.
>
> *Dean Rusk*

J'aimerais tout d'abord remercier mon directeur de thèse, Michaël Rusinowitch, pour sa supervision pendant les trois ans que j'ai passés dans l'équipe Pesto Inria Nancy-Grand Est et Loria et pour sa lecture méticuleuse de chacun des chapitres. Je remercie également Abdessamad Imine pour ses critiques constructives ainsi que tous les membres de l'équipe Pesto. Je tiens aussi à remercier tous les chercheurs et assistants de la DRH du Loria et Inria Nancy-Grand Est.

J'adresse aussi mes remerciements à Marc Rigolot et Jean-Marc Truffet ainsi qu' à la fondation MAIF pour son soutien à mes recherches.

Je tiens à remercier Fadila Bentayeb et Catherine Berrut pour avoir accepté d'être rapporteurs de cette thèse. Je remercie également Emmanuel Coquery, Abdessamad Imine et Kamel Smaili de faire partie de mon jury.

Mes remerciements vont aussi à mes chers parents, mes frères, Khalil et Khereddine, et ma sœur, Olfa, pour leur soutien.

Dedicated to my lovely self,
family, colleagues, friends
and all users of social networks.

Contents

List of Tables **xi**

Chapter 1
Introduction **1**

1.1 Research context 1
1.2 Structure of social networks 6
1.3 Research on social network privacy analysis 9
1.4 Motivations and challenges of this work 14
1.5 Contributions of the thesis 14
1.6 Outline 17

Chapter 2
Defining sensitive subjects

2.1 Introduction 19
2.2 Conducting a survey on the behaviour of French users of social media 20
2.3 Analysing responses and defining sensitive subjects 23
2.4 Possible attack vectors according to the behaviour of participants 32
2.5 Conclusions 36

Chapter 3
Disclosing friendship and group membership links

3.1 Introduction 39
3.2 Modelling social network for on-line link disclosure attacks 40
3.3 Problematics and objectives 40
3.4 Social networks group properties 41
3.5 Link disclosure attacks 44
3.6 Conclusions 51

Chapter 4
Overview of our implemented prediction system

4.1 Introduction 53
4.2 Architecture 54
4.3 SONSAI user guide 56
4.4 Examples of inference scenarios 59
4.5 Conclusions 59

Chapter 5
Sampling and modelling social networks

5.1 Introduction 61
5.2 Definitions 61
5.3 Sampling social network around a target user profile 62
5.4 Modelling discovered links and nodes by graphs 67
5.5 Anonymizing the social network graph models 69
5.6 Conclusions 71

Chapter 6
Cleansing the collected data

6.1 Introduction 73
6.2 Definitions 73
6.3 Motivations for cleansing data 74
6.4 Cleansing the sensitive graph 76
6.5 Cleansing the learning graphs 80
6.6 Cleansing results 85
6.7 Conclusions 92

Chapter 7
Analysing cleansed data and inferring target sensitive values

7.1 Introduction 95
7.2 Translating social attributed network to a text document 96
7.3 Applying Word2Vec to compute node embeddings 98
7.4 Inferring hidden sensitive values of the target user profile 102
7.5 Measuring inference accuracy 104
7.6 Experiments and results 105
7.7 Conclusions 110

Chapter 8
Conclusions and perspectives

Appendixs **113**

Appendix A Datasets **113**
A.1 Dataset 1 (D1) 113
A.2 Dataset 2 (D2) 113

Appendix B Questionnaire **119**

Bibliography **129**

List of Figures

1.1 Number of monthly active user on the top 10 most famous social network as of 1st quarter 2018(from [Tauzin, 2018]). . . . 3
1.2 Worldwide number of monthly active Facebook users from 2008 to 2017 (from [Statista, 2018]). . . . 4

2.1 Distribution of participants by gender and age. . . . 21
2.2 Distribution of participants by region. . . . 22
2.3 Distribution of participants by study discipline. . . . 22
2.4 Distribution of the age of the sample with regard to the mother population. . . . 22
2.5 Participant Net surfers activity rate on forums and websites. . . . 25
2.6 Example of reversed parental monitoring scenario. . . . 33
2.7 Example of scenario of attack of linking profiles across different social media. . . . 33
2.8 Examples of scenarios of sensitive information leakage. . . . 34
2.9 Scenario of leakage of sensitive information between several social networks. . . . 36
2.10 Scenario of leakage of sensitive information on the dame social network. . . . 36

3.1 Social graphs : (a) unipartite friendship graph, (b) bipartite group membership graph. . . . 40
3.2 Group distribution of a sample of 14,517 Facebook users. . . . 42
3.3 Results of analysis: (a) Variation of public density with respect to group declared size, (b) Expected number of disclosed links between the target and group members. 44
3.4 g3 is a real 3-hop distant group from t. . . . 45
3.5 An example of public n-hop distant groups and members. . . . 45
3.6 2-hop friendship disclosure attack. . . . 48
3.7 Undisclosed links by mutual-friend attack. . . . 49
3.8 The average number of undisclosed links by mutual-friend attack but disclosed by friendship attack. . . . 50
3.9 Results of attacks: (a) The average number of friendship request to disclose one friendship link, (b) The average number of mutual-friend request to disclose one friendship link . . . 51
3.10 Sample of 14,517 Facebook profiles: (a) Frequency of published group membership, (b) Frequency of list of friends size. . . . 51

4.1 Architecture of SONSAI. . . . 55
4.2 Collector GUI. . . . 56
4.3 Analyser settings GUI. . . . 57
4.4 Analyser results GUI - left screen. . . . 57
4.5 Analyser results GUI - right screen. . . . 58

5.1 Example of the evolution of the knowledges of the crawler about the social network. 63
5.2 Example of friendship graph. 68
5.3 Example of groups membership graph. 68
5.4 Example of pages like-ship graphs. 69
5.5 Example of relationship status and gender graphs. 70
5.6 Example of TSV files. 71

6.1 Example of clustering the values of the attribute "politician". 77
6.2 Variation of partitioned bipartite sub-graph similarities with respect to the minimal size of sub-graphs, (a) Users-Actors graph: 15k users, 364 actors, (b) Users-FastFoods graph: 15k users, 777 fast foods. 79
6.3 Example of cutting graphs for structure comparison. 80
6.4 Example of transforming bipartite graphs into weighted unipartite graphs. 82
6.5 Splitting graphs for comparing them. 84
6.6 Example of creating a dense and discriminant graph for gender inference. 85
6.7 The distribution of the learning graph with regard to the sensitive graph Users-Politicians. 87
6.8 The distribution of the 23 selected relevant graphs with regard to the sensitive graph Users-Politicians. 87
6.9 The distribution of the learning graph with regard to the sensitive graph Users-Genders. 89
6.10 The distribution of the 8 selected discriminant graph with regard to the sensitive graph Users-Genders. 90
6.11 Distribution of the learning graph w.r.t. the sensitive graph Users - RelationshipsStatus. 91

7.1 Example of multi graph random walk. 96
7.2 The skip-gram model [Mikolov et al., 2013a]. 99
7.3 The CBOW model [Mikolov et al., 2013a]. 100
7.4 The neural network of Word2Vec. 101
7.5 Example of 2-dimensional vectors that encode 8 nodes. 102

A.1 Frequencies of published attributes per user in dataset 1. 114
A.2 Friendship graph of dataset 2. 116
A.3 Frequencies of published attributes per user in dataset 2. 117

List of Tables

2.1 Distribution of responses by use of social media. 20
2.2 Margins of error of the studied sample. 23
2.3 Subjects ranked by increasing order of discussions on social networks. 25
2.4 Subjects ranked by increasing order of discussions on forums and websites. 26
2.5 Subjects ranked by decreasing rates of anonymous publication on forums and websites. 27
2.6 Subjects ranked by decreasing order of avoidance on social networks. 27
2.7 Notations. 28
2.8 Sensitive subjects. 29
2.9 Descending order of sensitive subjects. 31

4.1 Samples of values of attributes of the targets u_1 and u_2. 60

6.1 Details about the graphs used in experiment 1. 75
6.2 Details about the graphs used in experiment 2. 76
6.3 Details about the graphs directly related to politics in the dataset. 86
6.4 Comparison of 3 graphs directly related to politics with the politicians graph. . . 86
6.5 Distribution of the learning graphs parameters w.r.t. the sensitive graph Politicians. 86
6.6 Details about the distribution of the 23 selected graphs with regard to the sensitive graph Users-Politicians. 88
6.7 Statistic details about the distribution of all the learning graphs with regard to the sensitive graph Users-Genders. 89
6.8 Details about the distribution of the 8 selected discriminant graphs with regard to the sensitive graph Users-Genders. 90
6.9 Distribution of relationship status of user that publish their status in the dataset. 91
6.10 Distribution of learning graph parameters w.r.t. the sensitive graph Users - RelationshipsStatus. 92
6.11 Distribution of the selected discriminant graphs with regard to the sensitive graph Users-RelationshipsStatus. 93

7.1 Experimental results. 106
7.2 Relationship status of users in the datasets. 106
7.3 Selected attributes in D2 for relationship status inference. 107
7.4 Genders of users in the datasets. 107
7.5 Selected attributes in D1 for gender inference. 108
7.6 Processing times. 108
7.7 Impact of lr on inference accuracy. 109
7.8 Impact of hr on inference accuracy. 109

7.9 Impact of *cr* on inference accuracy. 109

A.1 Details about the dataset 1. 113
A.2 Details about the published attributes of crawled profiles in dataset 1. 115
A.3 Details about the dataset 2. 115
A.4 Details about the published attributes of crawled profiles in dataset 2. 115

1

Introduction

All human beings have three lives:
public, private and secret.

Gabriel García Márquez: a Life

Contents

1.1 Research context **1**
1.2 Structure of social networks **6**
1.3 Research on social network privacy analysis **9**
1.4 Motivations and challenges of this work **14**
1.5 Contributions of the thesis **14**
1.6 Outline **17**

1.1 Research context

The appearance of the first on-line social networks in the late 1990s was marked by the use of *avatars*. Interactions on on-line social networks like Classmates.com (1995) or OpenDiary.com (1997) were very limited. Moreover, Net surfers were careful not to publish sensitive information and not to communicate their ages, addresses or real names on the sites of *chats*. Today on-line social networks are full of personal photos and sensitive information spontaneously published by users. Even more, on-line social networks encourage users to provide their personal data in order to help them find new friends and enjoy rich social experiences. Link recommendation is a critical functionality for on-line social networks. It allows the network to evolve by increasing linkage and attracting new users. In [Yin et al., 2010], the authors demonstrate the importance of attributes in designing an effective link recommendation system for social networks. In [Barbieri et al., 2014], the authors distinguish between two types of links between users: (i) social links and (ii) topical links. Social links are links between users sharing several common friends. However, topical links are links between users sharing same interest toward common topics. They investigate both types of links to design an accurate link prediction system with topical explanation.

A social network can refer to on-line social network [van Schaik et al., 2018], opportunistic social network [Zakharya and Benslimane, 2018], scientific social network [Gimenes et al., 2014],

consumer social network [Leskovec et al., 2007], etc. In this work, we only analyse on-line social networks that we shortly refer to by social networks in the following.

While social networks are getting larger every day, new challenging big data and networking problems emerge. Moreover, privacy issues are becoming more difficult to resolve. Social networks, on the other hand, pay more attentions to technical problems and making profits than resolving privacy issues. To be free from any problems that can trigger privacy incidents, they make sure to add articles in the user charts to put all the responsibility on the end user in case of privacy leakage.

In this section, we shed the light on the importance of personal information in the business of social network and we discuss several aspects that have direct impact on the privacy of users.

Clumsy users' behaviours. In the absence of a clear definition of sensitive information and its impact on real life, Net surfers tend to seek popularity by collecting as many interactions as possible on social networks. In a frantic pursuit of fame, they reveal their personal information in the hope of catching the attention of the world. As a consequence, a simple search on Google is enough to reveal their most intimate information, to guide recruiters in their choice of the best candidate or to put an end to their careers.

Attempts to safeguard users' privacy. To help protect the unaware or clumsy social network users, Google dereferences sensitive information if requested for it. However, the French National Commission of Computing and Freedoms (CNIL) judged that the measures taken by Google were not sufficient and was fined 100,000 euros in March 2016 for not dereferencing sensitive information concerning French citizens on non European versions of the search engine (such as google.com) [CNIL, 2016]. It is then worth mentioning that the dereference of the information does not remove it from the web. It only limits the access to it by removing the links toward web pages holding sensitive information when displaying the result of a search by the name of the person in question.

On their side, social networks provide several solutions in order to safeguard the privacy of users [Guo and Chen, 2012]. However, their main deficiencies are related to complicated, non-uniform, periodically updated and unintelligible privacy policies, long and ambiguous user charters, and non-ergonomic privacy management interfaces. Although most of the social networks offer similar services (creating profiles, pages and groups, establishing links and interactions), their visibility management and the definition of links (symmetrical, non-symmetrical) are different. These design differences can be confusing for users of multiple social networks that do not take care of checking the settings of each network. In addition, the default choice of these parameters favours public display and increases the risk of sensitive information leakage.

Fast and vast spread of information. With the exponential growth of social network, information spreads rapidly between users and leaves no time for rectifications and takeover. It is hard to erase completely the traces of shared, downloaded or re-published data. They can hibernate on servers or terminals of users who downloaded them and reappear later to start a new life cycle. Facebook stressed the fact that "*When you delete intellectual property content, it is deleted in a manner similar to emptying the recycle bin on a computer*[Facebook, 2015]". Moreover, URLs pointing to contents may continue to exist on the Content Delivery Network (CDN) and can be downloaded through hotlinks even after their deletion by users [Whittaker, 2010].

In [Dow et al., 2013] authors study the cascades of reshares that generate some photos on Facebook. They shows that photos can virally spread on the network within a few hours even if

the original publisher is not a hub. For instance, Petter Kverneng had an important life lesson after sharing a joke photo on Facebook in 2013. He was on the photo with Catherine and held a sign saying that she will have sex with him if he gets one million likes. The photo rapidly reached more than 1 million likes in a few hours. He still keeps on receiving comments on his profile asking if she honoured her promise up to now.

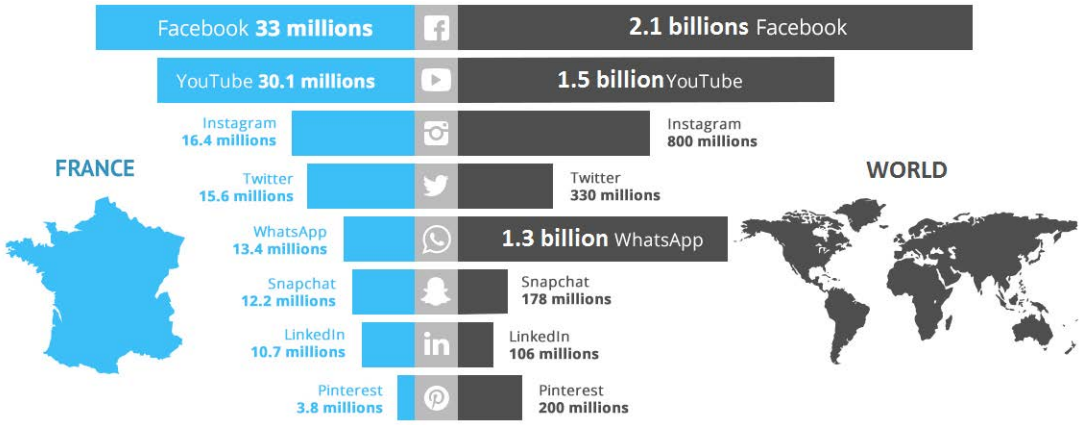


Figure 1.1: Number of monthly active user on the top 10 most famous social network as of 1st quarter 2018(from [Tauzin, 2018]).

Figure 1.1 details the number of monthly active users on the top 10 most famous social network in France and the world as of 1st quarter of 2018 [Tauzin, 2018]. Monthly active users are those which have logged in to the social network in question during the last month. Statistics show that about 28% of humans and half of french citizens are active on Facebook. Facebook is the first social network that surpassed 1 billion monthly active users in the third quarter of 2012. Figure 1.2 depicts the evolution of the number of Facebook users between 2008 and 2017 [Statista, 2018]. The authors of [Bhagat et al., 2016] show that the separation degree on Facebook is between 2.9 and 4.2 for the majority of users. Which means that every information on this network can reach all users (28% of human) after only 3 hops of shares. Similarly, the average separation degree between Twitter users is 3.43 [Bakhshandeh et al., 2011].

Profitability of personal data. Only a decade after its creation in 2004, Facebook hit one billion active users and gained the title of the largest social network. It then had to daily handle about 350 million photo uploads, 4.5 billion likes and 10 billion messages. Moreover, about 100 petabytes of analytic data and 500 terabytes of new data are generated every day [Feinleib, 2014]. Today Facebook counts more than 2 billion monthly active users and the challenge is increasing. To deal with the problem of big data, Facebook allows users to publish up to 5 posts per day as soft limit and 25 posts as upper limit before decreasing the reach of posts. Most social networks define daily posting limit as well. For instance, users cannot publish more than 730 tweets over 24 hour period. The daily posting limit on LinkedIn, Google+, Pinterest and Instagram is 100 [buffer.com, 2018].

Despite the challenge of processing it, this huge amount of data is indeed a great asset for social networks as it guarantees a large profit. Every second, Facebook is getting richer by getting permission to use any intellectual property posted on it. The permission ends only if the content is deleted by all users who shared it [Facebook, 2015]. Besides, social network are making large profit through advertising. Users enjoy free access to most of the social network services.

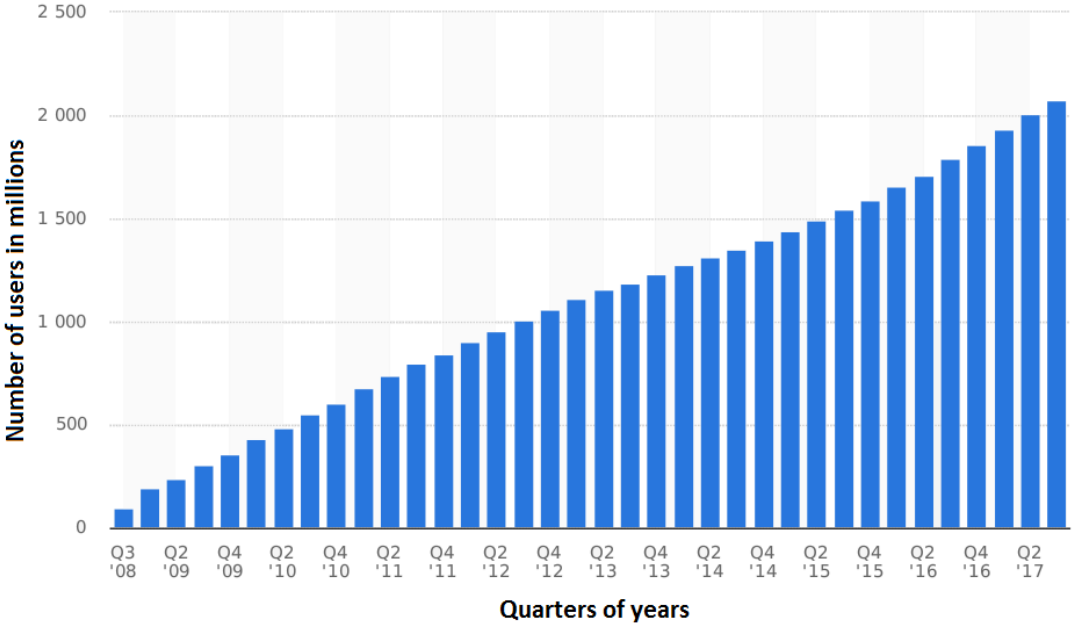


Figure 1.2: Worldwide number of monthly active Facebook users from 2008 to 2017 (from [Statista, 2018]).

However, by spontaneously revealing their personal data they become themselves the product. Based on their personal information such as age and location, social networks define several targeting options for advertisers. Today, Facebook counts over 5 million monthly advertisers and 2.1 billion monthly active users [Ingram, 2017]. The average Click Through Rate (CTR) across all industries in Facebook ads is 0.9%, where CTR is the ratio of users who click on an advertisement to the number of all users who view it. The average Cost Per Click (CPC) across all industries in Facebook ads is 1.72$ [Irvine, 2017] where CPC is the money paid by an advertiser to Facebook each time the ad is clicked. Consequently, new research problems emerged such as which set of users should be targeted or how to reduce advertising budget while maximizing influence [Zhang et al., 2016, Zhan et al., 2016].

Correlation between information. Knowledge accumulated by social networks about users goes beyond what is published. In an interview with Atlantic's journalist in 2010, former Google CEO Eric Schmidt said "*We don't need you to type at all. We know where you are. We know where you've been. We can more or less know what you're thinking about.*" [Thompson, 2010]. In [Griffith, 2008, Griffith, 2010] Virgil Griffith has proven high correlations between the education level of American Facebook users and their favourite musics and books. He crawled Facebook profiles and collected three attributes: (i) favourite musics, (ii) favourite books and (iii) colleges. Based on the SAT/ACT score of college, he shed the light on the correlation between intellectual milieu of colleges, musics and books. In [Mason, 2014], Winter Mason has mined the cultural similarities between American Facebook users and their political view (Democrats or Republicans). He sampled profiles of Facebook users who liked the campaign pages of some Democrat or Republican politicians. Then, he collected their lists of liked pages. Finally, he statistically identified the types of the pages that are most disproportionately liked by the supporters of one political view versus the other. The results of his work show that politics is highly correlated to

musicians, landmarks, authors, books and TV shows.

Revealing and exploiting such correlation is the cornerstone for designing accurate recommendation systems [Ricci et al., 2011]. We distinguish three categories of recommendation systems: (i) collaborative filtering [Elahi et al., 2016], (ii) content-based filtering [Son and Kim, 2017] and (iii) hybrid recommendation systems [Adomavicius and Tuzhilin, 2005].

Collaborative filtering systems collect information concerning a target user (for instance, his favourite musics and books). Then, they seek for users that like the same values in order to recommend their other values to the target (for instance, their favourite politicians). They are based on the assumption that users who share some common values of attributes are more likely to have similar taste concerning more attributes. On the other hand, content-based filtering systems require an accurate understanding of the attributes in order to generate textual description. Based on the textual description of the liked values of the target, the system will recommend the most similar values to them. Hybrid recommendation systems combine both collaborative filtering systems and content-based filtering systems. Either they perform sequential prediction based on each technique separately then combine the results or perform parallel prediction in one unified model.

It is then worth mentioning that recommendation systems represent a real threat to privacy as they can reveal sensitive information such as political view and ethnicity. However, they are not tremendous bulletproof and have many limitations. For instance, one typical issue of those systems is the cold start problem that is highly related to the data sparsity. When users do not provide sufficient attribute values a system cannot provide reliable recommendations.

To cope with those problems and improve their recommendation systems, many companies release their valuable data. However, before publishing their data they sanitize it. The main objectives of the sanitization is to decrease the accuracy of linking a particular information to a real person [Sweeney, 2002] and hide sensitive patterns [Telikani and Shahbahrami, 2018]. In order not to alter the utility of the data, sanitization tasks must be carefully carried on. In [Edgar, 2004], the author details the most used sanitization techniques.: (i) NUL'ing Out technique consists of deleting column from the dataset. (ii) Masking Data consists of deleting some characters from string values. (iii) Substitution consists of replacing values by randomly generated values. (iv) Shuffling Records consists of randomly switching between values in the same column. (v) Number Variance consists of varying numerical data within a defined range. (vi) Gibberish Generation consists of replacing texts by random ones while preserving the same statistical distribution of words or characters. (vii) Encryption/Decryption consists of encrypting the data and managing access to it. Those techniques are then used by several anonymization models such as k-anonymity [Sweeney, 2002, Singh et al., 2016], l-diversity [Machanavajjhala et al., 2007], t-closeness [Li et al., 2007] ...

The k-anonymity property ensures that each real person "sampled in the dataset" must be linked to at least $k-1$ possible records in the dataset. In addition to k-anonymity criteria, in l-diversity model, each real person "sampled in the dataset" must be linked to at least l possible values of each sensitive attribute. Besides, in t-closeness model, the distribution of the values of a given sensitive attribute A, that can be linked to each real person "sampled in the dataset", must be close by no more than a threshold t to its distribution in all the dataset.

Despite all the precautions taken, some privacy incidents may occur. Netflix released in October 2006 anonymized data concerning 100 million movie rates made by 500 thousand users [Bennett and Lanning, 2007]. Netflix said that user IDs on the released data cannot be de-anonymized because the data represents only $\frac{1}{8}$ of their records in 2005. They only kept movies titles, ratings and dates and anonymized user IDs . However, in [Narayanan and Shmatikov, 2006], the authors have successfully de-anonymized many user IDs from the released Netflix dataset by

using auxiliary information from the Internet Movie Database (IMDb).

In the same year, 2006, AOL released twenty million search keywords typed by 650 thousand users in 3 months [Arrington, 2006]. AOL deleted the dataset only 3 days after its release. However, the data was mirrored and is still available for download today. AOL removed user IDs from the dataset. However, by mining the searches conducted by users and their topics, the New York Times magazine [Barbaro and Zeller, 2006] managed to de-anonymized the IDs of several users by cross-linking information.

Impacts of disclosing personal information. Personal information if revealed may have serious consequences on clumsy users. They can be exploited to carry out personalized spam attacks [Garrett Brown and Borders, 2008], identity theft attacks [Bilge et al., 2009, Conti et al., 2012], cloning attacks [Kontaxis et al., 2011], Sybil attacks [Kayes and Iamnitchi, 2015], etc. They can cause serious damages to companies [Tanimoto et al., 2015, Shullich, 2012] such as degradation of reputation, malware attacks, copyright infringement, loss of intellectual property ...

Today there are places reserved for future social networks users, even before their birth, through information published by their relatives. A picture of pregnancy deemed to be unresponsive by a future mother may become sensitive tomorrow for the newborn. Indeed, a simple medical consultation on health forums, such as Doctissimo, can have consequences on the baby's future career. According to a survey conducted by Consumer Reports National Research Center in 2010 [Tapellini, 2010], about half of American users of social networks have published personal information exposing them to attacks. About 26% of parents published photos and names of their children on Facebook. About 23% of users do not pay attention to privacy settings. Furthermore, users do not pay attention to their geolocation coordinates when posting on social networks. Based on such information, burglars can plot a break-in when users are not at home. Moreover, recent studies [Ge et al., 2014] show that more than half of Pengyou users, one of the most popular social networks in China, have published personal information exposing them to attacks.

This demonstrates the imperative need to design applications to detect and minimize the dissemination and exploitation of users' personal information.

1.2 Structure of social networks

To study the privacy risks on social networks, it is important to understand their structure and the different types of publication and interaction that they allow. A social network anatomy will allow us to detect inconsistencies in their privacy policies. For example, administrators of a given group "g" on Facebook can choose to make it secret and hide the list of members. However, if members leave the default visibility settings on their profiles, their memberships to "g" will be displayed publicly on their profiles.

In this section we present a general model of On-line Social Networks (OSN). Then we present four examples of social networks among the most used in France in 2018 [Tauzin, 2018]: Facebook the preferred network in France with 33 million active users, Twitter with 15.6 million active users, Linkedin with 10.7 million active users and Viadeo with 3.5 million active users.

General social network model

A social network is a website that allows users to create a personal page called user profile to share information and communicate between them. These user profiles contain attributes specified by

their owner, publications, and interactions on publications. User profiles are interconnected through relationship links such as friendship links and through interactions such as comments.

Types of Link between user profiles. We distinguish two types of link between user profiles on social networks: (i) symmetrical links and (ii) unsymmetrical links. For both types only two user profiles are engaged. The symmetrical links are instantiated by a user profile and confirmed by the second one as, for instance, relationship on Linkedin, Contact-ship on Viadeo or friendship on Facebook. On the other hand, unsymmetrical links are established by one user to follow another as, for instance, follow-ship on Twitter or Facebook.

Personal information types We distinguish three types of personal information concerning user profiles on social networks: (i) attributes, (ii) posts and (iii) interactions.

Values of Attributes are defined by users to complete their profiles in order to better present themselves and expand their networks. We distinguish two types of attributes: (i) standard attributes and (ii) custom attributes. Users can choose the values of the standard attributes from predefined lists. For instance they can choose their genders from two options male or female. However, they are free to choose the value of the custom attributes. For instance they can upload customised profile pictures.

Posts are publications made by users on social network that allow them to express their points of view and initialize discussions. We distinguish three types of posts: (i) link, (ii) text, and (iii) multimedia. Users can include links in their posts such as URL links to web pages, hash-tags or profiles tags. They can just publish texts including alphanumeric characters and emoticons. Moreover, they can integrate multimedia contents into their publications such as photos, videos, animations or sound records.

The third type of personal information is interactions. We distinguish three types of interaction: (i) share, (ii) comment and (iii) rating. Users can share the posts of other users on the same social network such as retweets. They can also comment their own posts and those of others. Moreover, they can classify their own publications (posts or comments) and those of others. For instance users can put like, love, Haha, Wow, sad or Grrr on Facebook publications.

Pages and groups. In addition to user profiles some social networks allow the creation of pages and groups that unlike the user profile can be administered by several users. We distinguish two types of link between a user profile and a page or a group: (i) administrator and (ii) member. Administrators are the users who create the group or page and have the right to delete it. They manage its privacy settings and have the right to exclude or add members. Users can ask to become members of a group or like a page. They interact within the page or group through posts or interactions and receive notifications.

Visibility settings. We distinguish six visibility levels on social networks: (i) not connected, (ii) connected, (iii) first degree of connection, (iv) first and second degrees of connection (v) customised list of users and (vii) private. Users can choose to made their data accessible to all Net surfers including the not connected one to the social network in question. They can also restrict visibility to only connected users. Moreover, they can choose to reveal their data to only their direct connections or connections of first and second degrees. Besides, they can customise the list of users that can display their data. Finally, they can hide the data to all Net-surfers by making it private.

Social networks examples

In this section we present the structure of four social networks among the most used in France in 2018 [Tauzin, 2018].

Facebook. Facebook allows users to specify several attributes. According to the French law [CNIL, 2010], the most sensitive of which are: (i) religious Views, (ii) political views, (iii) hometown, (iv) works, (v) sexual orientation and (vi) relationship status.

Facebook users can declare their religious affiliation and explain their beliefs. They can also declare their political views and reference the official pages of the parties they support. It is also possible to reference home cities which may have official pages on Facebook. Moreover, users can specify the company, job title, city and work period for each job. Sexual orientation on Facebook can be inferred through two attributes (interested in and gender). Facebook defines eleven relationship status but allows users to choose only one status and mention the profile of the person concerned by this relationship.

Facebook defines four levels of visibility for these attributes: (i) public, (ii) friends, (iii) only me and (iv) custom. If the visibility setting is "public" then all Facebook users can see the value of the attribute. If it is "friend" then only the friends of the user can see the value of his attribute. By choosing the "only me" visibility option the user hide his values to all the network. Finally, he can customise the list of users who can view his values of attribute. The same network only defines two levels of visibility for the list of groups: (i) public and (ii) only me.

Group administrators on Facebook can set the visibility level of their groups as follows: (i) public, (ii) closed and (iii) secret. If the group is public then all Facebook users can see the list of its members and the publications that are made in it. However, if the group is closed then all Facebook users can see the list of its members. But only members can see the publications made in this group. Finally, if the group is secret then only the members can see the list of its members and the publications.

Twitter. Twitter allows users to specify only eight attributes: (i) location, (ii) birthday, (iii) website, (iv) name, (v) photos, (vi) biography, (vii) followers and (viii) following. Twitter users can specify their home addresses and birthday. They can also write their biographies and add links to their personal web pages on their profiles. Twitter users can choose a nickname of up to 20 alphanumeric characters. Nicknames may not be unique on Twitter. Users can personalize their profile photos and banner photos. Every Twitter user has a list of followers and a list of following. All attributes on Twitter are public by default. The user can only change the visibility settings of his tweets and his date of birth if he is over 18 year-old. Five visibility levels are defined for the date of birth: (i) public, (ii) followers, (iii) following, (iv) following each other and (v) only me. If the visibility setting is "public" then all users, even unidentified on Twitter, can see the date of birth. If it is "followers" then only the follower of the user can see his date of birth. Similarly, if it is "following" then only the followed person by the user can see his date of birth.

Only two visibility levels are defined for tweets: (i) public and (ii) followers. Moreover, Twitter does not support the creation of groups and pages.

Linkedin. Linkedin allows users to specify several attributes, the most important of which are: (i) experience, (ii) education, (iii) languages, (iv) featured skills & endorsements and (v) groups. Linkedin users can specify the company, the job, the city and the work period for each job. They can also specify the school or university they attended, the prepared degree

and education. Moreover, they can specify the languages they speaks with the mastery level. Connection between users on Linkedin are symmetrical. Only direct connections can recommend the user for his skills. Linkedin users can join multiple groups. The network contains about 58k listed french groups.

All attributes on Linkedin are public by default. Users only manage the visibility settings of their connections list. Two visibility levels are defined for this list: (i) connections and (ii) only me. If the "connections" visibility setting is selected then only the direct connections of the user can see his connections list. Otherwise, if the "only me" visibility setting is selected then the connections list will be hidden. The creators of a group on Linkedin can define the visibility settings of the group as follows: (i) listed or (ii) unlisted. If the group is listed then users can find it in Linkedin's group directory. The group also appears on the profiles of its members. Unlisted groups are not listed in Linkedin's group directory and they are not displayed on member profiles.

Viadeo. Viadeo allows users to specify several attributes, the most important of which are: (i) career, (ii) languages, (iii) skills, (iv)list of contacts and (v) list of groups.

Viadeo users can specify their university curriculum as well as the positions they have held during their professional career. They can also specify the languages they speak with the fluency level. Moreover, they can cite their skills and determine their proficiency level in each skill. The contacts on Viadeo are symmetrical and users can join more than one group. The network contains about 37k listed French groups. All attributes values are visible to all Viadeo users by default. Users can only manage the visibility settings of their own contact list. Three visibility levels are defined for this list: (i) all users, (ii) contacts and (iii) only me. If "all users" visibility setting is selected then the list of contacts can be seen by all users. If "contacts" visibility setting is selected then only user's contacts on Viadeo can view his list of contact. Finally, If "only me" visibility setting is selected then only the user can consult his own list of contacts.

A group initiator can define the visibility settings of the group as follows: (i) public, (ii) private or (iii) masked. Viadeo users can search for public and private groups in the Viadeo group directory. Only the content of public groups can be seen by non-members. Masked groups do not appear in the Viadeo group directory and only members of the group can see its contents.

1.3 Research on social network privacy analysis

In this section, we detail related works about privacy on social media. In order to safeguard privacy, it is important to study privacy breaches (or attacks) and the types of sensitive exposed information. A breach occurs when secret sensitive information about a user of social network is revealed.

Privacy attacks

We distinguish two categories of privacy attacks: (i) disclosure attacks and (ii) inference attacks.

Disclosure attacks. Disclosed information is revealed information with certainty. In this case, the revelations rely on undeniable and solid evidences. Hence, they can be used in a lawsuit for instance. In [Price, 2016] the author have listed 20 tales of employees who were fired after secret sensitive information about them was disclosed on social media. Some of them were alcoholic others smoke weed, lie to leave work early or use fake sickness to be on leave. From this angle, we define disclosure attacks as attacks that aim to bypass the visibility settings of sensitive information provided by the social network.

Inference attacks. On the other hand, inferred information is derived with uncertainty. Usually, the adversary needs to collect huge amount of information about the target. Then, he uses cross-correlation techniques in order to infer the secret sensitive information. From this angle, we define inference attacks as attacks that aim to guess sensitive information that may not be stored anywhere in the social network.

Types of sensitive exposed information

We distinguish three types of sensitive information exposed on social networks. Consequently, we define three prediction attacks: (i) identity prediction, (ii) link prediction and (iii) attribute prediction.

Identity prediction. Identity prediction consists of linking a set of social network profiles to a real person. To conduct such attacks, the adversary needs first to collect identifying and quasi-identifying information across social networks such as age, gender, zip-code and language. Several techniques have been investigated in order to anchor link profiles across different social networks. These techniques help to complete a target profile for increasing identification accuracy [Chen et al., 2012]. We distinguish two categories of anchor linking profiles across social networks: (i) anchor linking profiles across homogeneous social networks and (ii) anchor linking profiles across heterogeneous social networks.

In [Golbeck and Rothstein, 2008], the authors compare the "friend of friend" networks between homogeneous networks in order to anchor link profiles across them. In [Man et al., 2016] the authors have proposed a supervised model called PALE to anchor link profiles across two homogeneous social network based on their friendship structure. First, PALE computes a low dimensional embedding of each user profile in each social network. Then, the profiles known to be anchor linked across both social networks are used to train the model and compute the mapping function. The resulting mapping function is one-to-many. In other words, for each profile from the first network the model predicts a list of potentially anchor linked profiles from the second network. However, structure-based models fail to anchor link profiles across heterogeneous social networks. In fact, social graphs properties can be quite different between heterogeneous social networks [Wu et al., 2014]. For instance, the friendship graph on Facebook is undirected. On the other hand, the followship graph on Twitter is directed. Moreover, the structure of a user's professional connections on Linkedin and his following buddies on Instagram are not necessary similar.

To cope with this problem and anchor link profiles across heterogeneous social network, several solutions have been investigated. In [Jain et al., 2013], the authors use network structure, values of attributes and features as well as generated publications and post in order to cross link Facebook and Twitter profiles. In [Liu et al., 2013], the authors have proposed an unsupervised approach that analyses user-names to anchor link profiles. In [Kong et al., 2013], the authors compare the features of users across two heterogeneous social networks: Foursquare and Twitter, where features include location from where users have published posts, times slots when users have published posts, set of words used in posts ... Based on this comparison, they designed a one-to-one model to anchor link a given profile from one social network to only one profile from the other social network. In [Ma et al., 2017], authors have proposed a hybrid model called MapMe. MapMe uses both network structure and profile features to anchor link profiles across different social networks.

However, aligning profiles across heterogeneous social networks remains an open challenge as features are differently defined across them. For instance, Facebook users declare their relation-

ship status via a drop-down menu. On the other hand, Instagram users declare their relationship status though photos with caption and emoji [Lorenz, 2017].

Once several profiles across different social network are anchor linked, an adversary can piece together identifying and quasi-identifying attributes and features such as age, gender and location from where posts are made in order to reveal the real identity of the target user.

Link prediction. Link prediction consists of inferring (with uncertainty) or disclosing (with certainty) links between users of the same social network. Several link prediction methods have been investigated since the blossom of social networks [Wang et al., 2015b, Gao et al., 2015].

Researches on link inference have been first motivated to improve link recommendation algorithms in social networks. These research works consist of analysing off-line a sub-graphs of the social network. Their main objective is to evaluate the probabilities of new links emergence. However, they can also be used by an adversary to evaluate the possibilities of existing secret links. The first investigated link inference methods fall under the category of unsupervised link inference. They only focus on network structure. Most of them compute a link score between two given nodes. The most popular scoring functions are Katz [Katz, 1953], preferential attachment [Newman, 2001], Adamic/Adar [Adamic and Adar, 2003] and Jaccard [Liben-Nowell and Kleinberg, 2007]. These methods are usually used as a baseline for supervised learning methods that have appeared later.

In addition to network structure, supervised learning methods take into consideration users' attributes and features to infer links [Lichtenwalter et al., 2010, Cukierski et al., 2011]. Observed links and similarity between users are considered in the field of supervised learning as predefined labels used in training in order to infer new links. Several supervised inference algorithms have been recently investigated including random forest classifiers [Guns and Rousseau, 2014], adaptive boosting [Wang et al., 2015a] and link utility [Li et al., 2017].

The link disclosure works are more recent and have much to do with privacy. They consist of analysing social networks in on-line way. Their main objective is to bypass visibility settings and disclose with certainty the secret links between a target user and other users. We distinguish two types of attacks as defined in [Backstrom et al., 2011]: (i) passive attacks and (ii) active attacks. In passive attacks the adversary does not change the structure of the social network and does not create new links. However, in active attacks he creates new links that allow him to conduct his attack. In both attacks, adversary must interact with the social network in on-line way in order to disclose links.

In [Korolova et al., 2008], the authors analyse the vulnerability of social networks to link disclosure attacks. They investigate several strategies to conduct such attacks based on features provided by the social networks themselves. These features include lookahead, search interface, degree of users and complete user-list of all network. We recall that a social network has lookahead l if any user can see the friend list of any other user within distance l from himself. For instance, LinkedIn has lookahead 1 and the lookahead on Facebook depends on the visibility setting of each user (it can be 0, 1 or ∞). Search interfaces allow users to search for other users by identifying information. Moreover, an adversary can take advantage of several additional functionalities (APIs) provided by social network such as shortest path, length of shortest path, list of mutual friends and number of mutual friends. In [Jin et al., 2013] the authors investigate the mutual-friends query in order to disclose links. They analyse the success rate of active attacks on 1 lookahead social networks such as LinkedIn. Adversary needs to create new nodes and new links between them and other users in order to disclose links using mutual-friends query. This query is available in most social network with different restrictions policies. For instance, any Facebook

user can query mutual friends between any two other users. However, a LinkedIn user can only query mutual friends between himself and other users.

Attribute prediction. Attribute prediction consists of inferring or disclosing explicit attributes such as age, gender and relationship status as well as implicit attributes (also called features) such as time slot of connection, location of connections and characteristics of devices. We distinguish two categories of attribute prediction: (i) unsupervised prediction and (ii) supervised prediction. Unsupervised prediction relies on clustering algorithms. In this category, the data is not labelled (named) at the beginning. Clustering algorithms determine the groupings that will be labelled later. For instance, clustering algorithms yield several clusters of users based on the similarities between their published attributes. These clusters are usually considered as communities. Once the community of a given target is revealed, the values of attributes of its members can then be used to predict the missing values of attributes of the target [Mislove et al., 2010, Hu et al., 2017]. For instance, to predict the political orientation of a target user u, the algorithm defines several clusters of users in a way that maximize the similarities between users belonging to the same cluster. The favourite politicians of users that belong to the same cluster are marked with similar labels (right or left politicians for instance). The target user u will then inherit the label of the clusters to which he belongs. In other words, the algorithm discovers the labels (clusters) and assigns the right ones to the target.

On the other hand, supervised prediction relies on classification algorithms. In this category, all the classes (labels) are predefined at the beginning and a part of the data is initially classified (labelled). The classifier then relies on classified data to predict the classes of unclassified ones. In the field of machine learning, a first part of the initially classified data is called the training data. This part of data is used to design a model that minimize the error when predicting the classes. A second part of the initially classified data is called the validation data. This part of data is used to validate the model by tuning the optimal parameters in order to maximize the prediction accuracy. A third part of the initially classified data is then used to compute the model accuracy by comparing the predicted classes (labels) to the real classes. Finally, the unclassified data is called the application data.

In the case of attribute prediction problem in social network the set of classes is the set of values of a given attribute or feature. For instance the classes are "*Male*" and "*Female*" in the case of predicting the gender of users.

Following [Zheleva et al., 2012], we define three main categories of classifiers for attribute prediction in social networks: (i) content-based classifiers, (ii) link-based classifiers and (iii) content&link-based classifiers.

We recall that social network contents includes attributes (implicit and explicit) as well as posts. Content-based classifiers (also called local classifiers) do not take into consideration relationships and interactions between users. In this case, statistic models such as linear classifiers are used to predict the classes (the unknown values of sensitive attribute) of users based on their published attribute values. These types of classifiers are widely used in goods recommendation systems where interactions between customers are not available [Smith and Linden, 2017].

On the other hand, link-based classifiers (also called relational classifiers) rely on the assumption that *"birds of a feather flock together"*. These types of classifiers shed the light on the importance of social network formed by users' interactions and relationships. They exploit only structural information of the social graph to classify the nodes. In [Perozzi and Skiena, 2015], the authors design a supervised age predictor. The predictor uses a linear regression function that takes as input a vectorial representation of users (embeddings). The vectorial representa-

tion of each user is computed based only on the structural information of the social network [Perozzi et al., 2014].

Finally, content&link-based classifiers (also called collective classifiers) balance the importance of structural information and the content (attributes and posts).

In [Chester and Srivastava, 2011] the authors introduce a α-proximity notion to combat attribute prediction attacks in social network. The proposed solution consists in creating new links between users in order to reduce the variation gaps between all local networks and the global network, where each local network is formed by a given user and his friends. The variation in their work refers to the variation of the distribution of attribute values. A user is considered to be vulnerable towards attribute prediction attacks if the distance between the variation of his local network and the variation of the global network is less than α. In [Conover et al., 2011], the authors design a classifier to predict the political alignment of Twitter users based on the content of tweets (text mining) as well as the structure of the network of re-tweet and the network of mention (graph mining). They show that such classifiers widely outperform only content-based classifier. In [Zhang and Zhang, 2012], the authors introduce an information re-association attack in order to predict the values of sensitive attributes of users. This attack consists in combining web search (through search engines) with information extraction and data mining techniques. The study shows that the attack is more successful when it takes into consideration information about the network of the target such as the network of his universities. In addition, it shows that Facebook graduated users from top schools are more vulnerable under this attack than random users. In [Heatherly et al., 2013b], the authors propose a content&link-based classifier that outperforms both content-based classifiers and link-based classifiers when predicting the political views and the sexual orientation of Facebook users. In addition, they explore the effectiveness of sanitization techniques to combat such attacks concerning released data. In contrast to [Chester and Srivastava, 2011] where sanitization solutions consist of adding new link between users, in [Heatherly et al., 2013b], sanitization solutions consist of removing content and links. However, selecting the right contents and links to remove without altering the data utility remains an unsolved challenging problem. In [Ryu et al., 2013] the authors show that an adversary can infer sensitive attribute values of a target based only on the target local network (1-hop friendship network) and the public attribute within it. The proposed predictor takes into consideration the structure of the network by quantifying the importance of friendship. Then, it measures the power of each attribute value according to the importance of the target friend that publishes it. In [Gong et al., 2014], the authors extend the attribute-augmented social network model that is introduced in [Yin et al., 2010]. In the first model, attribute values are represented by nodes. The users that publishes a particular value of an attribute are linked to its representing node in the model. The extended model adds negative links between users and their hidden attribute values and mutex links between mutually exclusive values of the same attribute such as male and female. This model is used with both supervised and unsupervised method to predict links between users (link prediction) as well as links between users and values of attributes (attribute prediction). In [Vidyalakshmi et al., 2016], the authors design a classifier to predict the missing attributes values of a Google+ user. The classifier takes into consideration only the target local network (only 1-hop users from the target). In addition to the attributes, the designed classifier takes into consideration the direction of links (follower or followings), the type of links (acquaintance, family, friend ...) and the tie-strength of the links. In [Gong and Liu, 2018], the authors introduce a social-behaviour-attribute (SBA) network model that extend the attribute-augmented social network model [Yin et al., 2010] by adding behaviours nodes to the framework that already integrates user nodes and values of attribute nodes. Then, they design a vote distribution attack (VIAL) under the SBA network model to predict attribute values. They show that

by taking into account social friendship, attributes and behaviours, the accuracy of the attacks is considerably increased.

1.4 Motivations and challenges of this work

In this work we aim to provide social network users a tool to safeguard their privacies. To that end, we investigate potential privacy attacks, study their feasibilities and analyse their impacts. This approach allows us to put the hand on the origins of threats and design effective countermeasures. Concretely, we design on-line attacks on the world biggest social network, "Facebook". The attacks are tested on-line on several real volunteer profiles.

In order to effectively combat privacy leakage, it is of high importance to take into account the combination of attacks (identity prediction, link prediction and attribute prediction). In fact, these attacks are strongly related and when combined they present higher threats on privacy. For instance, an adversary can perform link prediction attacks in order to disclose the local network of his target (1-hop from the target). Then, he can perform attribute prediction attack based on the discovered local network. Finally, he can perform an identity prediction attack based on the disclosed values of attributes.

It is also important to take into account the feasibility of on-line attacks. For instance, in order to perform on-line link disclosure attacks on Facebook, an adversary may be tempted to check public friends lists of Facebook users with the hope of finding the target in those lists. However, Facebook counts about 2 billion monthly active users and a random approach may last for years. Moreover, Facebook is highly dynamic. For instance, every second 5 new profiles are created and 8 500 comments are posted [Noyes, 2018]. Thus, attacks based on network pattern recognition are quite difficult to mount.

On-line attack encompasses two steps: (i) data collection and (ii) data analysis. Data collection must be fast, selective, passive and unnoticed. In fact, social networks are highly dynamic and contain big data. Random collection may result in useless data. On the other hand, massive collection is time wasting. A fast and selective sampling algorithm must be used in order to guide the collector toward most important data and speed up the process. Moreover, the adversary must limit his interaction with his target. He must perform his attack in passive way in order to avoid raising the attention of the target to him. The adversary must also use only legal requests to collect data and should not exceeds thresholds in order to remain unnoticed by the social network.

Data analysis must be fast, accurate and deal with sparsity. We recall that the system (collection and analysis) is meant to help users safeguard their privacy against real attacks. Hence, data analysis must not exceed few minutes in order to rapidly put the hand on the origins of threats and quickly put countermeasures in action. Results of the analysis should be accurate in order to reduce false positive alerts and inspect all threats. As the collector only samples few data from an ocean of data, the analyser should deal with the fact that collected data may be sparse and incomplete.

1.5 Contributions of the thesis

In this work we focus on predicting personal sensitive information as they are responsible of the highest privacy damage. The main contributions of the thesis are as follows:

Attribute sensitivity measure

We have proposed a sensitivity measure to quantify the sensitivity of subjects (a subject may refer to a set of attributes). We have also estimated the percentage of french Net-surfer that are vulnerable to several attack scenarios though a questionnaire survey. This is in contrast to the definitions of sensitive attributes proposed in [Ryu et al., 2013, Vidyalakshmi et al., 2016] where all the unpublished attributes (masked or not specified) by a given user are considered sensitive for him. In other previous works, researchers select a few attributes and consider them sensitive such as political affiliation [Heatherly et al., 2013b, Conover et al., 2011], sexual orientation [Heatherly et al., 2013b] and age [Perozzi and Skiena, 2015].

Link disclosure strategy with certainty

We have designed an on-line link disclosure attack strategy (with certainty). The proposed strategy is passive: the adversary does not have to interact with his target. Our attack is performed on (1,2 or ∞)-lookahead social network and has been tested on-line with volunteer profiles. This is in contrast to off-line link inference (with uncertainty) attacks proposed in [Wang et al., 2015b, Gao et al., 2015] and the active link disclosure attack in [Jin et al., 2013] performed on 1-lookahead social network (Linkedin) that discloses friendship through mutual friend query but was not tested on-line. Furthermore, by effectively exploring the target group network, our proposed attack strategy is able to perform group-membership, friendship and mutual-friend attacks along a strategy that minimizes the number of queries. The results of attacks performed on active Facebook profiles show that 5 different friendship links are disclosed in average for each single legitimate query in the best cases.

Let us note that we can apply these on-line link disclosure attacks (with certainty) to prepare attribute inference attacks (with uncertainty) in the tool described below.

A tool to prevent attribute inference attacks

To increase user awareness about privacy threats we have designed a tool, SONSAI, for Facebook users to audit their own profiles. The system first crawls the network around the user while performing link disclosure attack. Then it predicts the values of sensitive attributes using a machine learning engine. The results provided by SONSAI, quantify the correlation between attributes and shows the public attributes of the user that have oriented the learning algorithm towards a particular sensitive attribute value. The tool is fully interfaced for Facebook, however it can be adapted to many other social networks. The system has been tested by several volunteer users for auditing their Facebook profiles. In each case a dataset was built from real profiles collected in the user neighbourhood network.

The whole analysis process in SONSAI does not exceed 5 minutes when analysing the target local network (containing a few hundreds of profiles and attributes). We notice that inference accuracy (measured by the Area Under the Curve "AUC") changes according to the sensitive attribute. It is about 0.83 for gender inference, 0.7 for relationship status inference and 0.79 for politicians inference .

For the approach to get feasible, we have solved several problems:

Selective crawling. First, data collection by crawling is limited both by the social network and by country regulations. Hence, we have designed a crawling exploration strategy that focuses only on meaningful representative network nodes.

Relevant attribute selection for learning sensitive ones. As in [Yin et al., 2010], we have used an attribute-augmented graph to model social networks. In our model, each attribute is modelled by a distinct bipartite graph. Since attributes are numerous, for the learning algorithms to scale, one has to select only the most relevant ones for inferring sensitive attribute values. To that end, we have introduced a relevance measure that is both accurate and easy to compute. This measure quantify the correlation between attributes through only graph structural analysis. We note that we cannot rely on semantic proximities since we notice that users who hide a sensitive attribute also hide semantically related ones. Moreover, we anonymize the data before analysing it. Since data processing could be delegated to a third party, semantic information is discarded to preserve users' anonymity.

Data sparsity and high range of attribute values. Collection tasks may result in sparse and incomplete data that alters the final results. To cope with this problem, we have designed a graph merging algorithm that derives new dense graphs by merging several sparse graphs. Additionally, some attributes such as favourite politicians have high range of possible values. To help end users manage their privacies and make inference results easy to understand, we have designed a greedy algorithm to cluster values of a given attribute by similarity of preferences.

Fast on-line inference. In order to perform attribute inference from possibly sparse datasets collected in short time by user in his local-network, we rather use random walk-based learning. The random walk technique has been applied to social representations in [Perozzi et al., 2014] and [Grover and Leskovec, 2016] but only to predict friendship links. In [Ryu et al., 2013] the authors also propose algorithms to detect whether a sensitive attribute value can be inferred from the neighbourhood of a target user in a social network. Heatherly et al. [Heatherly et al., 2013a] infer values of attributes in social network with Bayesian classification techniques. For the same purpose, Estivill-Castro et al. [Estivill-Castro et al., 2014] employ decision-tree learning algorithms. In these previous works, unlike ours, learning is performed off-line on large datasets.

Contributions have been published in:

- [Abid et al., 2016a] Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Marc Rigolot, Michaël Rusinowitch:Analyse d'activité et exposition de la vie privée sur les médias sociaux. EGC 2016: 545-546

- [Abid et al., 2016b] Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Michaël Rusinowitch: Online Link Disclosure Strategies for Social Networks. CRiSIS 2016: 153-168

- [Abid et al., 2016c] Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Michaël Rusinowitch: Stratégies de divulgation de lien en ligne pour les réseaux sociaux. BDA 2016

- [Abid et al., 2017] Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Michaël Rusinowitch: Two-Phase Preference Disclosure in Attributed Social Networks. DEXA (1) 2017: 249-263 2016

- [Abid et al., 2018] to appear. Younes Abid, Abdessamad Imine, Michaël Rusinowitch: Sensitive attribute prediction for social networks users DARLI-AP 2018 (EDBT-Workshop) to appear.

1.6 Outline

We detail now the organization of this thesis. We arrange the major topics covered by each chapters as follows:

In Chapter 2, we introduce subject sensitivity measure for social networks. Then, we inspect the vulnerability of french Net-surfers toward some privacy attacks based on the behaviour of participants in our survey. In Chapter 3, we design and test on-line link disclosure attacks. In this chapter, we address the problem of rapidly disclosing friendship and group membership links using only legitimate queries (i.e., queries and tools provided by the social network). We perform several on-line attacks on Facebook to test the feasibility and the results of our attacks. In Chapter 4, we present the architecture of our system SONSAI. Then, we give a user guide and depict two inference scenarios. In Chapter 5, we detail our sampling and crawling algorithms. Then, we throw light on modelling and anonymizing collected data. In Chapter 6, we detail our cleansing algorithms. These algorithms quantify the correlation between attributes, select the most relevant ones for inference and combat data sparsity. Finally, in Chapter 7, we explain our inference algorithm. Then we detail the results of the conducted inference attacks.

2

Defining sensitive subjects

Contents

2.1 Introduction **19**

2.2 Conducting a survey on the behaviour of French users of social media **20**

2.3 Analysing responses and defining sensitive subjects **23**

2.4 Possible attack vectors according to the behaviour of participants . **32**

2.5 Conclusions **36**

2.1 Introduction

In order to combat privacy leakage, it is important to define what personal information are sensitive. Some researchers consider that all the unpublished values of attributes (masked or not specified) by a given user are sensitive for him [Ryu et al., 2013, Vidyalakshmi et al., 2016]. While others subjectively select a few attributes and consider them to be sensitive such as sexual orientation [Heatherly et al., 2013b], political affiliation [Heatherly et al., 2013b, Conover et al., 2011] and age [Perozzi and Skiena, 2015]. It is also possible to rely on the definition of sensitive information given by law. However, social networks evolves faster than the law. For instance, health data were not considered sensitive by the French law of January 6, 1978 relative to computers, files and freedoms (version 1978). It was considered sensitive much later. It is also possible to rely on a definition of sensitive subject given by social media themselves. For instance, according to Google, sensitive data are "*relating to confidential medical facts, racial or ethnic origins, political or religious beliefs or sexuality*" [Google, 2018]. But how can we trust social networks in defining what is sensitive or not knowing that they make most of their profit using personal information for targeted advertising ?

In this chapter, we conduct a questionnaire survey to define sensitive subjects based on the behaviour of french Net-surfers. This method has the advantage of being fast, objective, accurate and up-datable. Only a few weeks are enough to conduct this survey. Sensitive subjects are defined by the users themselves instead of being imposed by social networks or laws. In addition, it is possible to take into account statistics results of more recent surveys in order to update the definition without repeating the whole process. Moreover, the survey has allowed us to assess the vulnerability of Net-surfers toward some privacy attacks.

2.2 Conducting a survey on the behaviour of French users of social media

In this section, we present the results of our survey as well as how we have checked the validity and consistency of the responses to eliminate the random ones. Moreover, we evaluate the representativeness and reliability of our studied sample. The questions of our survey are presented in detail in Annex B.

Cleansing the responses

The questionnaire was distributed between April 4, 2015 and August 21, 2015 and submitted to the University of Lorraine staff (researchers, teachers, administrative agents, doctoral students, etc.) and MAIF customers. We collected 345 responses including 85 incomplete ones. Among the 260 participants who provided complete answers, 27 of them said that they do not use social networks and do not have profiles on forums and websites.

Participants	Use social networks	Do not use social networks	Total
Use forums or websites	197	19	216
Do not use forums or websites	17	27	44
Total	214	46	260

Table 2.1: Distribution of responses by use of social media.

The answers to Questions 8 and 9 (see Annex B) show that among all respondents, 17 of them use exclusively social networks, 19 use of them use exclusively forums or websites and 197 of them use both. Table 2.1 summarizes these results. To focus on answers of interest to our study, we discarded responses of the 27 participants who do not use social media [1].

To identify random responses, we identified respondents whose behaviours were not dominant compared to most participants, and then analysed the consistency of their responses. Thus, from the answers to Questions 1 and 32 (see AnnexB), we found the following dominant behaviours:

- All the participants who followed the television programs “c’est mon choix” and “sacrée soirée” were over 9 years old during their broadcast respectively on France 3 in 2004 and TF1 in 2001.
- Among all 117 participants who watched the show “ 7 sur 7 ”, only 5 were less than 9 years old when it was broadcast on TF1 in 1997.
- All Atari 2600 users declare that they have also used the floppy disk.
- Only 9 out of 130 GSM users did not use the floppy disk.
- Only 11 out of the 119 carriers of the latest smart-phones from Samsung or Apple have not used the WiFi in public places.
- Only 9 out of 171 participants who handled the VHS player did not use the DVD.

[1] Social media refers to social networks, forums and websites.

- Only 3 respondents say that they follow the " jour de foot " TV-show and do not talk about sports on social media.

On the other hand, we found that only one participant has a very distinct behaviour from the other participants in several answers. When analysing his answers, we noticed that he checked only the choices in the middle of the lists of choices. For example, he stated that he is active on several networks and forums but he only discusses topics related to health and receives only ads on real estate. Thus, we have rejected his participation because his answers are random.

In the following, we limit our study to a sample of 232 complete and valid responses from social media users.

Sample representativeness

There are several definitions of the representativeness of a sample. In this work we adopt the viewpoint of Olivier Sautory [Gerville-Réache and Couallier, 2011]: "A sample is never representative in itself", it is representative with regard to certain variables. In this section, we study the representativeness of our sample with regard to the "age" variable.

Sample dispersion. For our sample, the age of participants varies between 20 and 78 years, the average age is 40 years and the standard deviation is equal to 12.64. The coefficient of variation is 31.6%. The distribution by gender and age of the 232 respondents to the survey is detailed in Figure 2.1. The participation of women accounted for 63.36% of the participants with an average age of 39.03 years. As for men, the average age is 41.67 years.

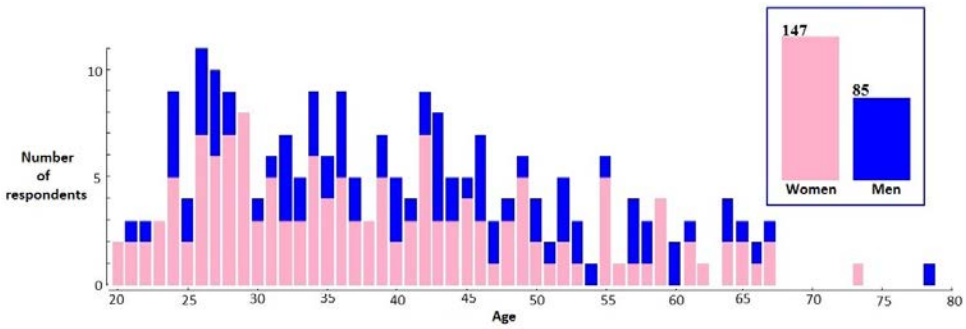


Figure 2.1: Distribution of participants by gender and age.

Our sample covers 90.9% of France's not overseas regions. Moreover 54.31% of participants are from Lorraine. This concentration is explained by the dissemination of the survey through emails addressed to subscribers of "expression libre" from the University of Lorraine. On the other hand, responses from other regions were collected thanks to the newsletter of MAIF Foundation. The distribution of participants by region is shown in Figure 2.2. The distribution by academic discipline of the participants is described by Figure 2.3. The participants who have completed studies in formal sciences [2] or natural sciences [3] account for 56.96 % of our sample.

[2]The formal sciences are: Logic, Mathematics, Computer Science, Theoretical Computer Science, Discrete Mathematics ...

[3]The natural sciences are: Biology, Chemistry, Physics, Physics Chemistry, Earth Sciences or Geoscience ...

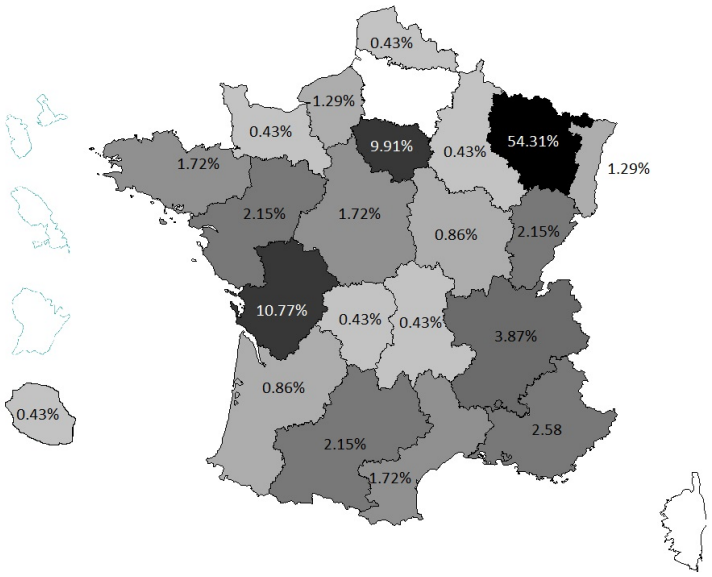


Figure 2.2: Distribution of participants by region.

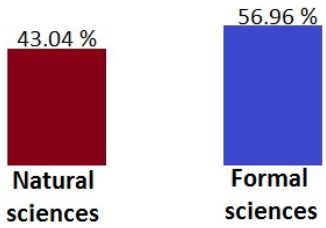


Figure 2.3: Distribution of participants by study discipline.

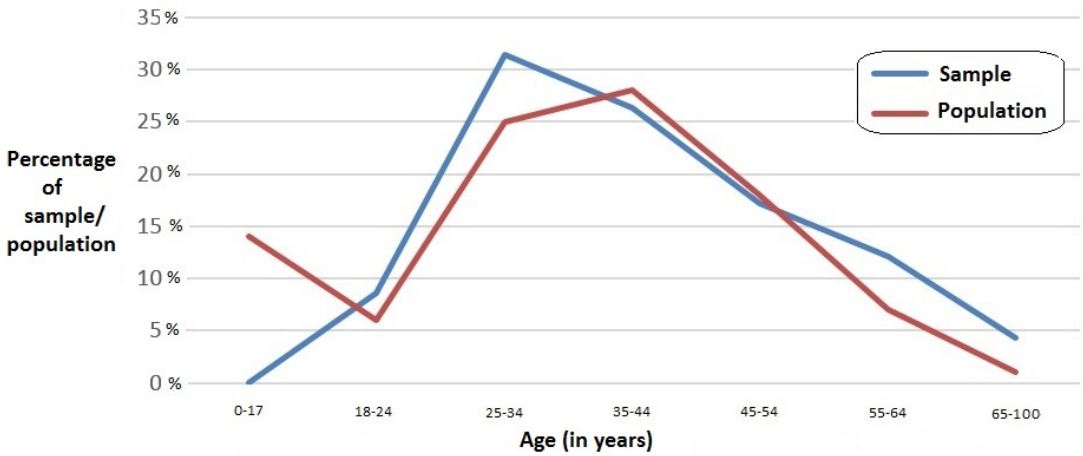


Figure 2.4: Distribution of the age of the sample with regard to the mother population.

Sample stratification. To evaluate the stratification of our sample, we compare the age distribution of our sample to the demographics of social network users in France [2803media.fr, 2010].

The correlation between the two distributions (see Figure 2.4) is 0.8, and the correlation between the sample and the population over 18 year is equal to 0.95. We note that the age proportion of our sample among the mother population of social network users is high. Therefore, we conclude that our sample is fairly representative of the mother population with regard to the "age" variable.

Sample reliability

In this section, we compute the error margins of our survey following different confidence thresholds. For a very large mother-population of social network users, we apply Equation 2.1 with the following notations:

- n: sample size
- s : confidence threshold
- t : margin coefficient deducted from the confidence threshold s
- e : margin of error
- p : estimated proportion of the population with certain characteristics

$$e = t\sqrt{\frac{p(1-p)}{n}} \tag{2.1}$$

It is difficult to calculate the exact proportions of all Internet users who share certain characteristics among the mother population of social network users. Thus, we use a proportion $p = 0.5$ in all our calculations. The margin coefficient t is deduced directly from the confidence threshold via the normal law table. Table 2.2 represents the error margins e of our sample ($n = 232$) for different levels of confidence and a proportion p of 0.5. In the following, we consider that

Threshold of confidence s	Margin coefficient t	Margin of error e
80	1.28	0.0420
85	1.44	0.0473
90	1.645	0.0540
95	1.96	0.0643
96	2.05	0.0673
98	2.33	0.0765
99	2.575	0.0845

Table 2.2: Margins of error of the studied sample.

the confidence threshold s is equal to 85%. Thus, the margin of error e is equal to 4.2% for all statistics.

2.3 Analysing responses and defining sensitive subjects

In this section, we analyse 18 different subjects discussed on social media. We classify these subjects according to four criteria: discussion on both forums and social networks, avoidance and anonymity. Moreover, we identify the favourite topics of Net surfers by gender and level of

study. Besides, we propose a formal definition of sensitive subjects and we identify the sensitive subjects among those studied in our survey. Finally, we define a sensitivity coefficient to classify subjects by sensitivity level.

Analysing the survey responses

We have analysed the social media discussions of 18 various subjects studied in the questionnaire to identify Net surfers' favourite subjects by gender and level of study. We classify these subjects according to four criteria: (i) rate of discussion of subjects on social networks, (ii) rate of discussion of subjects on forums and websites, (iii) rate of anonymity of publication and (iv) rate of avoidance of subjects.

Rate of discussion of subjects on social networks. Table 2.3 classifies the various subjects proposed in question 35 (see Annex B) in increasing order of discussion according to the type and level of study. Subjects that have global chat frequencies below the average frequency minus the standard deviation are: "*Money, Shopping Religion and Dating*".

Men's favourite subjects are "*News and Technologies*". As for women, they discuss more about "*News, Going out and Travel*". Men talk about "*Money, Religion, Technology and Dating*" twice as much as women. The latter talk about "*Fashion*" twice as much as men.

Net surfers who have been to graduate school are more inclined to discuss "*News and Going out*". The subjects "*Family, Travel and Going out*" are the most discussed by Internet users who have not completed higher education. They talk about "*Money and Shopping*" eight times as much and "*TV shows*" twice as much as those with higher education. On the other hand, the latter discuss "*Politics and Technology*" twice more and "*Studies*" thrice more than those who have not followed a higher education.

Rate of discussion of subjects on forums and websites. Table 2.4 ranks the forums and websites proposed in Question 12 (see Annex B) in increasing order of activity according to the gender and level of study. The forums and websites that have global activities rates below the average rate minus the standard deviation are "*Going out, Dating and Chat*" and "*Philosophy, Religion and Free Thinking*" sites and forums.

Men prefer forums and sites about "*Computer Science and Technology*". Conversely, women prefer forums and sites about"*Health, Shopping, Kitchen, House and Tip*" and "*Travel, Transport, Holidays, Insurance*". Although the overall activity on forums and sites of "*Philosophy, Religion, Free Thinking*" does not exceed 27%, 30.86% of men are active on these sites.

Participant Net surfers who have not received higher education are more present on the forums and sites of "*Health, Shopping, Kitchen, House and Tip*", "*Travel, Transport, Holidays and Insurance*" and "*Games, Music, Movie, Humour, Art and Book*". The forums and favourite sites of participant Net surfers who have received higher education are "*Computer Science and Technology*" and "*Health, Shopping, Kitchen, House and Tip*". Even if the forums and sites of "*Going out, Dating and Chat*" are globally less visited, the activity of participant Net surfers who have not followed higher studies on these sites exceeds the threshold of the average of all activities minus the standard deviation.

Rate of anonymity of publication. We distinguish three types of activity on forums and websites: search without publication, anonymous publications and non-anonymous publications. Table 2.5 details the rates for each type of activity on the forums and websites proposed in Question 12 (see Annex B)

Subjects	Rate of discussions in %					Average age
	Global	Men	Women	Graduate studies	Secondary studies	
Money	0.94	1.37	0.71	0.53	4.17	42.50
Shopping	1.88	1.37	2.14	1.06	8.33	34.75
Dating	5.16	8.22	3.57	5.29	4.17	39.63
Religion	5.63	8.22	4.29	5.29	8.33	39.83
Fashion	10.80	6.85	12.86	10.58	12.50	33.74
TV show	15.02	10.96	17.14	12.70	33.33	36.81
Health	17.37	9.59	21.43	15.87	29.17	35.73
Sport	21.13	23.29	20.00	19.58	33.33	36.64
Study	21.60	20.55	22.14	23.28	8.33	36.19
Politics	25.82	28.77	24.29	26.98	16.67	40.10
Kitchen	25.82	16.44	30.71	25.40	29.17	36.05
Technology	27.70	**38.36**	22.14	29.63	12.50	39.05
Work	30.52	27.40	32.14	30.69	29.17	39.35
Family	31.45	27.40	33.57	28.57	**54.17**	37.12
Travel	36.62	28.77	**40.71**	34.39	**54.17**	38.31
Going out	37.56	28.77	**42.14**	**36.51**	**45.83**	33.89
News	53.52	**50.68**	**55.00**	**52.91**	**58.33**	37.86
Other subjects	17.37	17.81	17.14	18.52	8.33	42.40

Table 2.3: Subjects ranked by increasing order of discussions on social networks.

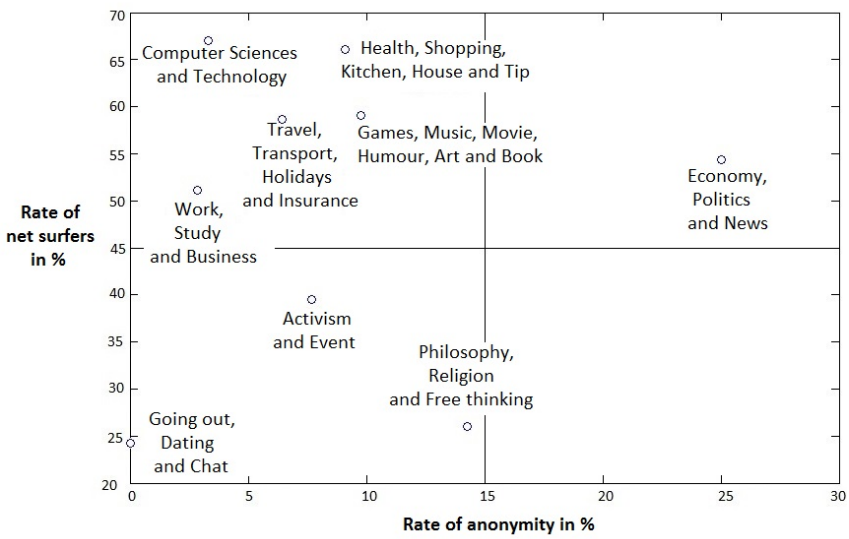


Figure 2.5: Participant Net surfers activity rate on forums and websites.

The rate of anonymous publications represents the percentage of publications made without identification or with anonymous profiles of all publications. It exceeds the threshold of 8.71 % (average of all anonymity) on websites and forums of "*Economy, Politics and News*", " *Philosophy, Religion and Free Thinking*", " *Games, Music, Film, Humour, Art and Book*" and " *Health, Shopping, Kitchen, House and Tip*".

Forums and websites	Rate of discussions in %					Moyenne d'âge
	Global	Men	Women	Graduate studies	Secondary studies	
Going out, Dating and Chat	24.19	23.46	24.63	22.92	34.78	38.05
Philosophy, Religion and Free thinking	26.05	30.86	23.13	26.56	21.74	40.68
Activism and Event	39.53	37.04	41.04	39.58	39.13	40.63
Work, Study and Business	51.16	54.32	49.25	51.56	47.82	39.53
Economy, Politics and News	54.42	56.79	52.98	56.25	39.13	41.83
Travel, Transport, Holidays and Insurance	58.60	46.91	**65.67**	56.25	**78.26**	41.48
Games, Music, Movie, Humour, Art and Book	59.07	59.26	58.95	57.81	**69.56**	39.99
Health, Shopping, Kitchen, House and Tip	66.05	50.62	**75.37**	**64.06**	**82.61**	39.83
Computer Science and Technology	66.98	**83.95**	56.72	**68.23**	56.52	40.37

Table 2.4: Subjects ranked by increasing order of discussions on forums and websites.

Unidentified searches represent more than 60% of all activities on all websites. They exceed 80% on the sites of "*Health, Shopping, Kitchen, Home and Tip*".

Figure 2.5 details the distribution of forums and websites according to the activity rates of participant Net surfers and non-anonymous publications. We notice that the anonymous publication rate and the activity rate are not related. For instance, the rate of anonymous publication on the forums and websites of "*Economy, Politics, and News*" is high. However, the rate of the active participant Net surfers on these forums and websites is higher than 50 %.

Rate of avoidance of subjects. We have analysed the answers to Question 36 (see Annex B) to identify avoided subjects on social networks. This question enabled us to simulate individual directive interviews. Participants have the opportunity to develop a free response in its form and length. Then, we have analysed these responses and defined 10 subjects based on the vocabulary used by the respondents to write their answers. Table 2.6 classifies these subjects in descending order of avoidance. Politics and religion are the most avoided subjects by users of social networks. One out of two participants avoids politics and one out of three participants avoids religious discussions.

Defining sensitive subjects

Sensitive subjects are often defined in relation to cultural, geographical, temporal and social factors. Thus, there are many interpretations and definitions of sensitivity. The data about health were not considered sensitive by the French law of January 6, 1978 relating to computers, files and

Forums and websites	**Rate of anonymous publication in%**	**Type of activity in %**			**Rate of active Net surfers in %**
		Search without publication	Anonymous publications	Non-anonymous publications	
Economy, Politics and News	25	76.07	**5.98**	17.95	54.42
Philosophy, Religion and Free thinking	14.28	75.00	3.57	21.43	26.05
Games, Music, Movie, Humour, Art and Book	9.76	67.72	3.15	29.13	59.07
Health, Shopping, Kitchen, House and Tip	9.09	**84.50**	1.41	14.08	66.05
Activism and Event	7.69	69.41	2.35	28.23	39.53
Travel, Transport, Holidays and Insurance	6.45	75.40	1.59	23.02	58.60
Computer Sciences and Technology	3.33	**79.17**	0.69	20.14	66.98
Work, Study and Business	2.85	68.18	0.90	**30.91**	51.16
Going out, Dating and Chat	0	61.54	0	**38.46**	24.19

Table 2.5: Subjects ranked by decreasing rates of anonymous publication on forums and websites.

Subjects	**Vocabulary**	**Participant in %**
Politics	Politics, War, Conflict and Conspiracy	50.72
Religion	Religion	33.33
Personal and family life	Family, Privacy, Phone Number and Contact	21.74
Sentimental and sexual life	Sex, Feeling, Love, Intimate and Child Porn	17.39
Financial life	Taxes and Money	10.14
News	Polemic, Grossness, Debate, Problem and news	10.14
Professional life	Work	7.24
Health and Sport	Health, Sport and Nutrition	5.80
Art	Tastes, Colours and Poetry	2.90
Holidays and Travel	Holiday departure	1.45

Table 2.6: Subjects ranked by decreasing order of avoidance on social networks.

freedoms (version 1978). The Social Security number is implicitly considered sensitive in France according to the same law since unlike other countries, it is not random and it carries information

about marital status, date and place of birth. The CNIL adopts the definition proposed by Article 8 of the 1978 Law on Computers, Files and Freedoms (2004 version). Sensitive data are "*personal data that directly or indirectly reveal racial or ethnic origins, political, philosophical or religious opinions or the trade-union membership of persons, or relating to the health or sexual life of those persons*"[CNIL, 2010]. According to Google definition the sensitive data are "*relating to confidential medical facts, racial or ethnic origins, political or religious beliefs or sexuality*" [Google, 2018]. However, the Court of Justice of the European Union insists that search engines are not obliged to remove these information from the search results if the person in question does not address such a request to them [CURIA, 2014].

In the following, we define the type of discussed subjects on social media. To this end, we use the notations presented in Table 2.7.

Symbols	**Significations**
S	Set of discussed subjects on social media
x	a variable that designates a given subject
M_a	a constant that refers to the mean rate of anonymous publication on forums and websites.
A	a function in S of arity 1. $(A(x) > M_a)$ is true if the anonymity rate of the subject x is greater than M_a.
$\theta_{networks}$	a constant threshold that refers to the average rate of discussion of subjects on social networks minus the standard deviation.
θ_{forums}	a constant threshold that refers to the average rate of discussion of subjects on forums and websites minus the standard deviation.
$D_{networks}$	a function in S of arity 1. $(D_{networks}(x) < \theta_{networks})$ is true if the discussion rate of the subject x on social networks is less than $\theta_{networks}$.
D_{forums}	a function in S of arity 1. $(D_{forums}(x) < \theta_{forums})$ is true if the discussion rate of the subject x on forums and websites is less than θ_{forums}.
$Delicate$	a relation in S of arity 1. $Delicate(x)$ is true if the subject x is delicate.
$Thorny$	a relation in S of arity 1. $Thorny(x)$ is true if the subject x is thorny on at least one social media(forums and websites or social network).
$Avoid$	a relation in S of arity 1. $Avoid(x)$ is true if the subject x is avoided by at least 1 user.
$Controversial$	a relation in S of arity 1. $Controversial(x)$ is true if the subject x is controversial.
$Sensitive$	a relation in S of arity 1. $Sensitive(x)$ is true if the subject x is sensitive.

Table 2.7: Notations.

Delicate subjects. A discussed subject on social media is **delicate**, if and only if, it is avoided on these media **or** [4] whose rate of anonymous publication on the forums and websites is above average. In other words:

$$\forall x(x \in S \Rightarrow (Delicate(x) \iff Avoid(x) \vee A(x) > M_a)) \tag{2.2}$$

Thorny subjects. A discussed subject on social media is **thorny**, if and only if, it is delicate **and** whose discussion rate on the forums/websites **or**[5] social networks are below the threshold of the mean of all discussions minus the standard deviation on that media, respectively θ_{forums}

[4]Or indicates an inclusive disjunction.
[5]Or indicates an inclusive disjunction.

and $\theta_{networks}$. In other words :

$$\forall x(x \in S \Rightarrow (Thorny(x) \iff Delicat(x) \wedge (D_{networks}(x) < \theta_{networks} \vee D_{forums}(x) < \theta_{forums}))) \tag{2.3}$$

Controversial subjects. A discussed subject on social media is **controversial**, if and only if, it is avoided on these media **and** whose rate of anonymous publication on forums and websites is above average. In other words :

$$\forall x(x \in S \Rightarrow (Controversial(x) \iff Avoid(x) \wedge A(x) > M_a) \tag{2.4}$$

Sensitive subjects. Based on the delicate, thorny and controversial subject definitions we define a sensitive subject in the following. A discussed subject on social media is **sensitive**, if and only if, it is delicate and thorny **or** [6] controversial. As thorny and controversial subjects are also delicate, the definition of sensitive subject is simplified as follows: A discussed subject on social media is **sensitive**, if and only if, it is thorny **or** controversial. In other words :

$$\forall x(x \in S \Rightarrow (Sensitive(x) \iff Thorny(x) \vee Controversial(x))) \tag{2.5}$$

Subjects	**Delicate**	**Thorny**	**Controversial**	**Sensitive**
Money	×	×	×	×
Religion	×	×	×	×
Shopping	×	×	×	×
Dating	×	×		×
Health	×		×	×
Politics	×		×	×
Family	×			
News	×			
Work	×			
Travel	×			
Sport	×			
Art	×			
Games	×			
Kitchen	×			
Fashion				
TV shows				
Study				
Technology				
Music	×	**Insufficient information**		
Movie	×			
Humour	×			
Book	×			
House	×			
Tip	×			
Sex	×			

Table 2.8: Sensitive subjects.

[6] or indicates an inclusive disjunction.

In our survey, the mean anonymous publication on forums and websites is "$M_a = 8.71\%$". The discussion thresholds on social networks and forums are respectively $\theta_{networks} = 7.53$ % and $\theta_{forums} = 33.47$ %. Table 2.8 identifies the sensitive subjects among the studied subjects in our survey, namely: "*Money, Religion, Shopping, Dating, Health and Politics*".

Sensitivity measurement

In this section, we use a normalization function to adjust the sets of values for sensitive subjects calculated in Tables 2.3, 2.4, 2.5 and 2.6. This function converts values to a common standard to make them comparable. It refers to the data transformation by dividing each value by the mean of the set.

Matrix M (6×4) summarizes the calculated percentages in the previous section for the sensitive subjects (see Matrix 2.6). The lines represent the sensitive subjects and the columns represent the four sensitivity criteria presented in the previous section. The normalization function is defined by Equation 2.7, where x_{ij} denotes the value on the line i and the column j. Matrix M' summarizes the normalized values of M (see Matrix 2.8).

$$M = \begin{bmatrix} & \textit{Rate of discussions on social networks} & \textit{Rate of discussions on forums and websites} & \textit{Rate of anonymity} & \textit{Rate of avoidance} \\ \textit{Money} & 0.94 & 54.42 & 25 & 10.14 \\ \textit{Religion} & 5.63 & 26.05 & 14.28 & 33.33 \\ \textit{Shopping} & 1.88 & 66.05 & 9.09 & 0 \\ \textit{Dating} & 5.16 & 24.19 & 0 & 21.74 \\ \textit{Health} & 17.37 & 66.05 & 9.09 & 5.8 \\ \textit{Politics} & 25.82 & 54.42 & 25 & 50.72 \end{bmatrix} \tag{2.6}$$

$$\begin{aligned} f(x_{ij}) &= \frac{x_{ij}}{\lambda_j} \\ \lambda_j &= \sum_{i=1}^{n} x_{ij} \end{aligned} \tag{2.7}$$

$$M' = \begin{bmatrix} & \textit{Rate of discussions on social networks} & \textit{Rate of discussions on forums and websites} & \textit{Rate of anonymity} & \textit{Rate of avoidance} \\ \textit{Money} & 0.02 & 0.19 & 0.30 & 0.08 \\ \textit{Religion} & 0.1 & 0.09 & 0.17 & 0.27 \\ \textit{Shopping} & 0.03 & 0.23 & 0.11 & 0 \\ \textit{Dating} & 0.09 & 0.08 & 0 & 0.18 \\ \textit{Health} & 0.31 & 0.23 & 0.11 & 0.05 \\ \textit{Politics} & 0.45 & 0.19 & 0.30 & 0.42 \end{bmatrix} \tag{2.8}$$

Given a subject, the less it is discussed on social media, the more sensitive it is. Thus, we define the *coefficient of sensitivity* C (see equation 2.9) oppositely to the rate of discussion on social media.

$$C(x_i) = (1 - x_{i1}) + (1 - x_{i2}) + x_{i3} + x_{i4} \tag{2.9}$$

Table 2.9 sorts subjects by descending order from the most sensitive to the least sensitive on social media.

In what follows, we discuss the statistics about sensitive subjects in our survey.

Subjects x	Coefficient of sensitivity $C(x)$
Religion	2.25
Money	2.18
Politics	2.08
Dating	2.00
Shopping	1.85
Health	1.63

Table 2.9: Descending order of sensitive subjects.

Religion. Religion is the fourth least discussed topic on social networks. Only 5.63% of participant Net surfers talk about it on social networks. It is also the second most-avoided topic with 33.33% of respondents not addressing it. About 26.05% of participant Net surfers visit forums and sites of religion. Only 25% of them make publications. The anonymous publication rate on these sites is about 14.28%. In addition, only 8.22% of men and 4.29% of women speak about religion on social networks and 30.86% of men visit forums and websites about religion compared to only 23.13% of women.

Money. Money is the least discussed topic on social networks. Financial subjects are avoided by 10.14% of the participants. The anonymous posting rate on the forums of " *Economy, Politics and News*" is 25%. Participant Net surfers who talk about money on social networks have the highest average age, i.e. 42.5 years. Their favourite networks are Facebook and Youtube. They have small and restricted friends lists, do not accept strangers, and use privacy settings when sharing content. They are familiar with several technologies and spend between 7 and 14 hours a week on social networks. However they are used to keep the same credentials and the same email addresses on different networks and they are active on the sites of "*Travel, Transport, Holidays and Insurance*" and " *Work, Study and Business*".

Politics. Politics is the most avoided subject by participant Net surfers. Indeed, 50.72% of participants avoid political discussions on social networks. However, 54.42 % visit forums and websites of "*Economy, Politics and News*" and 25.82% discuss political topics on social networks. Although it is avoided by more than half of participant Net surfers, politics is one of the most discussed subject on social media with a rate of discussion above the mean: 28.77% of men and 24.29% of women discuss it. In addition, 25% of the publications on the forums and sites of " *Economy, Politics, News and News*" are anonymous, where 76.07% of the activities on these sites are simply searches without identifications. The political orientation of Net surfers can be deduced through the type of information viewed on social media. More than 50% of participants that use social network discuss news. Besides, interactions, comments, "likes" and "shares" made by users on these media can be decisive to infer their political orientation.

Dating. Only 5.16% of participant Net surfers talk about dating on social networks. Men discuss this subject twice as much as women on these networks. More specifically, 34.78% of participant Net surfers who have not completed higher education visit "*Going out, Dating and Chat*" forums and websites, compared to only 22.91% of participant Net surfers who have completed higher education. In addition, 38.46% of participant Net surfers who visit these kind of websites and forums make non-anonymous publications.

Shopping. Only 1.88% of participant Net surfers talk about shopping on social networks. Women discuss this subject twice as much as men on these networks. Participant Net surfers who have not completed higher education discuss shopping eight times more than those who have completed higher education. It should be noted that 66.05% of participant Net surfers visit forums and websites of "*Health, Shopping, Cooking, Home and Tip*". But, 84.50% of activities are researches without publication. The anonymous posting rate on these forums and sites is about 9.09%.

Health. About 17.37% of participant Net surfers talk about health on social networks. Women discuss this topic twice as much as men, and 75.37% of women visit sites and forums of "*Health, Shopping, Cooking, Home and Tip*". The anonymous posting rate on these forums and sites is 9.09%. However, only 5.80% of participant Net surfers avoid this topic on social media.

Personal and family life. The personal and family life subjects are the third subjects avoided by participant Net surfers after politics and religion. About 21.74% of participants do not address it. Nevertheless, 31.45% of them discuss family topics on social networks. Those subjects are preferred by participant Net surfers who have not pursued higher education. 54.17% of them often address these subjects.

Sexual life. From all participants, 17.39% spontaneously mention that they avoid discussing sexuality on social networks. This subject is not included in the list of 18 proposed subject in question 35 (see Annex B). Additional statistical information are then needed to measure its sensitivity.

2.4 Possible attack vectors according to the behaviour of participants

In this section we discuss several scenarios of attack and information leakage. Each scenario exploits a vulnerability. Vulnerabilities are detected through the responses to the questionnaire.

Reversed parental monitoring

The responses of participant parents Net surfers show that only 1.44% of them do not know whether their children use social networks. Questions 30 and 31 focus on parental monitoring on social networks (see Annex B). These questions are only displayed to participants over 34-year-old and whose children use social networks. 71.05% of parents say that their children can view their publications against only 51.3% who can view their children's publication. In addition 5.26% of parents do not know if their publications are visible to their children. These results show that social networks have effective parental monitoring tools, but they are also within the reach of children to monitor their parents. About 76% of parents are vulnerable to the scenario detailed by Figure 2.6: Alice and Bob are friends on a social network; Alice has published a photo featuring Bob; Bob made a dirty joke about the photo in a comment; but he does not want his children to become aware of this photo or read his comment, and he does not know if they can do it. Although Alice is aware that her children can see all her publications, she did not realize Bob's inappropriate comment that is visible to her daughter too.

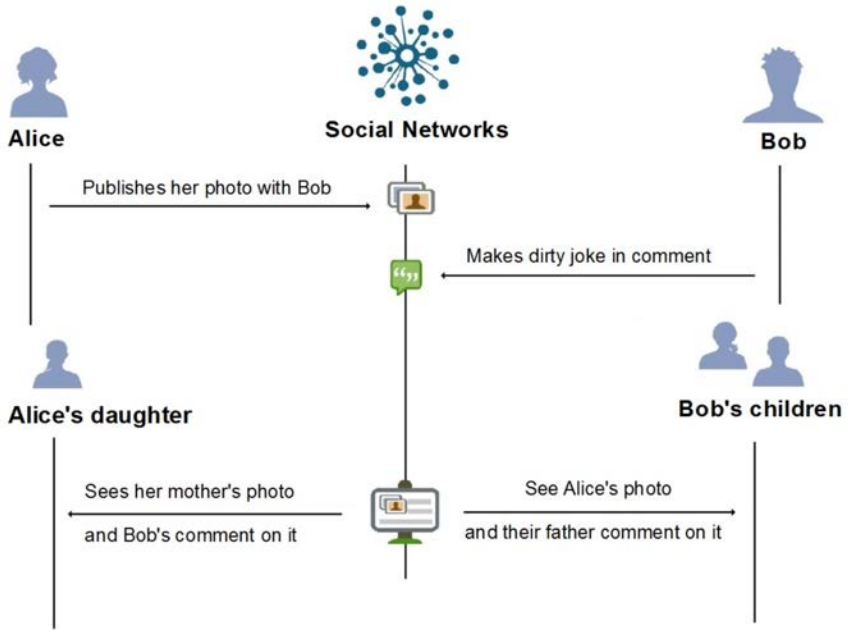


Figure 2.6: Example of reversed parental monitoring scenario.

Linking profiles across different social media

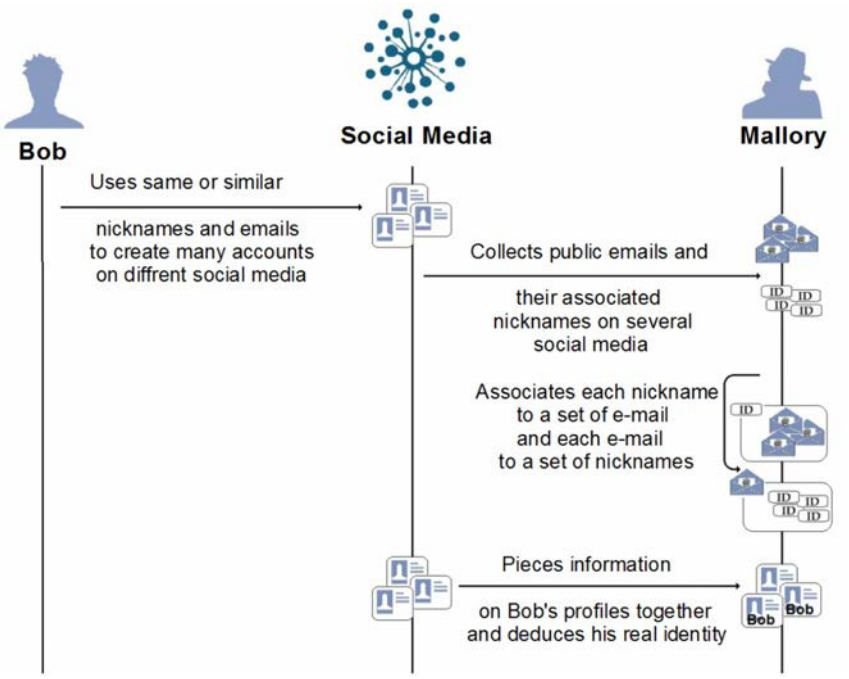


Figure 2.7: Example of scenario of attack of linking profiles across different social media.

Questions 18 and 19 are used to determine the probability of success of cross linking profile attack (see Annex B). The results of the answers to these questions show that 52.05% of participant Net surfers use the same e-mails to create different profiles, and 65.75% of them use the

same user-names or nicknames that are very similar on different social networks or websites (for instance, Mickey and Mickey.1). Hence, about 77.63% of users are vulnerable to cross linking profiles attacks by re-association of e-mail or nickname.

Figure 2.7 depicts an example of scenario of attack of linking profiles across different social media. Bob has several anonymous accounts on different social media (forums and social networks). He uses his accounts to talk about personal matters and seek the advice of specialists. He often uses nicknames or emails that he has already used to create other accounts on other social media. Since e-mails are often public on forums, Mallory was able to associate each nickname with a set of e-mail addresses and each e-mail address with a set of nicknames. He then discarded all profiles that do not belong to Bob and re-paired all of Bob's profiles. He gathers all information from various Bob's profiles to build a complete one and deduces his real identity.

Leakage of sensitive information

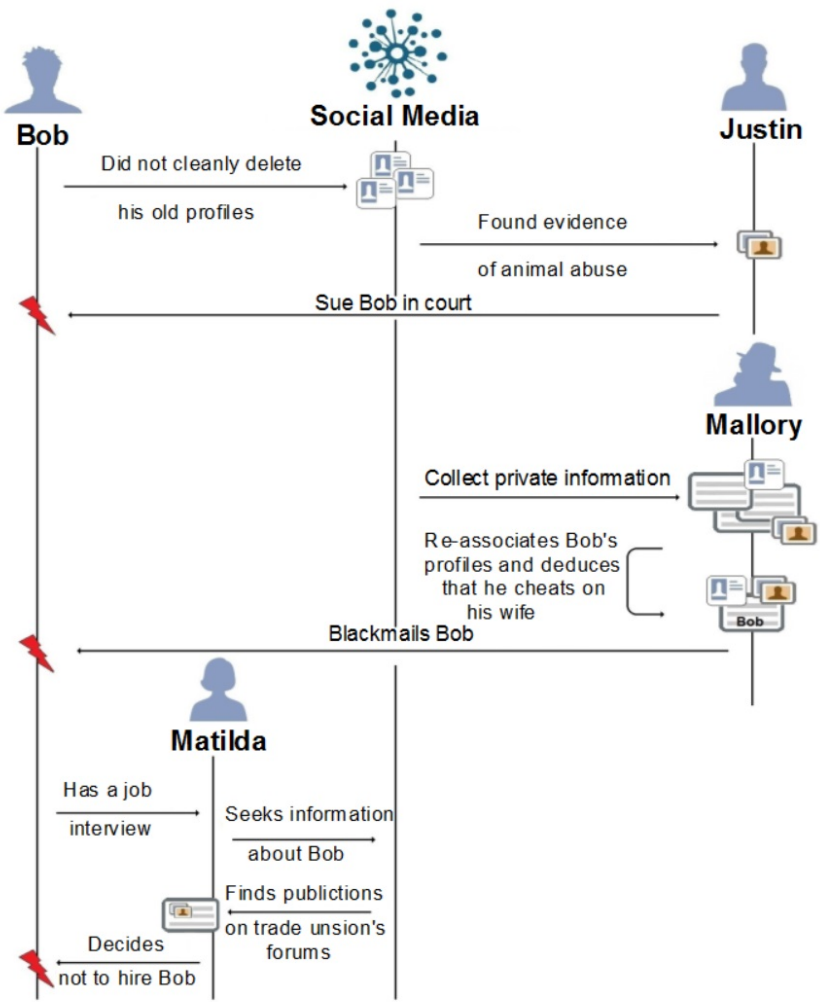


Figure 2.8: Examples of scenarios of sensitive information leakage.

To determine the probabilities of information leakage, we have analysed the answers to Questions 22, 23, 24, 27, 28 and 29 (see Annex B). The results confirm the inability of users to

manage information shared on social networks. Only 10% of participant Net surfers that use social networks do not have the same friends over different networks. On the other hand, 35% of participant Net surfers confirm that the majority of their friends are the same on different social networks. Moreover, 72.16% of participant Net surfers that are active on forums or websites confirm that they have created profiles that they do not use any more and do not know if these profiles are cleanly deleted. Hence, 76.29% of participant Net surfers that use social media are vulnerable to leakage of sensitive information between networks through old badly deleted profiles or through friends.

Figure 2.8 details three scenarios of sensitive information leakage. Bob was very active on forums and social networks when he was young. He decided to reduce his activity when he started looking for work. But he did not cleanly delete the old accounts he had created in the past. His personal data continue to exist on the web and are at a few clicks from curious Net surfers.

Scenario 1. Justin is a lawyer. He is active in animal welfare associations. He decided to sue Bob after discovering photos and videos of acts of cruelty committed by Bob against an animal.

Scenario 2. Mallory has collected several sensitive information that are anonymously published by Bob. After performing a cross linking profiles attack, Mallory has revealed the real identity of the publisher. Then Mallory has deduced that Bob cheats on his wife. He used the collected evidences to blackmail him.

Scenario 3. Matilda is a recruiter. She decided to look for information about Bob on internet after a job interview with him. She noticed that he was active on trade union forums. As a consequence, she decided not to consider his application despite his skills.

Figure 2.9 details a scenario of sensitive information leakage between several social networks. Bob took a sick leave to spend holidays with Alice. He then shared some photos with her on a social network. Alice has republished these photos on other networks. Bob's employer decides to fire him after seeing the photos taken during his sick leave.

From the collected responses , we also observed the following behaviours:

- 56.34% of respondents do not separate their professional and personal lives by adding colleagues, classmates, family members and neighbours to the same friend list and using the same profile.
- 15.96% publish photos without asking the consent of people appearing in these photos.
- 8.45 % add strangers to their friend lists only because they have common friends.
- In a test, 6.10 % are not able to recognize a person added randomly to their friend lists.

These behaviours show that about 65.25% of participant Net surfers that are active on at least one social network are exposed to the risk of sensitive information leakage on the same social network.

Figure 2.10 details a scenario of sensitive information leakage on the same social network. Bob has added his employer to his friends list on his favourite social network. The latter discovers that his employee is very active on social media. Bob used to complain about the working conditions and criticize his superiors. These publications prompted his employer to fire him.

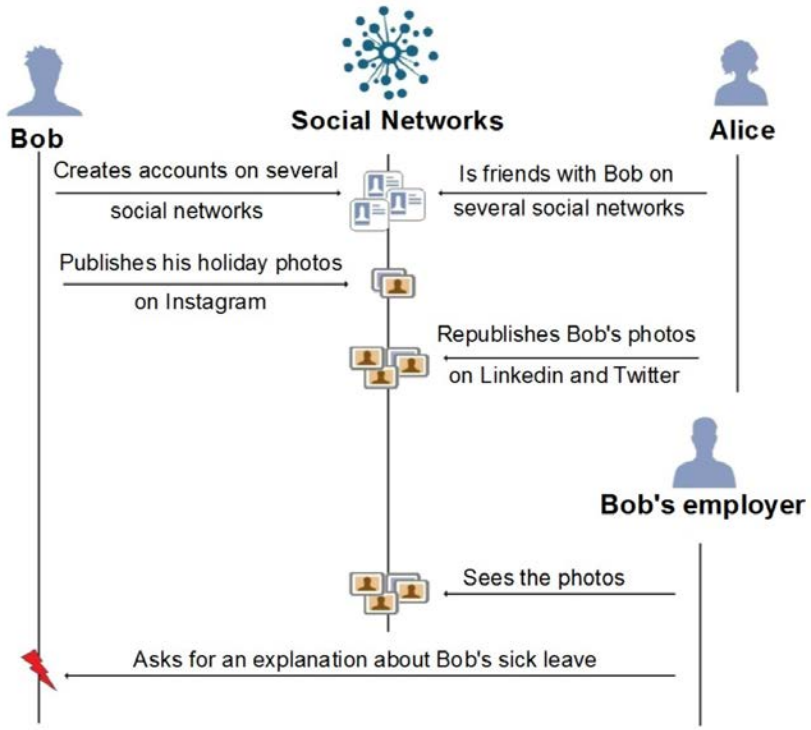


Figure 2.9: Scenario of leakage of sensitive information between several social networks.

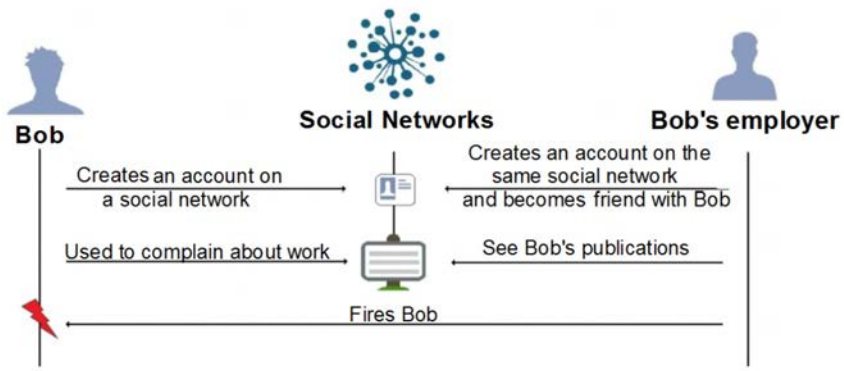


Figure 2.10: Scenario of leakage of sensitive information on the dame social network.

2.5 Conclusions

Our sample counts 232 social media users who provided valid and consistent answers. These users are located in 21 french regions and have followed more than 18 different disciplines of formal or social studies.

The coefficient of variation of the ages of the participants is 31.6%, and the correlation between the age distribution of the sample and the mother population is greater than 0.8. The margin of error, e, for a confidence threshold, s, that is equal to 85% is 4.2%. Thus, we conclude that our sample is reliable and fairly representative of the social media user in France with regard to the variable "age".

We have classified the subjects discussed on social media according to four criteria: rate

of discussion on social networks, rate of discussion on forums and websites, rate of anonymous publication and avoided subjects. Based on those criteria, we have proposed a definition of sensitive subjects. Then, we have computed the sensitivity coefficient of the studied subjects in our surveys. Our sensitivity-based method can be reused in other statistical studies targeting other topics (such as sexuality).

Finally, we have analysed the behaviour of participant Net surfers to identify some vulnerabilities on privacy. We then presented attack scenarios. Our study shows that more than 70% of social media users are exposed to the risk of sensitive information leakage that is mainly due to clumsy use of social media and unawareness about privacy issues.

3

Disclosing friendship and group membership links

Contents

3.1 Introduction **39**
3.2 Modelling social network for on-line link disclosure attacks **40**
3.3 Problematics and objectives **40**
3.4 Social networks group properties **41**
3.5 Link disclosure attacks **44**
3.6 Conclusions **51**

3.1 Introduction

While on-line social networks have become an important channel for social interactions, they also raise ethical and privacy issues. A well known fact is that social networks leak information, that may be sensitive, about users. However, performing accurate real world on-line privacy attacks in a reasonable time frame remains a challenging task. In this chapter, we address the problem of rapidly disclosing many friendship links using only legitimate queries (i.e., queries and tools provided by the targeted social network). Our study sheds new light on the intrinsic relation between communities (usually represented as groups) and friendships between individuals. To develop an efficient attack we have analysed group distributions, densities and visibility parameters from a large sample of a social network. By effectively exploring the target group network, our proposed algorithm is able to perform group membership, friendship and mutual-friend attacks along a strategy that minimizes the number of queries. The results of attacks performed on active Facebook profiles show that 5 different friendship links are disclosed in average for each single legitimate query in the best case.

To put the rest of this Chapter into context, we start by modelling social network for link disclosure purposes in Section 3.2. Then, we define the problematics and objectives of link disclosure attacks on On-line Social Networks in Section 3.3. After that, we analyse groups distribution, densities and visibility parameters in Section 3.4. Those properties are then used to perform group uncovering attack, membership attack, friendship attack and mutual-friend attack as detailed in Section 3.5.

3.2 Modelling social network for on-line link disclosure attacks

A social network can be defined as a website that allows users to create personal pages in order to share information with their friends and acquaintances. These pages are usually called profiles and contain personal information. Profiles are connected to each other through friendship links that can be either symmetric or asymmetric, depending on the network's policy.

In order to mimic real (i.e., non-cybernetical) societal interactions, some social networks such as Facebook, Linkedin and Viadeo support the creation of groups besides the profile creation. Accordingly, social networks can be modelled by two types of graphs –(i) friendship graph and (ii) group membership graph– as depicted by Figure 3.1.

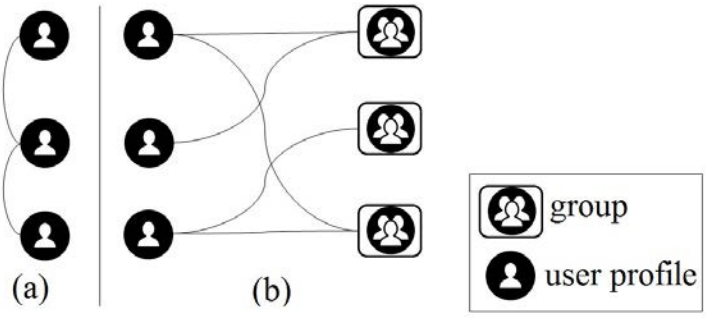


Figure 3.1: Social graphs : (a) unipartite friendship graph, (b) bipartite group membership graph.

The friendship graph (a) is unipartite and models the friendship links between users while membership graph (b) is bipartite and models the membership links between users and groups. Some of these links can be masked by users or group administrators. We call a friendship (resp. membership) attack a sequence of actions (e.g., queries) leading to disclose a masked friendship (resp. membership) link. Both kind of attacks are called link disclosure attacks. A mutual-friend attack discloses common friends to a target and other users. We call group uncovering attack a sequence of queries that disclose the membership network of the target and his acquaintances. In this work the attacker is limited to the usage of legitimate and minimal queries provided by the social networks APIs. Therefore the attacker model can be viewed as a passive one. We believe that these constraints are the cornerstones of successful real-world attacks that are difficult to detect because the traffic appears to be legitimate at first.

In [Dougnon et al., 2015] researchers propose a Partial Graph Profile Inference (PGPI) algorithm that exploit group memberships to infer profiles attributes. In [Zheleva and Getoor, 2009], relational learning approaches and group memberships are used to infer sensitive attribute of users such as locations.

3.3 Problematics and objectives

Problematics. In on-line attacks, the attacker is constrained by the network dynamicity and the time needed to scrap it. In fact, the dynamical network structure, with the addition/deletion of new links and nodes will ensure that the sampled graph does not reflect a real on-line social network at any given time. Therefore, crawling tasks for on-line attacks must be highly selective to collect only useful profiles and information and be as fast as possible.

For instance, [Elkabani and Khachfeh, 2015] show that homophilic attributes have significant influence on predicting friendship between users of Facebook. Thus, an attacker may be tempted to sweep the network for similar profiles to his target. He can also consider the friends of the target friends as potential friends and check these links. Although these general solutions may seem effective to gather many potential friends, they have major shortcomings. To understand these shortcomings let us recall the "six degrees of separation" phenomenon, that is the possibility to connect any two people in a maximum of six relationship steps. For example, the authors of [Bakhshandeh et al., 2011] show that the average degree of separation between Twitter users is 3.43 while the degree of separation on Facebook is between 2.9 and 4.2 for the majority of users [Bhagat et al., 2016]. Hence, considering friends of friends as potential friends is equivalent to considering at least tens of thousands users as potential friends for each single target [Ugander et al., 2011]. This is clearly impossible to handle and scale for real-world efficient attacks.

Objectives. Link disclosure attacks in on-line social networks aim to disclose hidden links by performing authorized requests.The attacks either reveal existing links or potential ones according to the employed method. We aim to disclose numerous links without having to verify a huge number of potential friends. In other words, we attempt to gather many potential friends but only those who have high probability to be friend with the target. The best way to achieve our objectives is to disclose the vicinity network of the target. To that end, we analyse groups' properties on on-line social networks since they reflect the way users are gathering within a network and uncover its structure. To keep our discussion simple, we aim to answer two questions in this work: Which groups leak useful information to meet previously detailed objectives? How to find and use them?

3.4 Social networks group properties

In this section we analyse some properties of Facebook groups. This analysis will guide crawling tasks in order to collect only data that leak more information about the target. Exploiting such data will increase the accuracy of link disclosure attacks and maximize the number of disclosed links. We stress that all experiments in this work were carried out on-line with real Facebook profiles. We have crawled 1,100 Facebook groups and all their members. Then, we have sorted the groups by declared size in sets. Each set contains at least 30 groups. Each group in the first set S_0 gathers between 2 and 10 members and each group in the set S_i gathers between $10i$ and $10(i+1)$ members.

Group distribution

We first study the distribution of groups in Facebook with regard to their sizes. We notice that the declared group size on this network is often different from the number of users published on the group member list. Moreover, crawling the same group using different IP addresses and accounts can result in slightly different listed members. This technique can reduce the gap between the two sizes by considering the union of all crawled member lists of the same group. However, it adds more complexity to attacks. To study groups distribution we have simulated a simple attack carried out using only one attacker node. All groups are crawled only once and we only rely on the declared group size to build the attack strategy.

Figure 3.2 shows that there are many more small groups on Facebook than larger ones. However, we notice the curve inflection for groups declaring between 30 and 70 members. By

checking these groups members lists we notice huge gaps between declared sizes and the numbers of listed members. Gaps reach 85% for some groups. Some groups are declared to have 60 members or more but they actually display less than 20 members on their members lists. These gaps can be explained by the fact that users unceasingly leave and join the group but size updates are not performed instantaneously. Henceforth, densities of such groups can increase if real sizes decrease since the less connected members are usually the first ones to leave the groups.

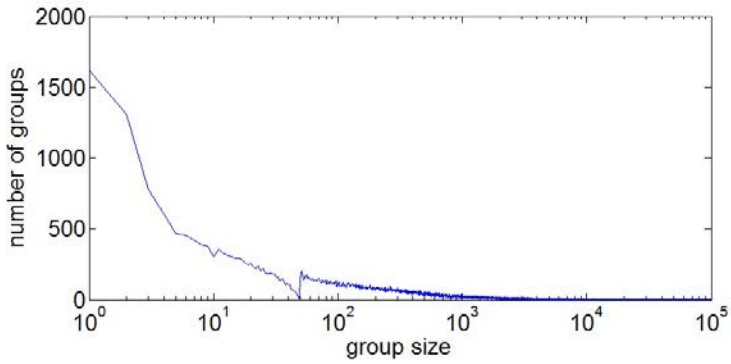


Figure 3.2: Group distribution of a sample of 14,517 Facebook users.

The result of our tests carried on 14,517 Facebook profiles shows that the probability of a given Facebook user to join at least one group gathering less than 50 members and publish his membership to it is 0.49. Thus, about half of analysed Facebook profiles are exposed to the danger of friendship link disclosure through groups they join and that gathers less than 50 members.

Group densities

In an undirected social graph, a friendship link between two user is considered public if at least one of them publishes it. It is considered hidden only if both users hide it.

In order to guide a strategy for disclosing hidden social links we first try to evaluate the probability that two members of a group are friends. We define three notions of group densities: public density, real density and maximal density, that we will use to estimate the number of friends that can be disclosed through link disclosure attacks. Given a group g, $PD(g)$ stands for its Public Density, $RD(g)$ stands for its Real Density and $MD(g)$ stands for its Maximal Density. The public density of g is the ratio of published friendship links between its members to all possible friendship links between them. It is defined by Equation (3.1) where $|g|$ is the number of members of g:

$$PD(g) = \frac{2}{|g|(|g|-1)} \sum_{\{m,m'\} \subseteq g} publicLink(m, m') \tag{3.1}$$

The real density of g is the ratio of all (public and hidden) friendship links between its members to all possible friendship links between them. It is greater or equal to the public density. It is defined by Equation (3.2)

$$RD(g) = PD(g) + \frac{2}{|g|(|g|-1)} \sum_{\{m,m'\} \subseteq g} hiddenLink(m, m') \tag{3.2}$$

The maximal density of g can be met only if all its members who hide their friend lists are friend with each other. It is greater or equal to the real density. It is defined by Equation (3.3) where p is the percentage of members who hide their friend lists among the members of g.

$$MD(g) = PD(g) + \frac{p^2|g| - p}{|g| - 1} \tag{3.3}$$

Thus we have:

$$PD(g) \leq RD(g) \leq MD(g) \tag{3.4}$$

Test results show that among 14,517 crawled Facebook profiles only 6,249 (43%) hide their friend lists or choose to reveal them only to their direct friends, friends of friends or some selected users. The rest (57%) leave the visibility setting by default and publish their friend lists. Hence, p can be considered equal to 0.43 if it is unknown by the attacker. Note that the attacker can easily verify the friend list visibility parameters of other users through the following Facebook request:

$$/ < nid_u > /friends \tag{3.5}$$

where nid_u is the numeric id [7] of the User u. In fact, this request returns the friend list of the User u if and only if he publishes it.

Figure 3.3 (a) shows that group densities decrease as the declared size of the group increases. It can be noticed that one can even estimate a given group density only from its declared size. This information is precious as it determines the number of links that can be disclosed between group members. In fact, the group real density can be viewed as the probability of the friendship link between a given member and another member from the same group. Hence, if the attacker discloses group membership of his Target t to a Group g, then all other members of g can be considered as potential friends of t with a probability in interval $[PD(g); MD(g)]$. Knowing the declared size of g, $PD(g)$ can be directly deduced from Figure 3.3 (a) and $MD(g)$ can be deduced from Equation (3.3). For instance, the average public density of groups gathering between 10 and 20 members is 0.343. Then, according to Equation (3.3) the real density of such groups belongs to interval $[0.343; 0.515]$ for p equal to 0.43. Expressively, the estimated accuracy of link disclosure attack is 0.343 and all the members of corresponding groups can be considered as potential friends with probability in $[0.343; 0.515]$.

Although popular groups gather many members, probabilities of friendship between them are very low. Crawling such groups is fruitful to seek a lot of potential friends of the target but with low probabilities. However, minute groups open small horizon for potential friends but with higher probability of friendship.

The relationship status between two members of a group g is a binary variable. Hence, assuming independence of friendship links in a first approximation, the expected number of published friendship links between a given member and all other members of the same group is the expectation of a binomial distribution of parameters $B(|g|, PD(g))$ which is $|g| \times PD(g)$. For example, Figure 3.3 (a) shows that the expected public density of groups gathering less than 11 members is greater than 35%. Hence, the expected number of friends of a target within a group he joins and that gather 6 members is 2 (since $0.35 \times 6 = 2.1$).

Figure 3.3(b) shows that the expected number of disclosed links between the target and group members slightly increases as the declared size of groups increases. Note that x-axis unit correspond to 10 members and y-axis unit correspond to 1 friendship link.

[7] Numeric id can be acquired through meta-data within an HTML code of the profile

Group visibility parameters

Groups and members can independently choose to publish or hide the membership relation. For instance Facebook users can choose to mask some groups from their list of groups. On the other hand, the administrators of groups can independently publish the entire lists of members. With that in mind, an attacker can build an attack strategy to disclose the groups that are masked by users or the membership lists of secret groups.

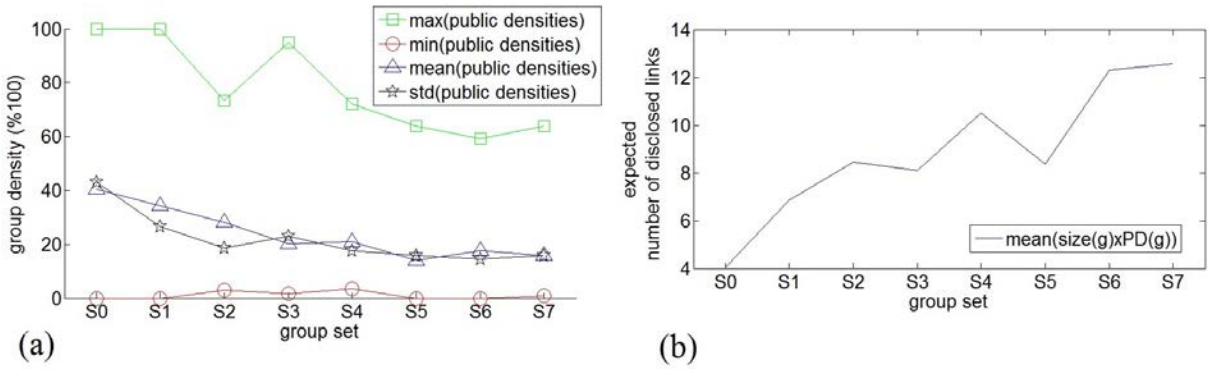


Figure 3.3: Results of analysis: (a) Variation of public density with respect to group declared size, (b) Expected number of disclosed links between the target and group members.

3.5 Link disclosure attacks

In this section we perform details four attacks: (i) group uncovering attacks, (ii) friendship attack, (iii) group membership attack and (iv) mutual friend attack.

Group uncovering attacks

In this section we exploit groups properties detailed in previous section to perform group uncovering attacks. To that end, we define real and public n-hop distant groups.

Real n-hop distant groups. Given a target t that joins Group g, g is considered as a real 1-hop distant group from t (denoted by $g \in RG_1(t)$) and all its members m are considered as real 1-hop distant members from t (denoted by $m \in RM_1(t)$). We define inductively $g \in RG_n(t)$ iff $g \notin RG_{n-1}(t)$ and there is $g' \in RG_{n-1}(t)$ with a non-empty intersection with g. For all m in $g \setminus RM_{n-1}(t)$ we have by definition $m \in RM_n(t)$. We can show the following symmetry rule:

$$u1 \in RM_n(u2) \iff u2 \in RM_n(u1) \tag{3.6}$$

where $u1$ and $u2$ are two different users. Figure 3.4 depicts an example of a real 3-hop distant group from the target node t.

Group $g1$ is a real 1-hop distant group from t. Consequently, all its members are real 1-hop distant members from t. Members $m6$, $m7$, $m8$, $m9$ and $m10$ are real 2-hop distant members from t since they join the same Group $g2$ as $m5$ who is real 1-hop distant members from t. Finally, $m11$, $m12$, $m13$ and $m14$ are real 3-hop distant members from t as $m9$ and $m10$ join their Group $g3$. Members $m5$, $m10$ and $m9$ act as gateway between groups.

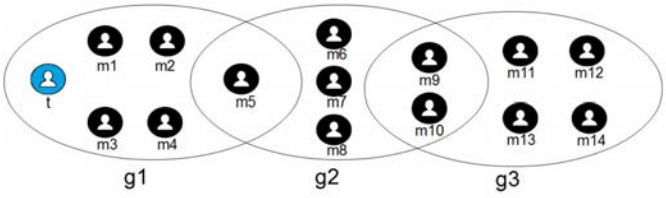


Figure 3.4: g3 is a real 3-hop distant group from t.

Public n-hop distant groups. Users can mask their membership to groups and groups can hide their members lists. Consequently, the public n-hop distant relation does not satisfy the symmetry rule.

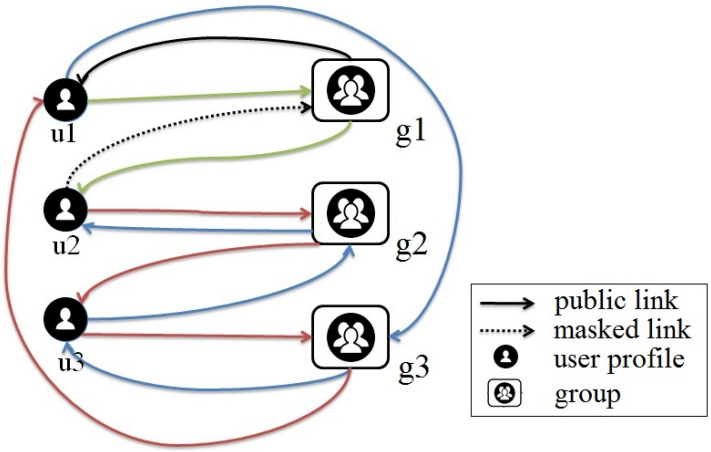


Figure 3.5: An example of public n-hop distant groups and members.

Figure 3.5 depicts an example of different public n-hop distant groups and members between two users. Arrows from user to groups stand for membership links while arrows on the opposite direction represent group members lists. Dotted lines represent masked links and solid lines represent public links. While both Users $u1$ and $u2$ join the same group $g1$, only $u1$ publishes his membership to $g1$. User $u2$ publishes only his membership to Group $g2$. User $u3$ acts as a gateway between $g2$ and $g3$ and publishes his membership to both of them. All groups $g1$, $g2$ and $g3$ publish their member lists. There are two public paths from User $u1$ to User $u2$. The first one, the green path, goes through $g1$ and is the shortest one with only one hop. The second one, the blue path, is two hops long. It goes through $g3$ then $g2$. Hence, $u2$ is a public 1-hop distant member from $u1$. On the other hand, there is only one public path, the red path, from $u2$ to $u1$ that goes through $g2$ then $g3$. Thus, $u1$ is a public 2-hop distant member from $u2$.

Social graph traversal algorithm. Let $u2$ be a target user who is friend with both users $u1$ and $u3$ and hides his friend list. Since he publishes his membership to $g2$, the attacker can reach $u3$ through $g2$ member list. If $u3$ publishes his friend list, the attacker can easily disclose the friendship link between $u3$ and $u2$ by checking the friend list of $u3$. Likewise, the attacker can reach $u1$ through $g3$ member list if $u3$ publishes his membership to that group nad he can search for $u2$ in $u1$ public friend list. Furthermore, next hop lead to $g1$ and hence the attacker can disclose group membership links between $u2$ and $g1$ by checking $g1$ public member list. Algorithm 1 gives more details about the graph traversal steps. The algorithm outputs are two

sets of groups and members. Its inputs are the number of hops and a set of seed groups of the target.

```
Data: gps: set of groups, h: number of hops
Result: d_m: set of distant members, d_g: set of distant groups
Procedure explore(gps, h, d_g, d_m)
    if (h > 0) then
        for each g ∈ gps do
            members.addAll(getMembers(g)) ;
        end
        members.removeAll(d_m);
        for each m ∈ members do
            groups.addAll(getPublicGroups(m)) ;
        end
        groups.removeAll(d_g);
        d_g.addAll(groups);
        d_m.addAll(members);
        explore(groups, h − 1, d_g, d_m);
    end
Return()
```

Algorithm 1: Groups uncovering attack through social graph traversal

To collect seed groups, the attacker can directly retrieve unmasked groups from his target profile. We note that among 14,517 attacked Facebook profiles 11,446 (78.84%) do not change group visibility parameters and publish their groups membership even to secret groups. Otherwise, if the target masks all his groups and attributes, the attacker can create a fake virgin profile, use it to only visit his target profile, send him friendship request and try to interact with him by liking and commenting his posts or sending him messages. Then, link prediction algorithms of the social network [Barbieri et al., 2014] will start suggesting groups and attributes to the attacker that are strongly related to his target. Hence, he can use the suggested groups as seeds or take advantage of network research features and uses suggested attributes to look for seed groups. For instance, one of this paper author hides all his attributes on Facebook. However, the social network suggested his home town and 10% of his friends to a newly created profile that he added as a friend.

By following Algorithm 1 steps the attacker can effectively crawl his target group network and avoid loops. However, some social networks do not allow robots to crawl their network. For instance, Facebook bans robot accounts for a week. To overcome this issue, we used many users accounts. Our robot is able to change IP adresses, simulate human behaviour, switch between accounts, manage connection loss and save data in XML format and SQL database to avoid loops and replay attacks offline.

In the following, we exploit the group uncovering attack to perform link disclosure attacks. We aim to disclose two types of link: friendship between users and membership between users and groups.

Friendship and membership attacks

The attacker can explore the group networks of his target then check the member lists of distant groups to disclose group membership links to the masked groups. However, results show that

less than 0.1 group membership in average can be disclosed by this attack. This can be explained by the fact that 78.84% of attacked profiles do not change group visibility parameters and even publish their memberships to secret groups. On the other hand, by exploring groups networks of 14,517 profiles we disclosed 430 different secret groups and 756 of their members. Secret groups can help to disclose communities if their member lists are disclosed. Moreover, their members can be taken into consideration to compute the probability of friendship between two users who hide their friend lists.

In this work we aim to disclose friendship links with certainty. In undirected social networks it is sufficient but not necessary that one of the two friends publishes his friend list to disclose the friendship link between them with certainty. In this perspective, an attacker can query all friend lists of the distant groups members of the target and check if he is listed in public ones. Opportunely, some social networks afford features that can be used to rapidly check friendships between users. For instance, friendship between two users of Facebook can be easily checked through the following PHP request (3.7):

$$/friendship/ < nid_1 > / < nid_2 > \tag{3.7}$$

$< nid_1 >$ and $< nid_2 >$ are numeric IDs of two different users. In fact, the request (3.7) returns the date of the link creation between two users if and only if there is a friendship link between them and at least one of them publishes his friend list. Taking advantage of this feature, attacker can easily follow Algorithm 2 to disclose both friendship and group membership links of his target. Algorithm inputs are the profile of the target, the number of hops and the minimum number of links to disclose.

```
Data: t: target profile, h: number of hops, th: disclosed link threshold
Result: d_f: set of disclosed friends, di_g: set of disclosed groups
seedGroups ← getSeedGroups(t);
sizeSort(seedGroups);                          ▷ list of set of groups sorted by size
while d_f.size() < th & seedGroups.length()>0 do
    d_m2.addAll(d_m);                          ▷ d_m2 contains all tested profiles
    d_g.clear(); d_m.clear();
    explore(seedGroups.pop(), h, d_g, d_m);    ▷ see algorithm 1
    d_m.removeAll(d_m2);                       ▷ remove already tested profiles
    for each m ∈ d_m do
        if friendship(m, t) then
            d_f.add(m);
        end
    end
                                               ▷ all newly explored groups are not tested yet
    for each g ∈ d_g do
        if getMembers(g).contains(t) then
            di_g.add(g);
        end
    end
end
```

Algorithm 2: Friendships and group membership attacks based on k-hop group graph traversal

We have attacked more than 100 active Facebook profiles that hide their friend lists from

each set detailed in Section 3.4. For each attack we only checked the groups belonging to the same set to disclose friendship links between the target and those groups members. Note that users can be members of many groups from the same set. Since tiny groups densities are higher than large ones, fewer requests are required to disclose friendship links with certainty between the former members than between the latter members.

1-hop attack results (Figure 3.9 (a), blue curve) show that the average number of required requests to disclose one link with certainty increases as the size of groups increases. Only 6 requests in average are sufficient to disclose a friendship link with certainty of a target joining groups gathering less than 40 members against more than 7 requests in average for larger groups.

However, the average number of requests to disclose one friendship link decreases if attacks involve 2-hop distant groups from the target. This does not mean that the ratio of published friendship links (PFLs), between the target and 2-hop distant groups members from him, is higher than the ratio of PFLs between the target and 1-hop distant groups members from him. But, the ratio of PFLs between the target and the union of both 1-hop and 2-hop distant groups members from him is higher than any of the two ratios.

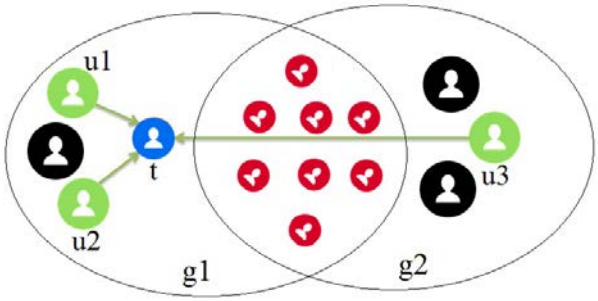


Figure 3.6: 2-hop friendship disclosure attack.

Observations. Figure 3.6 gives an illustration of an observed social phenomena. Users within the same network tend to crowd in small and highly overlapping groups. Thereby, small networks pop up within big networks. To put it in another way, some members joining the same group (e.g., $g1$) decide to create a new group (e.g., $g2$) of similar size and to add some of their acquaintances to it. Doing so they act as gateways between both groups (inclined nodes in Figure 3.6). Some newly added members to the latter group (e.g., $u3$) publish their friendship links to the former group members. Therefore, the ratio of published friendship links between the target t and all members of the two merged groups (e.g., 3/14 for $g1 \cup g2$) is greater than the ratio of published friendship links between him and any of the two groups taken alone (e.g., 2/11 for $g1$ and 1/11 for $g2$). Consequently, the average number of requests to disclose one friendship link decreases as well as the number of disclosed links increases.

However, Figure 3.9 (a) shows that 3-hop attacks are less effective than 2-hop attacks. This result can be explained by the fact that the ratio of members publishing their friendship to the target among 3-hop distant groups is low. On the one hand, crawling those groups may orient the attack toward adjacent networks and dramatically increase the number of requests to disclose one link in average. On the other hand, it may disclose masked groups of the target. With this in mind, attackers can perform 3-hop or above attacks to only disclose masked groups of the target by checking public member lists then perform 2-hop attacks to disclose friendship links. Moreover, they can reduce the size of attacked groups after each hop to avoid crawling adjacent networks. Thus, they can effectively uncover the group network of the target and minimize the number of requests to disclose friendship links.

Mutual-friend attacks

The term 'mutual friends' stands for friends in common between two users. Mutual-friend attacks are performed between the target who hides his friend list and another user to disclose a list of friends in common between them. In this section we exploit group uncovering attacks to perform mutual-friend attacks [Jin et al., 2013] between two members of the same network. Attacker can take advantage of the features afforded by social networks in order to list public mutual friends of two users. For instance, mutual friends of two Facebook users can be rapidly listed through the following Facebook request (3.8):

$$/browse/mutual_friends/?uid =< nid_1 > \&node =< nid_2 > \tag{3.8}$$

$< nid_1 >$ and $< nid_2 >$ are the numeric IDs of two different users. Thus, the attacker can follow Algorithm 2 steps while replacing lines from 8 to 17 by the function described by Algorithm 3 to disclose mutual-friend links between his target and other users. Similarly to Algorithm 2, this algorithm inputs are the target profile, the number of hops and the minimum number of links to disclose. But it discloses mutual friends between the target and the groupe members rather than friendships between them.

```
1 for each m ∈ d_m do
2 |   d_f.addAll(mutualFriends(m, t));
3 end
```

Algorithm 3: Mutual friend attack

In fact, a mutual-friend request (3.8) between two users returns the list of their mutual friends that publish their friend list if and only if at least one of the two given users publishes his friend list as well. Starting from the hypothesis that a mutual-friend attack is performed between the target who hides his friend list and another user, it is only successful if both the latter and the mutual friend publish their friend list. Moreover, it is not effective in the case of sparse networks since it does not disclose friendship link between two users that do not have mutual friends even if one of them publishes his friend list. The example depicted by Figure 3.7 shows that despite the fact that User $u1$ publishes his friend list, mutual-friend requests cannot disclose the friendship link between him and the target t. Dotted arrows represent masked links and solid

Figure 3.7: Undisclosed links by mutual-friend attack.

ones represent public links. In this example only User $u1$ publishes his friend list and both User $u2$ and the target t hide theirs. Hence, the results of all possible mutual-friend requests between Users $u1$, $u2$ and t are empty since two of them hide their friend list. However, friendship requests can disclose the friendship links between the target t and User $u1$ and between Users $u1$ and $u2$. Figure 3.8 depicts the average number of undisclosed links by a mutual-friend attack but disclosed by a friendship attack. We notice that this number increases with the number of hops.

Having said that, mutual-friend attacks can disclose more friends than friendship attacks if the target shares many mutual friends with his distant members.

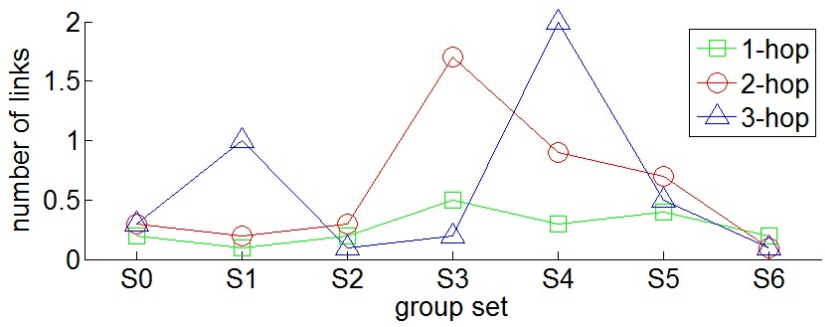


Figure 3.8: The average number of undisclosed links by mutual-friend attack but disclosed by friendship attack.

Figure 3.9 (b) shows that the number of mutual-friend requests to disclose one friendship link is quite similar for 1-hop and 2-hop attacks and increases for 3-hop attacks. However, it is far lower than the number of friendship requests depicted by Figure 3.9 (a) as mutual-friend request returns a list of friends.

To get better results the attacker can combine both attacks. For instance to maximize the number of disclosed links, he can sequentially perform a friendship attack after a mutual-friend attack. Hence, the number of attack requests will be equal to $2n - d$ where n is the number of distant groups members and d is the number of disclosed links between the target and them by mutual-friend attacks. Besides, he can alternatively perform both attacks to disclose friendship links between the target and his distant groups members. He can then follow Algorithm 2 steps while replacing lines from 8 to 17 by Algorithm 4 in order to focus his attack on distant groups members. Thus, the number of attack requests will belong to interval $[2; 2n]$. In fact, if mutual-friend requests do not disclose any friendship links between the target and his distant groups members then the number of attack requests will be equal to $2n$, by adding n friendship requests and n mutual-friend requests. On the other hand, if the target network is highly connected and the first mutual-friend request between the target and one of his distant groups members returns the rest of distant groups members then the number of attack requests will be 2, namely one friendship request and only one mutual-friend request.

```
1  for each m ∈ d_m do
2      if (!d_f.contains(m)) then
3          if (friendship(m, t)) then
4              d_f.add(m);
5          end
6      end
7      d_f.addAll(mutualFriends(m, t));
8      if (d_f.containsAll(d_m)) then
9          break;
10     end
11 end
```

Algorithm 4: Mutual friend and friendship attacks

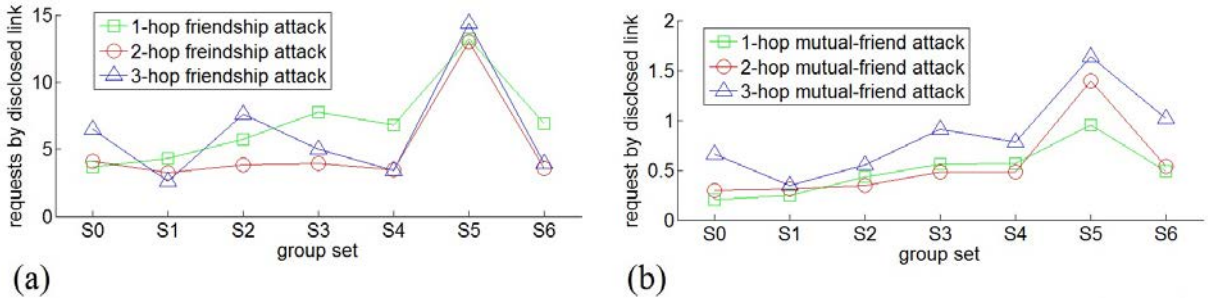


Figure 3.9: Results of attacks: (a) The average number of friendship request to disclose one friendship link, (b) The average number of mutual-friend request to disclose one friendship link

Dataset

We have performed on-line attacks on Facebook. We have crawled 14,517 profiles, 22,855 groups and 76,772 mutual-friend lists. The resulting graph contains 4,153,379 user nodes, 131,410 group nodes, 5,720,973 friendship links and 1,225,533 group membership links. We noticed that 78.84 % of crawled profiles do not mask their groups 56.95 % publish their friend lists and 47.77 % publish both. Among users who publish their friend list, the number of friends for a user in average is 530. Among all crawled profiles the number of unmasked groups for a user in average is 14.17. Figure 3.10 depicts the frequencies of published groups per user (a) and number of friends (b).

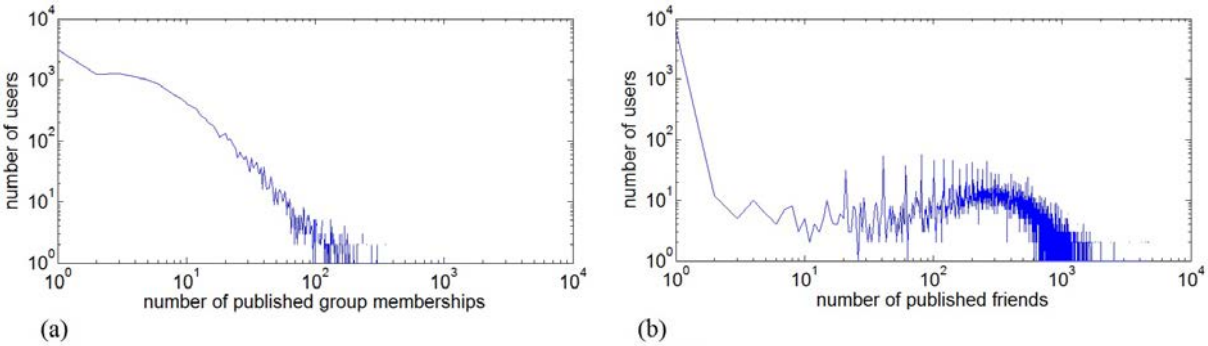


Figure 3.10: Sample of 14,517 Facebook profiles: (a) Frequency of published group membership, (b) Frequency of list of friends size.

3.6 Conclusions

Friendship links on social networks hold sensitive information about the community structure and affinity between users. Disclosing them can expose users to the highest danger of leaking personal sensitive information such as political orientation. In this Chapter we have tackled the problem of link disclosure with certainty. We have performed on-line attacks on active Facebook profiles and proved that attackers can easily and rapidly disclose many hidden links with certainty taking advantage of social network APIs.

4

Overview of our implemented prediction system

Contents

4.1 Introduction **53**
4.2 Architecture **54**
4.3 SONSAI user guide **56**
4.4 Examples of inference scenarios **59**
4.5 Conclusions **59**

4.1 Introduction

In order to benefit from the social power of social networks, users tend to be more active and share more content in a pursuit of fame, wealth, job or simply social interactions. However, they are unable to assess the risks of inferring sensitive information about themselves. Even well aware users who care about their privacy may be exposed to the risk of leaking sensitive information about themselves such as political views and sexual orientation based on seemingly harmless but correlated information such as favourite colours, musics and authors.

In this chapter we summarize several techniques we have designed and implemented in order to help social network users to evaluate the risk that a third-party infers values of their sensitive attributes. Attribute sensitivity is a subjective notion that may differ from one user to another. Some users may consider political views or ethnic origin sensitive while others consider them innocuous. We have identified in Chapter 2 through surveys the most sensitive attributes for a sample located in France.

Users must handle with care their publications concerning correlated attributes to the sensitive attribute in order to safeguard their privacies. For instance, if music is very correlated to politics in the social network then users must be careful about the musics they publish in order to keep their political views secret. Correlation might be complex to understand and/or unexpected for standard users. This is why we propose here a tool to help them managing their publications: once a user has defined his sensitive attribute that he want to be as much as possible hidden, the tool will check whether other published attributes give hints about this sensitive attribute value and which attributes are the most betraying. If this is the case the user can modify or delete this attribute in order to decrease or cancel the correlation. The tool is

designed to perform reasonably with the limited resources of a personal computer, by collecting and processing a relatively small relevant part of network data.

In the developed *SOcial Networks Sensitive Attribute Inference* system (SONSAI) that we detail in this part, a sensitive attribute is initially specified by the system user and chosen from the list of discovered attributes by the collector say a few hops around the user in the social network. Sensitive attribute can be pages of politicians or any other attribute judged sensitive by the user. SONSAI helps users simulate an attack of inferring sensitive information about themselves in order to assess their protection levels. Moreover, it helps the users understand where the threat comes from by generating a sorted list by importance of correlated attributes. SONSAI is composed of two main tools that respectively collect and analyse the data separately. All conclusions made by the analyser depends only on the collected data around the user. Thus, generated privacy rules are personal and specific for each user. In fact, by avoiding using general privacy rules we avoid making wrong conclusions about the privacy of users of different communities. Conclusions are more adapted to each specific user context. In Chapter 7 we compute a score to evaluate the risk of inferring values of a given sensitive attribute s based on the correlation of s and the values of attributes published by the user. We consider that the risk of disclosing values of sensitive attribute s is high if the score is higher than 65%. It is moderate if it is between 50% and 65%. It is considered to be a low risk if the score is lower than 50%.

In this chapter we detail the architecture of SONSAI to assess privacy leaks on social networks. Then we detail its functionalities.

4.2 Architecture

The architecture of SONSAI is detailed in Figure 4.1. SONSAI is composed of two main tools: a Collector and an Analyser. The Collector summarizes the functionalities of the Crawler. It is written in Java 8. and it counts about 5k lines of code. The Collector saves the collected data as XML files.

The Analyser encompasses the functionalities of the Anonymizer, Cleanser, Random Walker, Word2Vec and Ranker. It is programmed in about 2.5k lines of Python 2.7 code. The Analyser inputs are XML files generated by the Collector and its output includes (i) sensitive values ranked according to their proximity to the user profile, (ii) an evaluation of the user interest for the analysed sensitive attribute, (iii) an evaluation of the risk of inferring sensitive secret values and (iv) a list (sorted by importance) of attributes correlated to the analysed sensitive attribute.

The functionalities of the components of the Collector and Analyser are as follows:

Crawler. The crawler uses Web Browser Automation (WBA) to simulate human behaviour in order to explore the social network. It drives a Firefox 58.0b4 navigator through a Selenium 3.5.3 server [8]. Collected information from each profile, group and page are stored in separated XML files. The crawler algorithms are detailed in Chapter 5.

Anonymizer. The Anonymizer component parses all the XML files and generates anonymized graphs by replacing all the IDs by integers. Each graph models a different attribute. For instance the pages of politicians is modelled by the graph $G_p^{'} = (U, P, L)$. U is the set of user profiles, P is the set of pages of politicians and L is a set of links between user profiles and pages of politicians. Anonymized graphs are then sorted in separated TSV files. The Anonymizer algorithms are detailed in Chapter 5.

[8] http://www.seleniumhq.org/

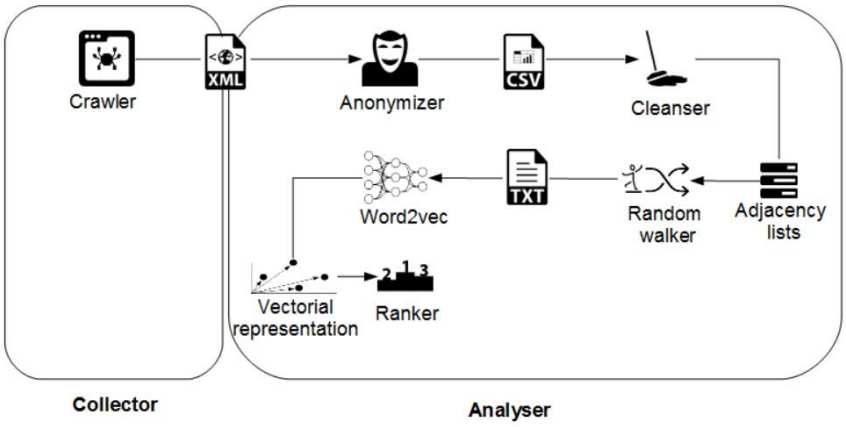


Figure 4.1: Architecture of SONSAI.

Cleanser. The Cleanser main objective is to select the best attributes (modelled by graphs) to help inferring user preferences about a given sensitive attribute. The Cleanser analyses only the structure of the anonymized graphs and does not rely on any semantic information. We have defined two techniques to select the best attributes depending on the properties of the sensitive attribute.

Another objective is to reduce whenever possible the number of attribute values by clustering similar values. The selected graph by the cleanser are processed and stored as adjacency lists. The Cleanser algorithms are detailed in Chapter 6

Random walker. The Random walker browses the adjacency lists and stores the steps of performed walks as word in a text document. We define two ways of performing random walks depending on the used technique to cleanse and select the best attributes for inference. The generated text document holds information about paths in the graphs and their frequencies. It will be processed to predict missing links between the target user profiles and the values of the sensitive attribute. The Random walker algorithms are detailed in Chapter 7.

Word2Vec. We use the Python gensim [9] implementation of Word2Vec to parse the text document and compute a vectorial representation of social network nodes encountered in the walks. Word2Vec is a Natural Language Processing (NLP) model that processes text documents in order to generate vectorial representation (embeddings) of vocabulary. It relies on a secession of related algorithms to compute the vectorial representation of each node. It uses a shallow neural network with one hidden layer to compute vectors and reduce their dimension. The Word2Vec algorithms are detailed in Chapter 7.

Ranker. The Ranker component classifies the sensitive nodes according to their similarity to the target user profile. We use cosine-similarity between the vector that represents the user profiles of the target and the vectors that represents the values of the sensitive attribute as metric to rank them. The higher the cosine-similarity between a user profile and a sensitive value is, the higher the probability of a hidden link between them is. The Ranker algorithms are detailed in Chapter 7.

[9] www.radimrehurek.com/gensim/index.html

4.3 SONSAI user guide

SONSAI is composed of two separated tools a Collector and an Analyser. They can be launched simultaneously and they communicate through XML files generated by the Collector. The Graphical User Interface (GUI) of the Collector and the Analyser are as follows:

Collector. The Collector encompasses the Crawler functionalities. The Collector tool GUI is depicted in Figure 4.2. In order to connect to Facebook network and collect data, the user needs to provide his login and password. The Collector will then crawl Facebook network through the user account. The user can also choose to crawl his data using an adversary account and perform link disclosure attack that target his own profile by checking the option "link disclosure attack".

The user must then set the collection duration to a non null value. He can continue the collection that he previously began by clicking on the "Collect" button. He can update the already collected data by clicking on the "Update" button. The oldest data are updated first. Finally he can stop the collection by clicking on the "Stop" button.

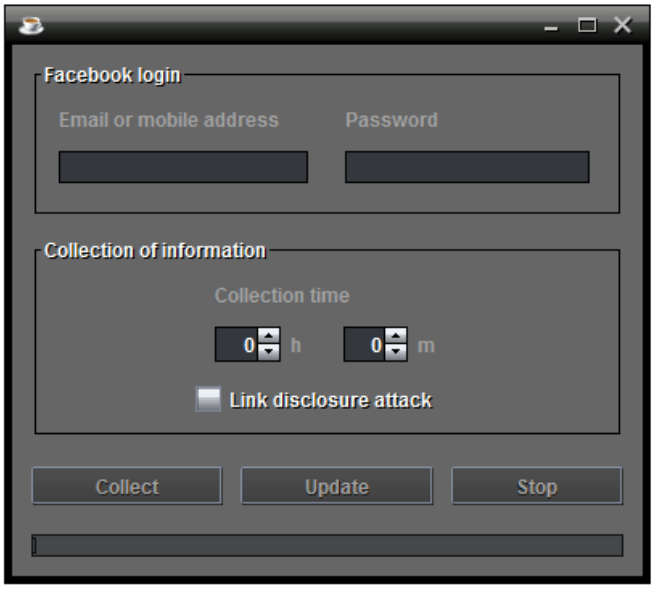


Figure 4.2: Collector GUI.

Analyser. The Analyser encompasses the functionalities of the Anonymizer, Cleanser, Random walker, Word2Vec and Ranker. The setting GUI of the Analyser tool is depicted in Figure 4.3. In order to anonymize the recently collected data and generate corresponding graphs, the user must click the "Update the dataset" button. He can then select the sensitive attribute from the list of all discovered attributes around his profile. He can choose the analysis accuracy. The accuracy is actually given as the percentage of selected attributes by the cleanser for analysis (random walk and Word2Vec) from all available attributes in the user network; these selected attributes are the ones employed to infer the closest values of the sensitive attribute to the user through Ranker algorithms. When the user clicks the "Analyse" button the Results page will be displayed.

The results GUI of the Analyser tool is depicted in Figures 4.4 and 4.5.

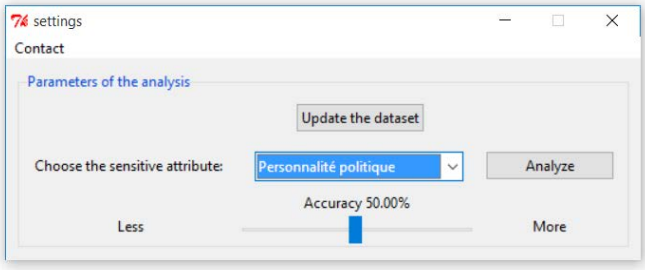
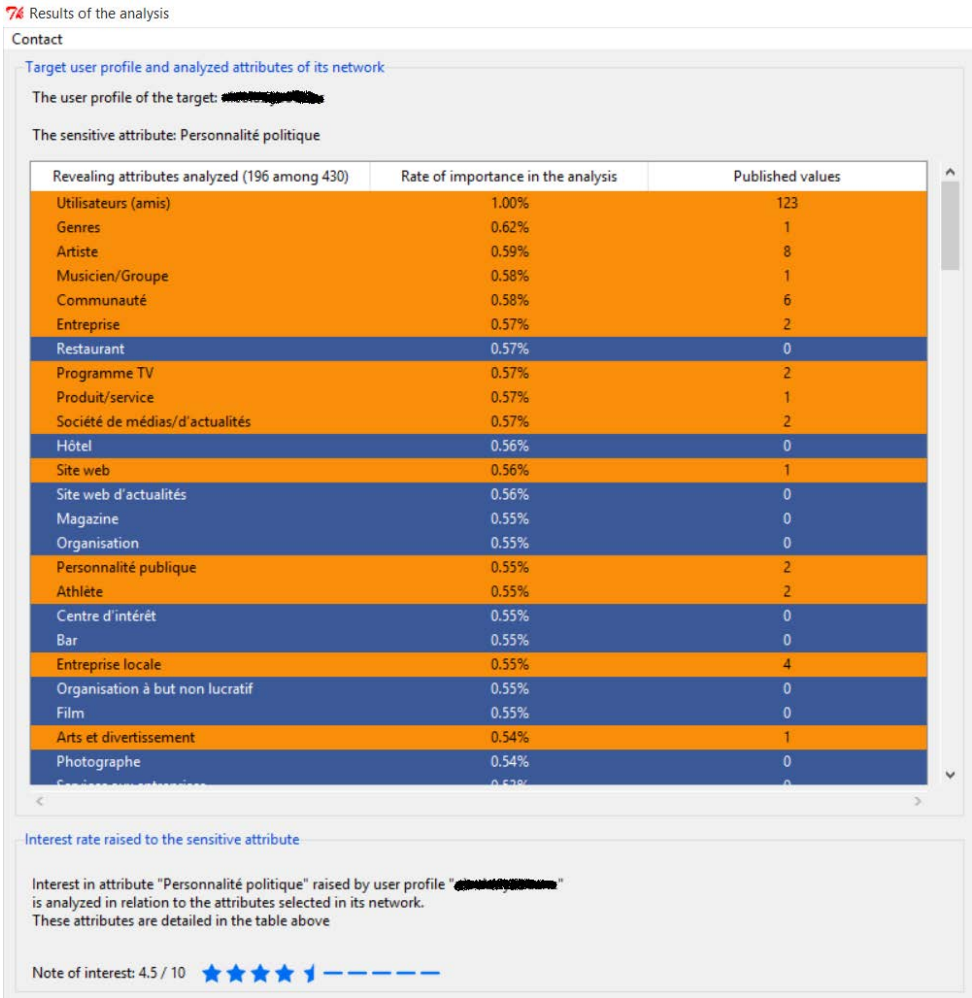


Figure 4.3: Analyser settings GUI.



Revealing attributes analyzed (196 among 430)	Rate of importance in the analysis	Published values
Utilisateurs (amis)	1.00%	123
Genres	0.62%	1
Artiste	0.59%	8
Musicien/Groupe	0.58%	1
Communauté	0.58%	6
Entreprise	0.57%	2
Restaurant	0.57%	0
Programme TV	0.57%	2
Produit/service	0.57%	1
Société de médias/d'actualités	0.57%	2
Hôtel	0.56%	0
Site web	0.56%	1
Site web d'actualités	0.56%	0
Magazine	0.55%	0
Organisation	0.55%	0
Personnalité publique	0.55%	2
Athlète	0.55%	2
Centre d'intérêt	0.55%	0
Bar	0.55%	0
Entreprise locale	0.55%	4
Organisation à but non lucratif	0.55%	0
Film	0.55%	0
Arts et divertissement	0.54%	1
Photographe	0.54%	0

Figure 4.4: Analyser results GUI - left screen.

The table in the top left part of the screen (Figure 4.4 summarizes the list of the selected attributes by the Cleanser and their importance in the analysis. The rows corresponding to attributes of which the target user has published some values are in orange colour.

In the bottom left part of the screen (Figure 4.4 is the evaluation of the user interest in the sensitive attribute. This evaluation is given as a score computed with respect to the Cleanser selection of attributes.

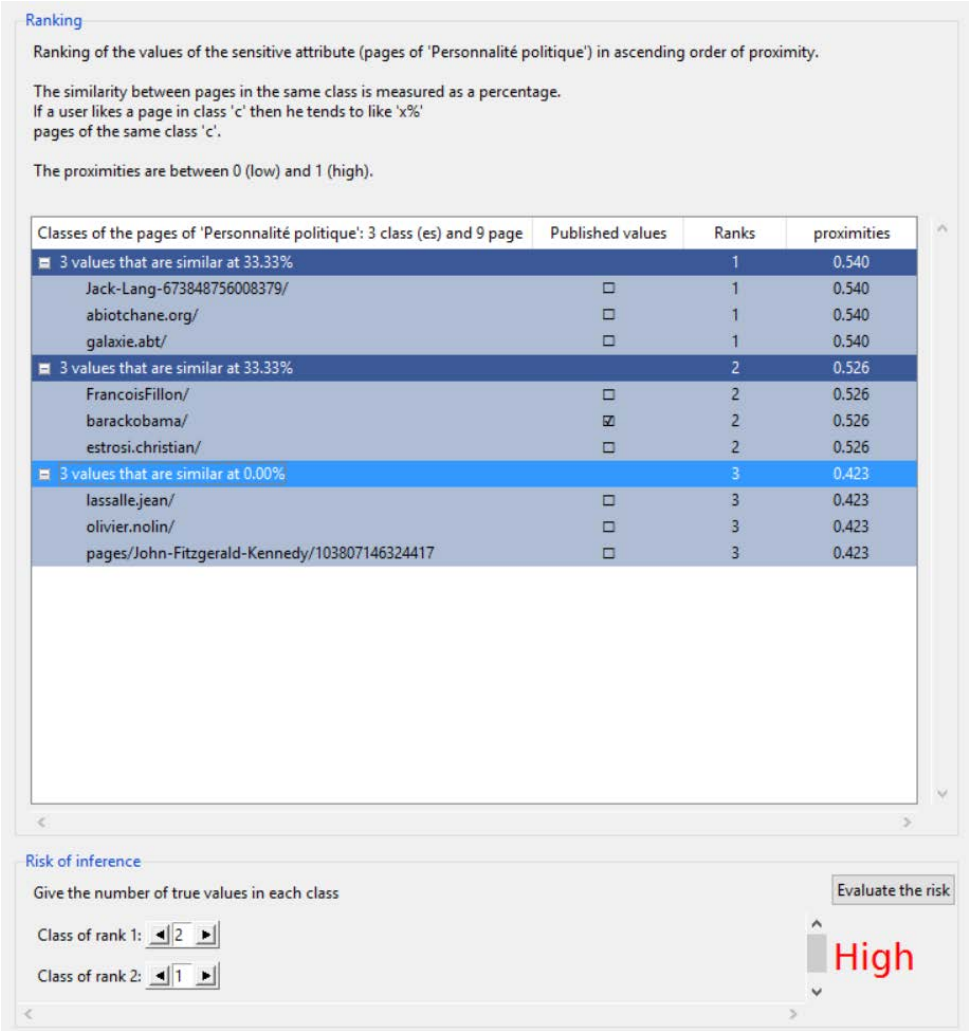


Figure 4.5: Analyser results GUI - right screen.

In the example depicted by Figures 4.4 and 4.5 the attribute labels are in French as the language of the user account on Facebook is french and we rely on the labels given by Facebook to the attributes in the user standard language.

The second table displayed in the top right part of the screen (Figure 4.5) sorts the values of the sensitive attribute according to their proximities to the user. The higher the proximity, the higher the probability of the value to be the real value of the user. Values are clustered in size-similar classes in a way that maximizes the similarity between values inside a cluster and minimizes it between values from different clusters. The similarity is measured in percentage. If a user has a value in a cluster "c" then he tends to like "x%" of values of the same cluster "c". The user can open the classes to display the values they contain. He can double-click the value to open its Facebook page and change his privacy setting. When the user publishes his values, their corresponding boxes in the column "Published values" are checked. In the bottom right part of the screen (Figure 4.5), the user can evaluate the risk of his sensitive values being inferred (by third-party). He must first specify the number of his real values inside each class even if he did not publish them on Facebook. Then he clicks the button "Evaluate the risk". The algorithm measures the accuracy of the ranking using the Area Under the Curve (AUC) as

detailed in Chapter 7. We define three levels of inference risk as follows:

If the accuracy in AUC is greater than 0.65 then the inference risk is high. On the other hand, if it is less than 0.5 then the inference risk is low. If it is between 0.5 and 0.65 the inference risk is considered to be moderate.

4.4 Examples of inference scenarios

We detail two scenarios of inferring the liked pages of politicians by a target Facebook user. We have deleted the "like" links between the users u_1 and u_2 and the pages of politicians then tried to infer them. We give more details on our attributes modelling in Chapter 5. The Ranker has to sort 4589 pages of politicians occurring in the dataset (see D1 in Annex A) in decreasing order of probability of being liked by the user. When the rank of the pages of politicians actually liked by the target is close to 1 the accuracy is close to 1. We give more details about inference accuracy computation in Chapter 7.

The target u_1 is a french user politically right-oriented. The disclosure accuracy of the pages of politicians that he likes on Facebook is 0.72. The target u_2 is a canadian user of left political orientation. The disclosure details of the politician pages he likes on Facebook is 0.97. In these scenarios, from 1928 attributes discovered in the social network by the Collector, the Cleanser selects only the top 20 attributes (0.01%) that are most correlated to pages of politicians. We give more details about how to bring to light the most important attributes for inference in Chapter 7. In Table 4.1 we summarize a sample of values of the pages of *Politicians, Gastronomy, Musicians/Bands, News/Media Websites and Communities* attributes that are liked by u_1 and u_2. Those attributes are highly correlated to the pages of politicians in Dataset D1 (see Chapter 7).

We notice that some pages do not have a label in adequacy with their contents. For example, the pages "Music Playlists" is labelled as news/media websites while it is a music page. In fact, labels are specified by the administrators of the pages. Thanks to Facebook's regular verification, the rate of mislabelled pages is negligible and does not affect the accuracy of the proposed method.

4.5 Conclusions

In this chapter we have presented the architecture and the functionalities of SONSAI. SONSAI performs two main tasks. It collects data around a given Facebook user and it analyses them in order to check the privacy level.

SONSAI simulates attacks by inferring sensitive information. It increases the awareness of users toward their personal information leaks. SONSAI helps them to take concrete steps to safeguard their privacy by identifying the sources of threats that are specific to their cases.

In Chapter 5 we will detail the social network sampling algorithm used by SONSAI in order to collect information around a given user. Data are then modelled by graph and anonymized. In Chapter 6 we cleanse the data to speed up analysis tasks and increase inference accuracy by discovering correlations between attributes. In Chapter 7 we detail the analysis algorithm used by SONSAI in order to infer the sensitive values of a given user attribute based on the cleansed data. Finally we will present experimental results on real Facebook profile to validate our algorithms.

Targets	AUCs	Politicians	Gastronomies	Musicians/Bands	News/Media Websites	communities
u_1	0.72	Marine Le Pen	Kitchen Trotter	Clean Bandit	Spi0n	Pour le retrait du timbre Femen
		Jean-François Copé	Granola	Lana Del Rey	Music Playlists	Rallye Jouvence
		Laurent Wauquiez	Kinder	DJ Antoine	Le Monde	La connerie est universelle
		Bruno Le Maire	Orangina	Timmy Trumpet	Frenchweb	Les Veilleurs
		Jean-Marie Le Pen	Foodora Global	Amxire	20 Minutes	Sauvons La PRÉPA
		Nicolas Dupont-Aignan	McDonald's	Carl Cox	Le Parisien	Parlons la mort
		Xavier Bertrand	Nestor	Walk Off The Earth	Demotivateur	Antiokia
		Nathalie Kosciusko Morizet	Bud Light Platinum	Lemaitre	Le360	La langue dans la poche
		Francois Fillon	Popchef	The Chemical Brothers	Merci Alfred	Paski's Givrés
		Marion Maréchal-Le Pen	Michel et Augustin	Cascada	CitizenPost	L'Eau Vive
			Le Boeuf Français	Imagine Dragons	MusiqueAuClaireDeLune	MA VOIX
				alt-J	Océan Surf Report	Parlons Astrologie
				French Fuse	OKLM	EXTREME
				Bondax	Ixène	Aumônerie Centrale Nantes
				Klosman	Delighted	La Transnépalaise
				A State Of Trance	ZoomOn Paris	Droite TV
				London Grammar	Les répliques	Innov' ECE
				Dillon Francis	BuzzFilGeek	Bordel De Droit
				Monsieur Monsieur	My Little Paris	Soigner Dans la Dignité
				Max Vangeli	confidentielles.com	ADDM - Respect it Enjoy it
				DJ Fresh	Boiler Room	Entourage - Réseau Civique
				Cazzette	MY Secret NY	Soutien au bijoutier de Nice
				Charlotte de Witte	Street FX Motorsport & Graphics	Valls Dégage
				RL Grime	Hitek	Banamak
				Bassjackers	Le Figaro	Fab Bike
				Tritonal	PIX GEEKS	Pour la démission de Hollande
u_2	0.97	Simon Marcil	Microbrasserie Les Grands Bois	Justice	TED	Keep Calm & Be Real
		Martine Ouellet	Budweiser	The Prodigy	InfopresseJobs	Es-tu game?
		Amir Khadir	On bouffe pour toe	Queen	Too Close To Call	Nos casseroles contre la loi spéciale
		Jack Layton	La fabrique - Brasserie artisanale	Crystal Castles	Radio-Canada Information	The Voyage North
		Jocelyn Beaudoin		Le husky	Isarta - Emplois	Arcade MTL
		Alain Therrien		Daniela Andrade	NowThis	Nous sommes les 68%
		Justin Trudeau		Boys Noize	Infos Insolites	Larping.org
		Bernard Drainville		Gunther	Faits et Causes	Pierre Céré
		Robert Aubin		Heroik	Progrès Villeray - Parc-Extension	Commodore 64
		Alexandre Boulerice		Les Poignards		Astuces de Mac Gyver

Table 4.1: Samples of values of attributes of the targets u_1 and u_2.

5

Sampling and modelling social networks

Contents

5.1 Introduction **61**
5.2 Definitions **61**
5.3 Sampling social network around a target user profile **62**
5.4 Modelling discovered links and nodes by graphs **67**
5.5 Anonymizing the social network graph models **69**
5.6 Conclusions **71**

5.1 Introduction

In this chapter we aim to sample a social network around the target user profile. Our objective is to collect data that help infer sensitive informations about the target user. To that end we design a sampling algorithm that takes into consideration three parameters: the closeness of sampled nodes to the target user profile, the type of sampled nodes and the centrality of sampled nodes as defined in this chapter. The sampled social network is then modelled by graphs. The friendship network is modelled by a unipartite graph. Each attribute is modelled by a different bipartite graph. Every graph is then anonymized before being processed as detailed in the following. Although these techniques apply to various socnets we focus here on Facebook.

5.2 Definitions

In this section we define the terms used to describe the crawling and sampling tasks.

Nodes. We distinguish three different types of nodes on Facebook: user profiles, pages and groups. User profiles are personal and managed by a unique user. Users can only create one user profile per account. On the other hand, users can create several pages and groups using the same user profile. Pages and groups can be administrated by several user profiles.

Links. Links on Facebook express relationships between nodes. User profiles can be linked to any other type of node. However, groups and pages can only be linked to user profiles. Links between user profiles can be friendship or follow-ship. In this work, we only investigate friendship

ones. Friendships on Facebook are symmetric and the friendship graph is not oriented. In this work we do not investigate administrator-ship between user profiles and pages and administrator-ship between user profiles and groups. Administrators of groups are therefore considered as standard members.

We only analyse like-ships between user profiles and pages. By liking a page a user profile will receive actualities about it. We also analyse member-ship between user profiles and groups. By becoming a member in a group a user profile will receive news concerning the group.

Crawler. A social network crawler is a robot that visits different nodes on the social network. It parses the HTML code of visited nodes and collects data. Our crawler is programmed to behave like a human user. It visits a web page at a time and can only access public data. Hence, the knowledge of the crawler about the social network evolves as it visits a new node and collect public information from its corresponding web page. For instance, at the beginning the crawler only knows the URL of the target user profile. After visiting that node and crawling it the crawler discovers new URL adresses of its friends, liked pages and groups.

Crawled node. To crawl a node the crawler accesses its web page and collects public information available on it. To crawl a user profile public values of its attributes such as the relationship status and the gender are collected as well as its friends list, its group list and its liked pages list. Crawling a group amounts to collect the list of its members (including the administrators). Crawling a page amounts to sample users that like or comment publications made on the page. Since the list of likers of pages is not available on Facebook, the crawler collect the latest n_l publication made on the page. Then, for each publication, it collects the first 50 likes, loves, Hahas, Wows, Sads or Angries. In fact, interactions spread by waves and the users who like the page are the first who receive notifications about new publications. Thus, the first interactions are more likely to be made by them. On the other hand, latest interactions have more chances to be made by users that do not like the page but they are friend with someone who do.

Collected data. To collect a particular data from a visited web page, the crawler looks for its xpath in the parsed HTML code. For instance if the crawler aims to collect the public friends list of user profile u_1, it will search the xpath $//div[@class =' fsl\ \ fwb\ \ fcb']$ in the HTML code of the web page $www.facebook.com/u_1/friends$

Discovered node. A discovered node is a node whose corresponding URL address is known by the crawler. For instance, if the crawler crawls a particular user profile and collects its public friends list, then the friends of that particular user profile are discovered even if they are not crawled yet.

Discovered link. A discovered link is a real link on social network that is known by the crawler. It represent a relationship between two discovered nodes by the crawler.

5.3 Sampling social network around a target user profile

In [Gjoka et al., 2011] authors take into consideration the different types of nodes (groups, profiles ...) and the different types of links (friendships, memberships...) when sampling social networks. In [Li et al., 2015], authors propose two random walks based sampling techniques to combat large deviation problem and sample rejection problem. In this chapter we focus on the utility

of sampled data instead of it representativeness. We aim to increase the density of the sampled network around the target while manipulating a limited budget of nodes to sample.

In this section we explain the strategy to discover nodes and links around the target user profile u_1. We also explain how to sample the discovered nodes before crawling them.

Reducing the crawling space around a target user profile

Since the separation degree of Facebook is 3.5 [Bhagat et al., 2016] and the network counts more than 2 billion monthly active users in 2018 [Statista, 2018], we need to reduce the crawling space around a target user profile. We only crawl nodes at distance two at most from the target user profile u_1.

Discovering links and nodes around the target node

The crawler does not have a complete view of all nodes and links in the network. Each crawled node extends its knowledge about the network. For instance after crawling a particular user profile, the crawler discovers the friends, the pages and the groups of that particular user profile.

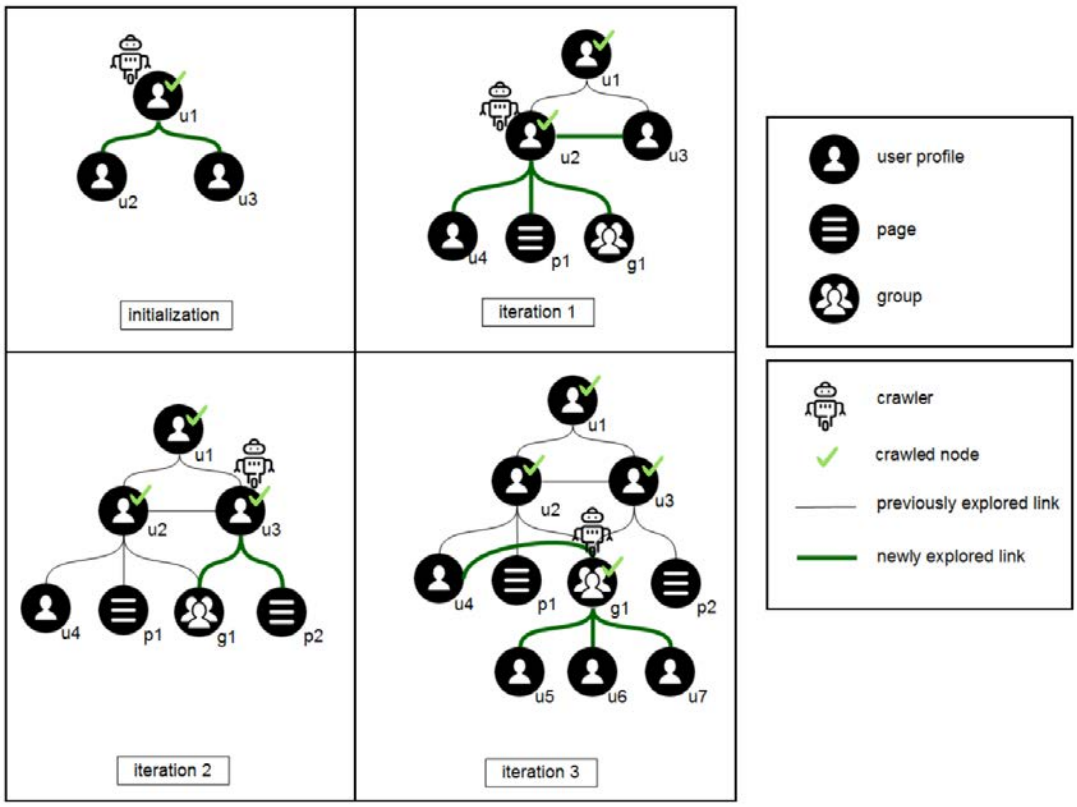


Figure 5.1: Example of the evolution of the knowledges of the crawler about the social network.

Example. Figure 5.1 depicts an example of the evolution of the crawler view of the social network. In this example, we suppose that the list of liked pages and the list of groups are not public on user profile u_1. At initialization step the crawler collects only the public friend list of

u_1. At this level the view of the crawler includes only the friends of u_1 and the links between them and u_1. At iteration 1 the crawler collects the public friend list, the public group list and the public liked page list of user profile u_2. Hence, the view of the crawler is enriched by nodes u_4, g_1 and p_1 and links (u_2, u_4), (u_2, g_1), (u_2, p_1) and (u_2, u_3) at this level. We note that although user profiles u_2 and u_3 are discovered at initialization step the link between them is only discovered at iteration 1 after collecting the friend list of u_2. After collecting the public group list and the public liked page lists of u_3 at iteration 2, the view of the crawler is enriched by node p_2 and links (u_3, g_1) and (u_3, p_2). At iteration 3, the crawler collects the lists of members (including the administrators) of group $G1$. Hence, the view of the crawler is enriched by nodes u_5, u_6 and u_7 and links (g_1, u_4), (g_1, u_5), (g_1, u_6) and (g_1, u_7).

Crawling algorithm

Objective. Our objective is to crawl a number n_c (a given parameter) of different nodes that are at most at distance 2 from the target user profile u_1. Moreover, we aim to increase the probabilities of crawling some selected types of nodes on behalf of the other types of nodes. For instance, we aim to crawl more user profiles than pages and we aim to increase the probabilities of crawling nodes at distance 1 from the target user profile u_1 on behalf of nodes at distance 2. Additionally, we want to increase the probabilities of crawling nodes that plays a central role for the target environment. The centrality of a node is defined here as the number of paths between the node and the target user profile u_1. Note that here the length of a path is at most 2. To achieve these goals, we have designed an iterative algorithm that calls a crawling function n_c times to crawl a new node.

Selecting the next node to crawl. The choice of the next node to crawl is guided by the probability of a random walker to finish on it after only two steps from the target user profile u_1. To take into consideration the centrality of nodes we compute the transition matrix by transforming the adjacency matrix of the network into a right stochastic one. Each entry V_{ij} of the transition matrix represents the probability of walking from the node i to the node j (assuming uniform selection of a link). However, the nodes in social networks are not connected to themselves. Consequently, the diagonal values of the adjacency matrix are null and a random walker cannot reach nodes at distance 1 from the target user profile u_1 after 2 steps. To cope with this small problem, we add a positive coefficient λ to the diagonal entries of the transition matrix to specify the probability of staying in the same node. Therefore with the modified matrix nodes at distance 1 are reachable.

After permitting self-loops, we count two ways to stop on a node at distance 1 from the target user profile u_1 after 2 steps. The first way is to loop on the target user profile u_1 then walk to a directly connected node to it. The second way is to walk to a directly connected node to the target user profile u_1 then loop on it. However, the only way to finish on a node at distance 2 from the target user profile u_1 after 2 steps is to first walk to a node at distance 1 then walk to a node at distance 2. Consequently, for nodes with the same centrality, the probability of crawling nodes at distance 1 is $2 \times \lambda$ greater than the probability of crawling nodes at distance 2. In order to increase the probabilities of crawling the closest nodes to the target user profile u_1 with respect to the farthest nodes, λ is chosen larger than 0.5.

In order to specify the probabilities of crawling a type of node i we define a positive coefficient α_i. For the non crawled nodes of the privileged type i, we multiply the entries of the corresponding column except for the diagonal ones by α_i in the transition matrix to get a biased matrix.

Then we multiply the first line of the biased adjacency matrix by the biased adjacency matrix. We note that the first line of the biased adjacency matrix correspond to the target user profile u_1. To guarantee that the crawler will not crawl twice the same node, we reset to zero the entries corresponding to the crawled nodes in the vector obtained by the previous multiplication. Finally, we normalise the resulting vector to obtain the transition probabilities to the next nodes TP.

Example. For the example in Figure 5.1 we consider that user profile u_1 is the target node. The algorithm is initialized by crawling u_1. To choose which node to crawl at iteration 1, we first compute the adjacency matrix A_0 of the discovered network at initialization as follows.

$$A_0 = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \end{array} \begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_3 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

At this level the crawler has discovered only nodes at distance 1 from u_1. Then, to stop on a discovered node after 2 steps we replace the diagonal values of A_0 by a positive value λ. The biased adjacency matrix $A_0^{'}$ is computed as follows with $\lambda = 1$.

$$A_0^{'} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

At this level the crawler has discovered only one type of node (user profile). Then, we directly compute the transition probabilities to next nodes, TP_0, based on $A_0^{'}$. To compute TP_0, we first multiply the first line of the biased adjacency matrix $A_0^{'}[1:]$ by $A_0^{'}$. Then we set to zero the entry that corresponds to the already crawled profile u_1. Finally, we normalise the resulting vector as follows.

$$A_0^{'}[1:] \times A_0^{'} = \begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_3 \end{array} \\ \begin{bmatrix} 3 & 2 & 2 \end{bmatrix} \end{array}$$

$$TP_0 = \begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_3 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array}$$

For iteration 1 in Figure 5.1, the crawler has discovered nodes at distance 1 and 2 from the target user profile u_1. In this example, for nodes that have the same centrality, we want the probability of crawling nodes at distance 1 to be the double of the probability of crawling node at distance 2. Thus, we take λ equal to 1. We obtain $A_1^{'}$ by making the diagonal entries of the adjacency matrix A_1 equal to λ as follows.

$$A_1 = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ p_1 \\ g_1 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & p_1 & g_1 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad A_1^{'} = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ p_1 \\ g_1 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & p_1 & g_1 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Moreover, for nodes at the same distance from u_1 and that have the same centrality, we want that the probability of crawling pages to be three times greater than the probability of crawling the other type of nodes. At this first iteration, the nodes u_4, p_1 and g_1 have the same centrality and they are at the same distance from u_1. The only discovered pages that is not crawled at this iteration is p_1. Thus, we multiply the entries of its corresponding column except for the diagonal one by $\alpha = 3$ in the biased adjacency matrix $A_1^{'}$. Consequently, we obtain $A_1^{''}$ from $A_1^{'}$ as follows:

$$A_1^{''} = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ p_1 \\ g_1 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & p_1 & g_1 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

To compute the transition probability TP_1 we first multiply the first line of the modified adjacency matrix $A_1^{''}[1:]$ by $A_1^{''}$. Then we set to zero the values corresponding to already crawled profiles u_1 and u_2. Finally, we normalise the resulting vector as follows:

$$A_1^{''}[1:] \times A_1^{''} = \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & p_1 & g_1 \end{array} \\ \begin{bmatrix} 3 & 3 & 3 & 1 & 3 & 1 \end{bmatrix} \end{array}$$

$$TP_1 = \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & p_1 & g_1 \end{array} \\ \begin{bmatrix} 0 & 0 & \frac{3}{8} & \frac{1}{8} & \frac{3}{8} & \frac{1}{8} \end{bmatrix} \end{array}$$

In this example, for the discovered network at iteration 1, we notice that the centrality of the node u_3 is 2 and it is at distance 1 from u_1. While the centrality of the node p_1 is 1 and it is at distance 2 from u_1. But the probability of selecting either the node u_3 or the node p_1 to be crawled at iteration 2 is the same. In fact, on the one hand we aim to increase the probability of crawling both nodes closer to u_1 and nodes that have high centralities. On the other hand we aim to increase the probability of crawling the pages.

Stop conditions. The algorithm stops when the number of crawled node reaches n_c (given as a parameter). It also stops when all the discovered nodes at distance at most d (given as a parameter) from the target node are crawled. In other word, it stops when the transition probabilities vector at iteration j is a null vector with $j \leq n_c$.

Algorithm steps. Because matrix multiplication is costly and we are only interested in computing the transition probabilities of paths starting from the target user profile u_1, we have implemented an iterative algorithm with nested loops. The procedure `crawl_nodes` of Algorithm 5 crawls at most n_c nodes at distance $\leq d$ from the target node u_t. Each iteration of the outer loop samples a node, crawls it and update the sets of discovered and crawled nodes. The sampling is done by random walks of length $\leq d$. The function `random_select` is designed to select randomly and return either the current node c or one (that is not a sink, see below) from the set of neighbour nodes n. The random selection law depends of the node type. We can imagine that each node c' in the pool of candidates is assigned a segment of length α depending of its type. All segments are joined end to end to obtain a large segment. Then a point is chosen uniformly at random in this final segment: if the point come from node c" sub-segment then c"

is selected by the procedure. For the special case where node c is not a sink we assign a segment of length λ to c and join it to the other segments.

Function `sinks`(j) returns the set of sinks, i.e. crawled nodes such that all discovered nodes at distance $\leq j$ from them are also crawled. A sink of depth 0 is a crawled node. Sinks are avoided by the random walks to guarantee that the final node has not been crawled yet.

```
1  Procedure crawl_nodes(u_t, d, n_c)
2      crawl_node(u_t);
3      while |crawled_nodes| < n_c do
4          c ← u_t;
5          s ← {};
6          for i ← 1 to d by 1 do
7              s.addAll(sinks(d − i));
8              c ← random_select(c, s);
9          end
10         crawl_node(c);
11     end
12 Function random_select(c, s)
13     if c ∈ s then
14         X ← 0;
15     else
16         X ← λ;
17     end
18     c.min ← 0;c.max ← X;
19     for c' ∈ c.n \ s                          ▷ c.n are connected nodes to c
20      do
21         c'.min ← X;
22         X ← X + c'.α;▷ c'.α is the parameter of selecting the node c' depending on its type
23         c'.max ← X;
24     end
25     rand ← random(0, X);
26     for c' ∈ {c ∪ c.n} \ s do
27         if c'.min ≤ rand ≤ c'.max then
28             return(c');
29         end
30     end
```

Algorithm 5: Crawling nodes around a target user.

5.4 Modelling discovered links and nodes by graphs

We use graphs to model the discovered data (links and nodes) around the target user profile. We note that in our case we limit the random walk to only 2 steps. Thus, crawled nodes can only be at distance 2 from the target user profile. However, discovered nodes can be at distance 3 from the target user profile. For the example in Figure 5.1, the crawler has crawled the target profile and 3 nodes around it after 3 iterations. However the discovered data from the network includes 10 nodes. The 3 nodes u_5, u_6 and u_7 are at distance 3 from the target user profile u_1.

We note that user profiles are connected to several types of node. We consider each type of node as a different attribute. For instance, groups, politician pages and music pages are three different attributes. In the following we use graphs to model the subnetwork associated to each attribute. Each attribute has several values that are represented by different nodes of the same type.

Modelling friendship relations

Since friendship on Facebook is symmetric, we model friendship between user profiles by undirected and unweighted graph. Let $G_f = (U, F)$ be the friendship graph where U is a set of users profiles and F is a set of friendship links between them. For instance the discovered friendship network depicted at iteration 3 of Figure 5.1 is modelled by the graph detailed in Figure 5.2.

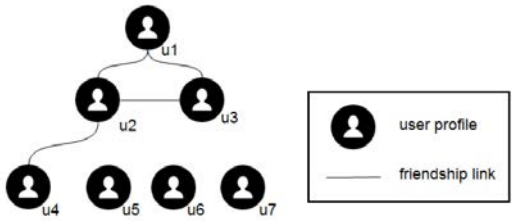


Figure 5.2: Example of friendship graph.

Although the discovered network is connected the discovered friendship graph can be disconnected as depicted by Figure 5.2. Moreover, the crawler may not have discovered all the public links between the discovered user profiles. For instance user profile u_4 may have published his friendship with u_5. But since the crawler has not crawled u_4 and u_5, the link between them remains undiscovered

Modelling group membership

We model user profile group memberships by a single undirected and unweighted bipartite graph. Let $G_m = (U, G, M)$ be the graph of group membership where U is a set of users profiles, G is a set of groups and M is a set of membership links between them. For instance the discovered membership network depicted at iteration 3 of Figure 5.1 is modelled by the graph detailed in Figure 5.3. We note that user profiles can be members of several groups.

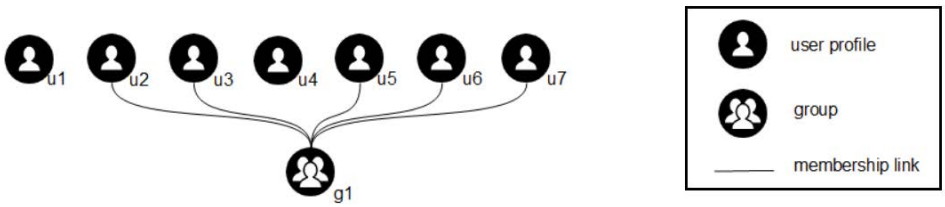


Figure 5.3: Example of groups membership graph.

Modelling page like-ship

Pages are of several types such as pages of politicians, pages of musics and pages of books. We model like-ship between user profiles and pages by several undirected and unweighted bipartite graphs. Each graph models a different type of pages. Let $G_{li} = (U, Pi, L)$ be the graph of page like-ship of type i where U is a set of users profiles, Pi is a set of pages of type i and L is a set of like-ship links between them. For instance, Figure 5.4 depicts an example of page like-ship modelled by two graphs. The graph (a) models liked pages of music type. The graph (b) models liked pages of book type. We note that user profiles can like several pages of the same type at the same time.

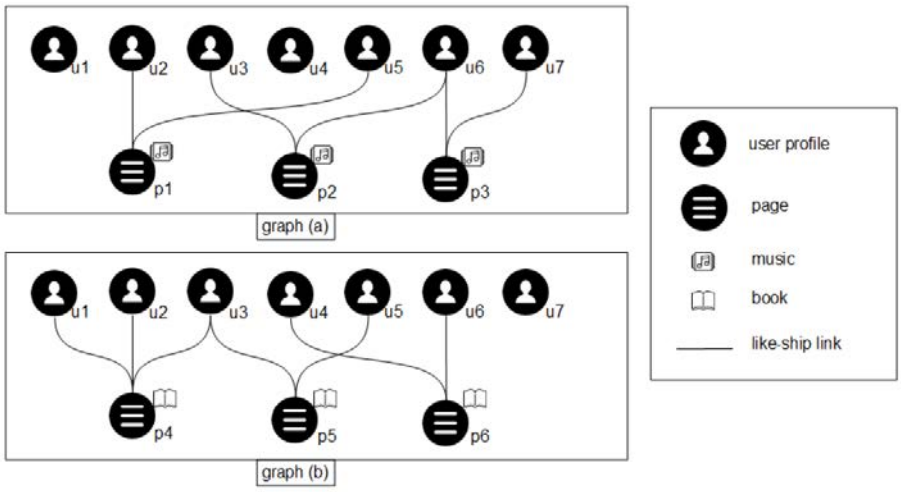


Figure 5.4: Example of pages like-ship graphs.

Modelling relationship status and gender

We model both relationship status and gender information by undirected and unweighed bipartite graphs. Let $G_{rs} = (U, RS, L_{rs})$ be the relationship status graph where U is a set of users profiles, RS is a set of relationship status and L_{rs} is a set of links between them. Let $G_g = (U, G, L_g)$ be the gender graph where U is a set of users profiles, G is a set of genders and L_g is a set of links between them. For instance, Figure 5.5 depicts a relationship status and gender network modelled by two graphs. Graph (a) models gender and Graph (b) models relationship status. We note that Facebook allows users to select at most one relationship status among eleven possible options (single, in a relationship, married, engaged, in a civil union, in a domestic partnership, in an open relationship, it's complicated, separated, divorced, or widowed). Users can select at most one gender from two possible ones (male or female).

5.5 Anonymizing the social network graph models

In this work we focus on the structure of social networks rather than the semantic of nodes. The meaning of the types of nodes is not taken into consideration in our analysis. Hence, the anonymisation task do not affect the analysis process. Moreover, we safeguard the privacy of users in released datasets by anonymising information that leads to their user profiles in the real Facebook network.

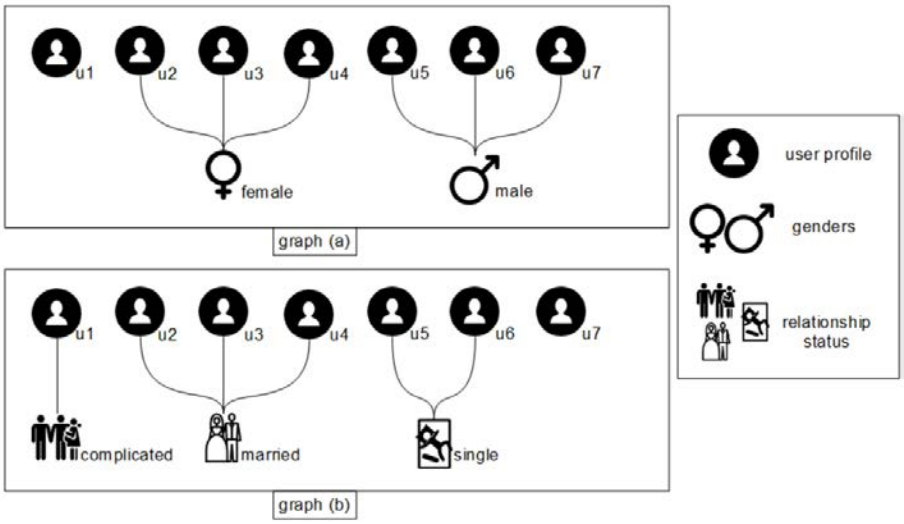


Figure 5.5: Example of relationship status and gender graphs.

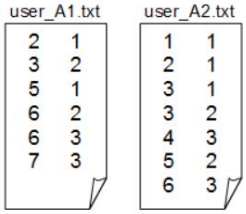
Facebook identifiers are replaced by fresh identifiers. Each node in the network is then identified by a unique integer ID replacing its Facebook ID. The anonymized IDs are sorted according to the type of the nodes. For instance, the first set of IDs identifies user profiles while the second set of IDs identifies music pages and the third one identifies book pages. We also anonymize the types of nodes that are not user profiles. Users profiles are labelled as *user* while other nodes are labelled as Ai, where i is an integer.

We use the tab-separated value (TSV) format to save the anonymized graphs. The TSV format is one of the most general delimiter-separated values format (DSV) and it is widely used in graph exchange. Each graph is separately anonymised and saved in a different TSV file of two columns. The first column represents the unique IDs of the users profiles. The second column represents the indexes of the nodes that are connected to the users profiles. To obtain the unique ID of a node from the second column, we sum its index and the offset of its label. The offset O_i of the label A_i is equal to the sum of the number of users and the total number of nodes of all the attribute A_j with $j < i$. It is computed as follows.

$$O_i = |users| + \sum_{j=1}^{i-1} |A_j|$$

where $|users|$ is the number of user profiles and $|A_j|$ is the number of nodes of attribute A_j. Figure 5.6 depicts an example of two TSV files corresponding to graphs (a) and (b) of Figure 5.4. The IDs of the user profiles range from 1 to 7. The label of the music pages is $A1$ and its offset is 7. The label of the book pages is $A2$ and its offset is 10.

Here we focus on the collected data analysis rather than data anonymisation. We are aware that a zero risk of data de-anonymisation does not exist. However, we aim to reduce it by hardening the de-anonymisation task. For instance, in 2006 Netflix has released anonymised data concerning 100 million movie rates made by 500 thousand users. Netflix argues that the released micro-data cannot be de-anonymised because they have removed all users IDs. They only kept movies titles, ratings and dates. Moreover, the released data represents only $\frac{1}{8}$ of their records in 2005. However, in [Narayanan and Shmatikov, 2006], the authors have successfully de-anonymised many user IDs from the released Netflix dataset by using auxiliary information



Facebook labels	anonymized labels	number of nodes	offsets
users	users	7	0
Music	A1	3	7
Book	A2	3	10

Figure 5.6: Example of TSV files.

from the Internet Movie Database (IMDb).

In contrast to the dataset released by Netflix where attribute values (movies title, rating, date of rating) are not anonymised and only user IDs are anonymised, all attribute values are anonymised in our datasets. In this work, we focus on analysing the social network structure. We discard all information about the semantics of attribute values. To de-anonymise our data the adversary should then analyse the graph structures in our database.

The first challenge in the de-anonymisation journey is to de-anonymise the label of the attributes. For instance, to figure which graph models the network of like-ship between movies and users, the adversary can only rely on graph properties such as density, connectivity and centrality. For example, based on auxiliary sources of knowledge, the adversary may guess that this graph must be very sparse. The second challenge is to de-anonymise the value of the attributes. For example, the adversary may have auxiliary information about the most popular movies and guess that they must have the highest degrees in the graph. Finally, the adversary can proceed to de-anonymise the IDs of users based on their de-anonymised attribute values.

Since the graph that models friendship between users in our dataset has the label $user_user$, an adversary may analyse its structure in order to detect communities and then de-anonymise users IDs based on their communities [Nilizadeh et al., 2014]. However, the fast evolution of the friendship network introduces noise and makes this task harder for this adversary. In fact, since 2015 and every 30 minutes, about one million friendship links are created on Facebook [Wits, 2015].

We note that Facebook counts more than 2 billions profiles, thousands of attributes and hundreds of millions of values of attributes [Noyes, 2018]. Since our dataset did not exceed 0.0001% of Facebook nodes, the rate of false positive would be high in de-anonymisation tasks.

5.6 Conclusions

Since the number of Facebook users is huge and the degree of separation between them is low, we crawl only nodes that are at distance two at most from the target node. However, a discovered node can still be at distance three from the target node. The discovered network is larger than the crawled network because the crawler collects the list of connected nodes to each crawled node. We also bias the sampling task to crawl nodes that are closer to the target node or to orient the crawler towards a particular type of nodes. For instance, we orient the crawler towards pages of restaurant, pages of fast-food and pages of drinks if we are willing to analyse the eating habits of the target user.

The discovered network is modelled by different graphs. Each graph is separately anonymised and saved as a TSV file. This way of handling the data facilitate the preprocessing tasks that will be detailed in the next chapter.

6

Cleansing the collected data

Contents

6.1 Introduction **73**
6.2 Definitions **73**
6.3 Motivations for cleansing data **74**
6.4 Cleansing the sensitive graph **76**
6.5 Cleansing the learning graphs **80**
6.6 Cleansing results **85**
6.7 Conclusions **92**

6.1 Introduction

In this chapter we first aim to blow the lid off correlations between attributes. Our objective is to select the best attributes to help inferring user preferences about a given sensitive attribute. We recall that attributes are modelled by graphs as depicted in the previous chapter. In order to select the best graphs for learning tasks we define two techniques depending on the sensitive graph. We recall that nodes are anonymized and their labels are not taken into consideration when selecting the most relevant graphs for learning. Moreover, we transform the friendship graph $G_f = (U, F)$ into a bipartite graph $G_f^{'} = (U, U, F)$ by duplicating the set of nodes. All considered graphs are then bipartite ones and processed uniformly.

Another objective is to speed up link prediction (detailed in next chapter) by reducing whenever possible the number of attribute values. This is obtained by clustering similar values.

6.2 Definitions

We use several techniques to cleanse the collected data. In this section we define the terms used to describe those techniques.

Sensitive graph. The sensitive graph is the graph that models a sensitive attribute.

Learning graphs. The learning graphs model non sensitive attributes. They are available for learning in order to predict hidden links in the sensitive graph.

Relevant graphs. Relevant graphs are special learning graphs that guarantee the machine learning to infer secret links in the sensitive graph with high accuracy. Relevant graph approach to cleanse data is used when a given user can be linked to many attribute values in the sensitive graph. The relevance of a given learning graph depends on its learning rate lr, confidence rate cr and Hamming distance rate hr.

Discriminant graphs. Discriminant graphs are learning graphs that hold discriminant information about the sensitive graph. Discriminant graph approach to cleanse data is used when a given user can be linked to at most one attribute value in the sensitive graph. The discrimination of a given learning graph depends on its learning rate lr, confidence rate cr and discriminant rate dr.

Learning rate. The learning rate lr quantifies the information that can be learned from a given learning graph concerning a given sensitive graph. The higher the learning rate is, the more relevant/discriminant the learning graph is.

Confidence rate. The confidence rate cr quantifies the amount of information that can be compared between a given learning graph and the targeted sensitive graph. The higher the confidence rate is, the more relevant/discriminant the learning graph is. .

Hamming distance. The Hamming distance between two weighted graphs with the same set of nodes is the sum of the weight differences between their corresponding edges.

Hamming distance rate. The Hamming distance rate hr quantifies similarity between a given learning graph and the targeted sensitive graph. The lower the Hamming distance rate is, the more relevant the learning graph is. A weighted graph WS where nodes are user profiles will be derived from the sensitive graph S and another weighted graph WL will be derived from a given learning graph L. The Hamming distance rate between WS and WL is the quotient of the Hamming distance between WS and WL to the maximal distance between WS and any weighted graph admitting the same set of nodes and where weights are real numbers between 0 and 1.

Discriminant rate. The Discriminant rate of a given learning graph l quantify the highest percentage of users that are connected in l and connected to a particular attribute values in the sensitive graph. For instance, if the majority of users that like cosmetic product pages are female, then the graph that models the link-ship between user profiles and cosmetic products is discriminant for the sensitive graph that models user gender. The higher the discriminant rate is, the more discriminant the learning graph is.

6.3 Motivations for cleansing data

Data cleansing consists in discarding less relevant/discriminant attribute graphs from the data available for inferring sensitive attribute values on one hand and selecting or synthesising (by merging) pertinent ones.

Focusing on relevant/discriminant data increases the accuracy of the final results. Moreover by deleting useless data the training time is considerably shortened. However data cleansing must be fast and effective. In this chapter we design several simple low cost solutions to cleanse the data and achieve our objectives.

In preliminary work, we have carried out two experiments without cleansing the data in order to assess the effectiveness of the inference process. Those experiments have confirmed the advantages of applying cleansing tasks before the inference process.

The dataset (D1) contains 15012 crawled user profiles and 1926 different type of pages considered as attributes. More details about the dataset (D1) is given in Annex A. For each experiment we have selected a small set of graphs that model attributes with a number of values close to the number of values of the sensitive attribute. The experiments have consisted in hiding several links in the sensitive graph then trying to infer them based on information from the learning graphs. For each experiment we have performed more than 500 tests. For each test we have randomly hidden links in the sensitive graph and changed the learning quota (aka importance level) for each learning graph. We have used Bayesian optimization as depicted in [Snoek et al., 2012] to automatically find the best configuration of quotas. The algorithm gives higher learning quotas to the most important graphs. Those graphs result in better accuracy when the algorithm learns more on them. They have low rank in Tables 6.1 and 6.2 that detail the learning graph of respectively Experiment 1 and 2.

Each test lasted for one hour on a 3.30 GHz processor. The accuracy (AUC) of inferring the hidden links of the first experiment range from 0.67 to 0.68. The accuracy (AUC) of inferring the hidden links of the second experiment range from 0.87 to 0.88. The whole experiment process lasted for a month and though could only reveal the best graphs from a small set of graphs. In this chapter we details several cleansing algorithm that allow us to select the best graphs from all the available graphs (1926 graphs) in only few minutes (less than 10 minutes).

Experiment 1. In this experiment the sensitive graph models the link-ship between user profiles and pages of travel agencies. We selected 6 learning graphs. Details about the graphs are given in Table 6.1.

Ranks	Attribute graph	♯ Connected user profiles	♯ Attribute values	Density $\times 10^{-4}$
	Travel Agencies	3 370	4 827	6
1	Users	13 155	13 155	12
2	Politicians	2 554	4 589	9
3	Causes	2 547	4 410	6
4	Small Business	2 386	4 350	5
5	Consulting Agencies	2 288	4 176	7
6	App Pages	4 396	4 244	8

Table 6.1: Details about the graphs used in experiment 1.

Experiment 2. In this experiment the sensitive graph models the link-ship between user profiles and pages of politicians. We selected 26 learning graphs. Details about the graphs are given in Table 6.1.

Ranks	Attribute graphs	♯ Connected user profiles	♯ Attribute values	Density $\times 10^{-4}$
	Politicians	2 554	4 589	9
1	Gastronomies	4 763	8 422	6
2	Bars	3 433	7 038	5
3	News/Media Websites	5 550	9 247	7
4	Arts And Leisure	2 776	5 255	6
5	Healths/beauty	3 016	8 073	5
6	Pleasures	3 704	9 454	4
7	Animators	3 788	6 826	5
8	Travel Agencies	3 370	4 827	6
9	Magazines	4 733	9 955	6
10	Retail Business	2 489	5 050	5
11	Business Services	2 288	4 176	7
12	Photographers	2 773	6 252	4
13	Entertainment Websites	5 669	8 319	5
14	Fictional Characters	3 200	5 306	6
15	Causes	2 547	4 410	6
16	Personal Blogs	3 534	8 607	4
17	Community Organizations	2 895	5 227	6
18	Application Pages	4 396	4 244	8
19	Books	3 003	8 019	5
20	Games/toys	2 878	5 468	7
21	Disc Houses	1 566	5 425	12
22	Authors	2 867	5 827	7
23	Local Enterprises	2 386	4 350	5
24	Restaurants	2 771	5 600	5
25	Schools	3 613	5 555	4
26	Users	13 155	13 155	12

Table 6.2: Details about the graphs used in experiment 2.

6.4 Cleansing the sensitive graph

Values of attributes to which user profiles are connected can be pages created on Facebook or any other web page connected to Facebook via the Open Graph protocol (OGp). The OGp is a Facebook invention that allows any web page to become an object in a social graph [10]. Some attributes such as community topics and groups of music have a huge number of values. For instance, we count 137k community topics, 84k different groups of music and 31k different artists liked by only 15k different users. Therefore, we reduce the space of values by clustering them to save computational cost when applying unsupervised learning in the next chapter. Values that are more likely to be connected to a same random user profile from the network are clustered together. The problem is then alleviated and it can be solved in two steps instead of directly inferring the exact value of attribute among huge number of values. The first step consists of inferring the cluster of values that are most likely to be connected to the target profile. The second step consists of inferring the most likely values to be connected to the target profile among

[10] https://developers.facebook.com/docs/sharing/opengraph

the selected cluster.

Clustering constraints

Semantically clustering the values of attributes into contexts requires huge up-to-date knowledge about many fields and many cultures. For instance, Eddie Murphy movies are linked to comedy in 2017 but in 2007 his name was correlated to drama for his role in Dreamgirls for which he picked up his only Oscar nomination. To cope with this problem we cluster the values of attributes based only on user profiles to which they are connected. However, we do not cluster user profiles simultaneously since it is obvious that they can have very different preferences at the same time. For instance, the same user can like both horror and documentary movies. Furthermore, we aim to infer secret values connected the target profile by exploiting information from different graphs, including the friendship graph.

Let $G_i = (U, V_i, L_i)$ be the graph that models the network of links L_i between the set of user profiles U and the set of values V_i of the attribute A_i. The Figure 6.1 depicts an example of clustering of the pages of politician, $A_i = politician$. In this example, we partition $G_{politician}$ into $n_l = 2$ sub-graphs of almost equally sized disjoints sets of politicians.

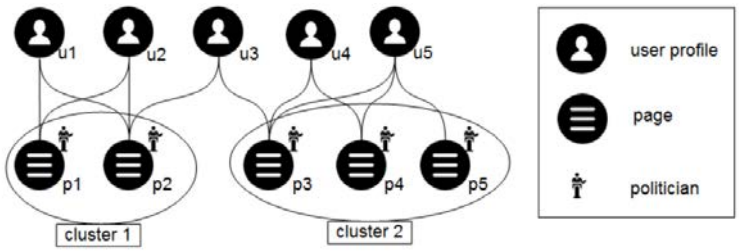


Figure 6.1: Example of clustering the values of the attribute "politician".

The problem can be related to a k-way graph partitioning problem since the goal is to divide the set of attribute values into k cluster of about equal size. We also aim to maximize the likelihood of being connected to a same user profile between the values of attribute that belong to the same cluster. Thus, different approaches of dense sub-graph discovery could be applied to iteratively seek and cut the densest sub-graph from the original graph [Lee et al., 2010]. The set of values of attribute of each sub-graph define a cluster. However, due to the sparsity of the social graph we consider, the dense sub-graphs are usually small and the algorithms mentioned in [Lee et al., 2010] end up partitioning the graph into a large number of not equally-sized sub-graphs with decreasing densities. To cope with this issue, we propose a greedy algorithm that adds constraints on the size of sub-graphs and the similarity between the values of attribute of each sub-graph. In the following we denote by $|S|$ the cardinal of a set S.

Objective function. Our objective is to find a partition π_l of attribute values in n_l clusters that maximize the similarity between values inside each cluster. We define the similarity between two attribute values v and v' to be the Jaccard index that measures the ratio of their common linked user profiles to the union of their linked user profiles. The set of user profiles linked to a value v of the attribute A_i in the graph $G_i = (U, V_i, Li)$ is defined by $|\{u \in U \text{ s.t.}(u, v) \in L_i\}|$ and denoted by $links(v)$. That is,

$$similarity(v, v') = \frac{links(v) \cap links(v')}{links(v) \cup links(v')} \tag{6.1}$$

For computational efficiency the number of clusters n_l must be small. But if n_l is too small the neural network detailed in the next chapter will be doomed to learn from insufficient data. On the other hand, if n_l is too large the neural network predictions will not be reliable due to over-fitting. Moreover, clusters must be almost equally-sized to avoid fostering a particular label. Therefore we only consider partitions $(c_1, \ldots, c_{n_l})$ of the attribute values satisfying $\sqrt{m} \leq \ |c_k| \leq \ 2\sqrt{m}$ for $1 \leq k \leq n_l$, where m is the number of all attribute values. Consequently, the number n_l of clusters satisfies $\frac{\sqrt{m}}{2} \leq n_l \leq \sqrt{m}$. The set of partitions satisfying the constraints above is denoted by Π_l. A good criteria for a candidate cluster c is to maximize the average similarity $similarity(c)$ between all couples of attribute values inside this cluster. Hence the objective function is given by Expression 6.2.

$$\max_{(c_1,\ldots,c_{n_l})\in\Pi_l} \frac{1}{n_l} \left(\sum_{k=1}^{n_l} similarity(c_k)\right) \tag{6.2}$$

Details of the clustering algorithm

Computing the average similarity of a cluster c is expensive due to the quadratic number of couples of values in c. Moreover, the algorithm needs to find the cluster of maximal average similarity among the numerous ones of size between $\sqrt{m}$ and $2\sqrt{m}$. To get around this problem, we propose a greedy algorithm that computes only the similarity between a cluster of movies and an unlabelled attribute value (that is a value not assigned yet to a cluster). Therefore we define:

$$similarity(c, v) = \frac{\sum_{v' \in c} similarity(v', v)}{|c|} \tag{6.3}$$

The idea now is to seek, from the set of unlabelled attribute values, an attribute value with maximal similarity with the cluster c. The function `seek_max_similar` returns the `max_similar` value and its $max_similarity$. Then, we add the chosen attribute value (`max_similar`) to c. The algorithm keeps adding attribute value to c until it reaches the stop conditions. It then defines next clusters sequentially the same way as detailed in Algorithm 6 until all attribute values are labelled.

Stop conditions. The algorithm stops adding attribute values to the current cluster c when the size of the cluster c is equal to $\texttt{int}(2\sqrt{m})$ or is in $[\sqrt{m}, 2\sqrt{m}-1]$ and one of the two following additional conditions is fulfilled: i) the similarity between c and any of unlabeled attribute values is less than $\frac{1}{2}$; ii) the number of unlabeled attribute values is higher than $\sqrt{m}$. In other word, there exists no sufficiently similar attribute value to add to the current cluster and there is enough unlabeled attribute values to create new clusters. There is also a stopping condition (line 11) when the number of unlabelled attribute values is $\texttt{int}(\sqrt{m})$ to guarantee that the size of the last cluster will be at least $\sqrt{m}$. Finally, the main loop stops when all attribute values are labelled.

Size of partitions. We have analysed the performance of the proposed algorithm with respect to the minimal size of computed clusters, where no cluster can have twice the size of other cluster from the same partition. Tests depicted by Figure 6.2 show that the choice of the minimal size to be the root square of the size of the set of attribute values yields good results for both very sparse graphs like Users-FastFoods graph (density = 0.0018) and less sparse graphs like the Users-Actors graph (density = 0.012). We note that this choice yields some clusters of high similarity (≥ 0.7), few sub-graphs (less than the square root of the number of attribute values) and relatively high

Data: $G_A = (U_A, V_A, P_A)$,
Result: π_l ▷ decomposition of V_A into l clusters

```
1  π_l ← ∅
2  B ← √|V_A|
3  V ← V_A                                  ▷ V contains values not assigned to a cluster
4  while |V| > 0 do
5      c ← one_most_liked(V)                ▷ initialisation of a new cluster with one element
6      while |c| < 2B  and |V| > 0 do
7          max_similar, max_similarity ← seek_max_similar(c, V)
8          if B ≤ |c| then
9              if max_similarity < 1/2 and |V| > B then
10                 break
11             end
12             if |V| = int(B) then
13                 break
14             end
15         end
16         c ← c ∪ max_similar
17         V ← V \ max_similar
18     end
19     π_l ← π_l ∪ {c}
20 end
```

Algorithm 6: Partition of a set of attribute values into clusters.

mean similarity compared to all partitions similarities (larger than the mean of the means of all similarities).

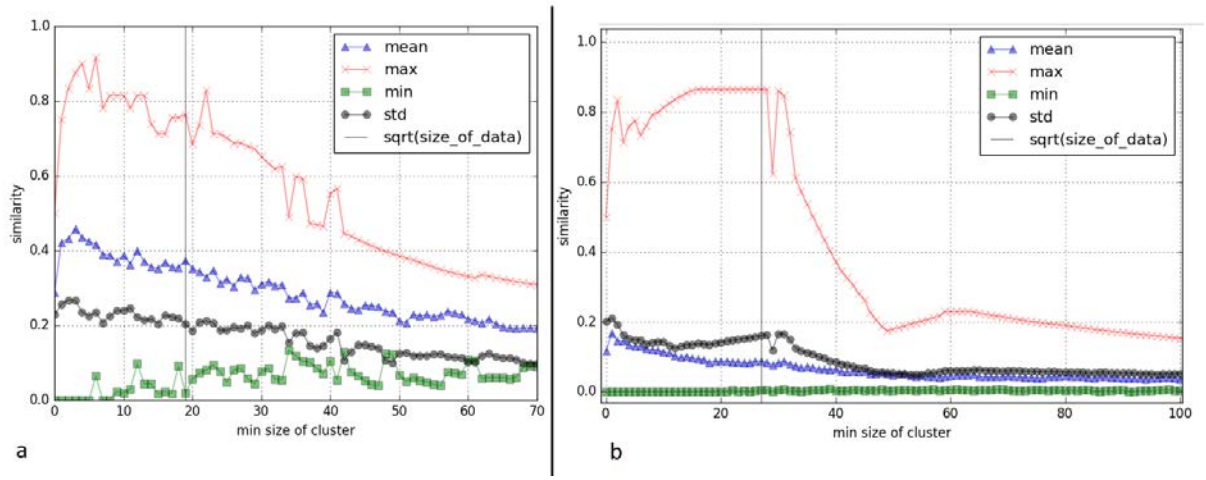


Figure 6.2: Variation of partitioned bipartite sub-graph similarities with respect to the minimal size of sub-graphs, (a) Users-Actors graph: 15k users, 364 actors, (b) Users-FastFoods graph: 15k users, 777 fast foods.

6.5 Cleansing the learning graphs

In this section we details two approaches to select the best graphs for learning tasks. Selected graphs speed up and improve inference results. The choice of the selection approach depends on the properties of the sensitive graph. If user profiles can be linked to many values in the sensitive graph then the so-called relevant graph approach to cleanse the data will be applied. However, if user profiles can be linked to at most one value in the sensitive graph then the so-called discriminant graphs approach will be applied instead.

For both approaches we compare the structure of a given learning graph to the structure of the sensitive graph. We compute the learning rate, lr, and the confidence rate, cr, related to each learning graph in both approaches. Let $l = (U, V_l, L_l)$ (resp. $s = (U, V_s, L_s)$) be a learning (resp. sensitive) graph, V_l (resp.V_s) the set of learning (resp. sensitive) values. Let $\texttt{deg}_l(u)$ (resp. $\texttt{deg}_s(u)$) be the degree of node u in graph l (resp. s) and U_l (resp.U_s) be the set of users u with $\texttt{deg}_l(u) > 0$ (resp.$\texttt{deg}_s(u) > 0$). To compare l with s we first split each graph into two parts: the first one contains user profiles hiding their links in the sensitive graph, and the second one holds user profiles publishing their links in the sensitive graph. The ratio of user profiles that publish their links in the first part (resp. second part) of the learning graph represents the learning rate, lr, (resp. the confidence rate, cr). The learning rate, lr, and the confidence rate, cr, are computed as follows:

$$\begin{aligned} lr(l) &= |U_l \cap (U \setminus U_s)| / |U \setminus U_s| \\ cr(l) &= |U_l \cap U_s| / |U_s| \end{aligned}$$

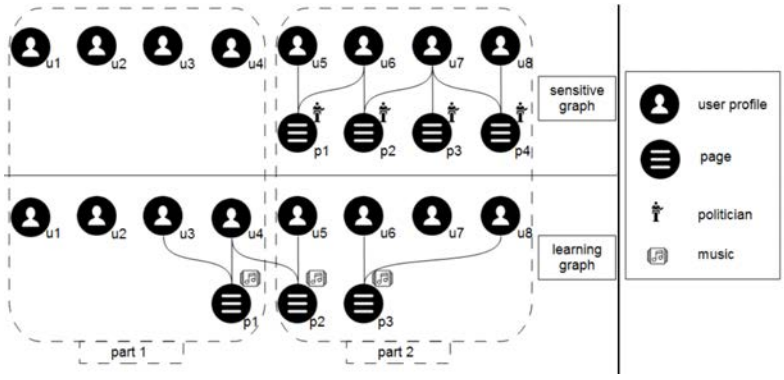


Figure 6.3: Example of cutting graphs for structure comparison.

Figure 6.3 depicts an example of splitting two graphs for comparison. The graph that models the link-ship between user profiles and pages of politicians is the sensitive graph. The graph that models the link-ship between user profiles and pages of musics is the learning graph. The learning rate, lr, for this example is equal to 50% and the confidence rate, cr, is equal to 75%.

Relevant graph approach to cleanse data

User profiles can be linked to many values of the same attribute such as politicians and musics. For instance, for a sample of 15012 user profiles we count 2554 user profiles that publish the lists of their liked politicians. The graph that models the like-ship between user profiles and pages of politicians counts 4589 pages of politicians. After discarding the user profiles that are not connected to any pages of politicians, the graph density gets equal to 9×10^{-4}. In order to infer

hidden links in such sparse graphs, we need to select graphs with similar structure. We define 3 steps to compute Hamming distance rate, hr, between graphs in order to bring to light these graphs.

Step 1: Cutting a given learning graph and the sensitive graph then transforming them into two weighted unipartite graphs. In this step,we discard user profiles that have a null degree in the learning graph or in the sensitive graph.Then, we transform the remaining parts from both the learning graph and the sensitive graph into two weighted unipartite graphs where nodes are user profiles. We use the Jaccard index to weight the links between user profiles in the unipartite graph according to their common attribute values in the initial bipartite graph. The Jaccard index between two user profiles u_1 and u_2 in a given graph,A_i, is computed as follows.

$$Jaccard_{A_i}(u_1, u_2) = \frac{|links_{A_i}(u_1) \cap links_{A_i}(u_2)|}{|links_{A_i}(u_1) \cup links_{A_i}(u_2)|} \tag{6.4}$$

The function $links_{A_i}(u_j)$ returns the set of nodes to which the user profile u_j is connected in the graph A_i.

Algorithm 7 computes two weighted unipartite graphs from a learning graph and the sensitive graph. From line 3 to line 5 the algorithm selects the user profiles that are connected in both the learning graph and the sensitive graph. The selected user profiles are then added to the weighted unipartite graphs. The Jaccard functions in line 7 and 9 return the Jaccard index of a couple of users given as parameter in the learning graph and the sensitive graph respectively. Those indexes represent the weight of the links that will be created between the same user profiles in respectively the weighted unipartite learning graph and the weighted unipartite sensitive graph.

```
   Data: l, s                                  ▷ learning graph and sensitive graph
         U                                     ▷ set of user profiles
   Result: lw                                  ▷ weighted unipartite learning graph
           sw                                  ▷ weighted unipartite sensitive graph
1  U' ← {};                                    ▷ set of kept user profiles
2  foreach u ∈ U do
3      if degree_in_s(u) > 0 ∧ degree_in_l(u) > 0 then
4          lw.add_node(u);
5          sw.add_node(u);
6          foreach u' ∈ U' do
7              w ← Jaccard_l(u, u');
8              lw.add_link(u, u', w);
9              w ← Jaccard_s(u, u');
10             sw.add_link(u, u', w);
11         end
12         U'.add(u)
13     end
14 end
```

Algorithm 7: Cutting and transforming the learning graph and the sensitive graph into two weighted unipartite graphs.

For the example in Figure 6.3, we cut the graphs containing user profiles u_5, u_6 and u_8 as they publish their links in the sensitive graph and the learning graph. The two cut graphs are then transformed into weighted unipartite graphs as depicted by Figure 6.4.

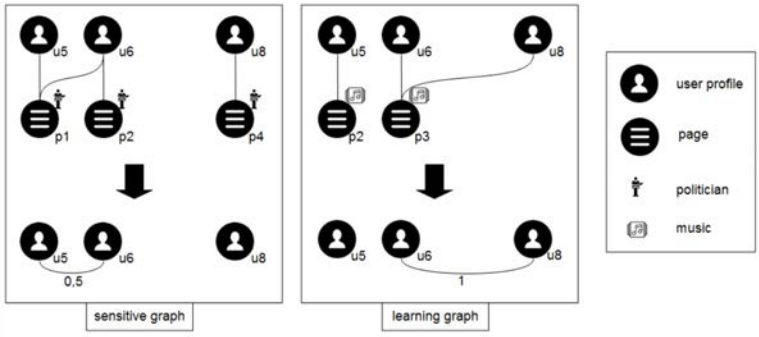


Figure 6.4: Example of transforming bipartite graphs into weighted unipartite graphs.

Step 2: Computing the Hamming distance rate between two unipartite weighted graphs. In this step, we discard user profiles that have a null degree in the learning graph or in the sensitive graph. The Jaccard index between two user nodes u_1 and u_2 in a given graph A is computed as follows, where the function $links_A(u_j)$ returns the set of nodes to which user node u_j is connected in the graph A.

$$Jaccard_A(u_1, u_2) = \frac{|links_A(u_1) \cap links_A(u_2)|}{|links_A(u_1) \cup links_A(u_2)|} \tag{6.5}$$

The Hamming distance Hd between graphs l and s is defined by:

$$Hd(l, s) = \sum_{\substack{u_k, u_j \in U_l \cap U_s \\ k \neq j}} |Jaccard_l(u_k, u_j) - Jaccard_s(u_k, u_j)| \tag{6.6}$$

In order to compare learning graphs with different sets of common connected profiles $Ul \cap Us$, we divide this distance by the maximal Hamming distance that can be obtained on such a set: $hr(l; s) = Hd(l; s)/M(l, s)$ where

$$M(l; s) = \sum_{\substack{u_k, u_j \in Ul \cap Us \\ k \neq j}} |Max(Jaccard_s(u_k, u_j), 1 - Jaccard_s(u_k, u_j))| \tag{6.7}$$

For instance, the Hamming distance between the sensitive and learning graph in Figure 6.4 is 1.5. In order to transform the learning graph into the sensitive graph, the link weighting 1 between u_6 and u_8 must be deleted and a link between u_5 and u_6 with weight 0.5 must be created. The maximal Hamming distance between the sensitive graph and the farthest graph that has the same set of users is 2.5. Since the Jaccard index of u_6 and u_8 is 0 then the farthest possible Jaccard index to it is 1. Similarly, the Jaccard index between u_5 and u_8 is 0 and the farthest Jaccard index to it is 1. Besides, Jaccard index between u_5 and u_6 is 0.5. Then, the farthest possible Jaccard indexes to it are 0 or 1. Consequently, the Hamming distance rate between the two graphs depicted in Figure 6.4 is $hr = \frac{1.5}{2.5} = 0.6$

Step 3:Selecting most relevant graphs from learning graphs w.r.t. a sensitive graph. In this final step, we aim to select the best graphs for learning based on their learning rate, lr, confidence rate, cr and Hamming distance rate, hr.

Thresholds selection method. We define three thresholds for these criteria: θ_{lr}, θ_{cr} and θ_{hr}. We first discard the learning graphs that have a learning rate lr lower than threshold θ_{lr} since they do not convey enough information. We then discard the graphs that have a confidence rate cr lower than θ_{cr} since they are considered as unreliable. Finally, from the remaining graphs we only select graphs that have a Hamming distance rate hr lower than θ_{hr} since they are the most similar to the sensitive graph.

Mahalanobis selection method. Each learning graph, l_i, is represented by a 3 dimensional vector $V_{l_i} = [lr(l_i),\ cr(l_i),\ 1 - hr(l_i)]$. The relevance of the graph l_i is its Mahalanobis distance to the null vector [0,0,0]. It is computed as follows:

$$relevance(l_i) \quad = \sqrt{V_{l_i}^T \Sigma^{-1} V_{l_i}} \tag{6.8}$$

with Σ the 3×3 covariance matrix over the set of selected graph vectors. We select the top k most relevant graphs.

Discriminant graph approach to cleanse data

For specific attributes such as gender, age and relationship status, user profiles are never linked to multiple values. Moreover the sets of values for these particular attributes are much smaller than for other attributes. Consequently, the graphs that model these attributes are denser than the other graphs. For instance, the density of the graph that models gender (as most users publish their gender) is close to 0.5. In order to infer hidden links in such dense graphs we need to learn from dense graphs. However, most of the available learning graphs are sparse. To cope with this problem, we derive new dense graphs using clusters of nodes from different graphs. First, we select most discriminant graphs. Then, we cluster their nodes and use those clusters to create new dense graphs. We define 3 steps to create new dense graphs:

Step 1: Computing the discriminant rate of a learning graph w.r.t. a sensitive graph. Let $l = (U, V_l, L_l)$ (resp. $s = (U, V_s, L_s)$) be a learning (resp. sensitive) graph, V_l (resp.V_s) the set of learning (resp. sensitive) values. Let $\texttt{deg}_l(u)$ (resp. $\texttt{deg}_s(u)$) be the degree of node u in graph l (resp. s) and U_l (resp.U_s) be the set of users u with $\texttt{deg}_l(u) > 0$ (resp.$\texttt{deg}_s(u) > 0$).

The discriminant rate dr is the maximum on $v \in V_s$ of the ratio dr_v of user profiles in U_l that publish the sensitive value v among those that publish their sensitive value. dr is computed as follows:

$$\begin{aligned} dr(l) \quad &= \max_{v \in V_s} dr_v \\ &\texttt{where}\ dr_v = |U_l \ \cap\ \{u \mid (u,v) \in L_s\}| / |U_l \ \cap\ U_s| \end{aligned}$$

Figure 6.5 depicts an example where the relationship status attribute is sensitive. In this example, we have $lr(car) = 50\%$, $cr(car) = 75\%$ and $dr(car) = 66\%$. In fact, 66% of users who like car pages and that publish their relationship status are married.

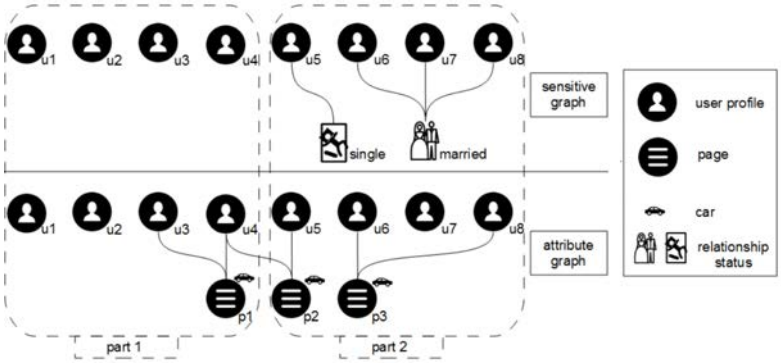


Figure 6.5: Splitting graphs for comparing them.

Step 2: Selecting most discriminant graphs from learning graphs w.r.t. a sensitive graph. In this step, we aim to select the best graphs for learning based on their learning rate, lr, confidence rate, cr and discriminant rate, dr.

Thresholds selection method. We define three thresholds for these criteria: θ_{lr}, θ_{cr} and θ_{dr}. We first discard the learning graphs that have a learning rate lr lower than threshold θ_{lr} since they do not convey enough information. We then discard the graphs that have a confidence rate cr lower than θ_{cr} since they are considered as unreliable. Finally, from the remaining graphs we only select graphs that have a discriminant rate dr higher than θ_{dr}.

Mahalanobis selection method. Each learning graph, l_i, is represented by a 3 dimensional vector $V_{l_i} = [lr(l_i),\ cr(l_i),\ dr(l_i)]$. The discrimination of the graph l_i is its Mahalanobis distance to the null vector [0,0,0]. It is computed as follows:

$$discrimination(l_i) \quad = \sqrt{V_{l_i}^T \Sigma^{-1} V_{l_i}} \tag{6.9}$$

with Σ the 3×3 covariance matrix over the set of selected graph vectors. We select the top k most discriminant graphs.

Step 2: Generating a dense graph from discriminant graphs. In order to build a new dense and promising graph for learning we first abstract selected discriminant learning graphs then we merge them. To abstract a selected discriminant learning graph we collapse all the attribute values into one super-value. User profiles remain unchanged and the links between them and the super-value in the abstracted graph are assigned a weight equal to the degrees of the corresponding user profiles in the original discriminant learning graph. Then we merge the abstracted graphs to build a new dense and promising graph for learning.

We can change the graph selection threshold (θ_{lr}, θ_{cr} and θ_{dr}) to generate a new dense graph for each threshold while avoiding selecting previously selected graphs.

Figure 6.6 depicts an example of generating a dense graph for gender inference. The selection criteria for this example are $\theta_{lr} = 100\%$, $\theta_{cr} = 55\%$ and $\theta_{dr} = 60\%$. In other words, all the profiles of users that do not publish their gender are connected in the selected graphs. At least 55% of the profiles of the users that publish their gender are connected in the selected graphs and at least 60% of user profiles connected in the selected graph and the gender graph are male or female. In this example the jewellery and the fast-food graphs are discriminant for the gender

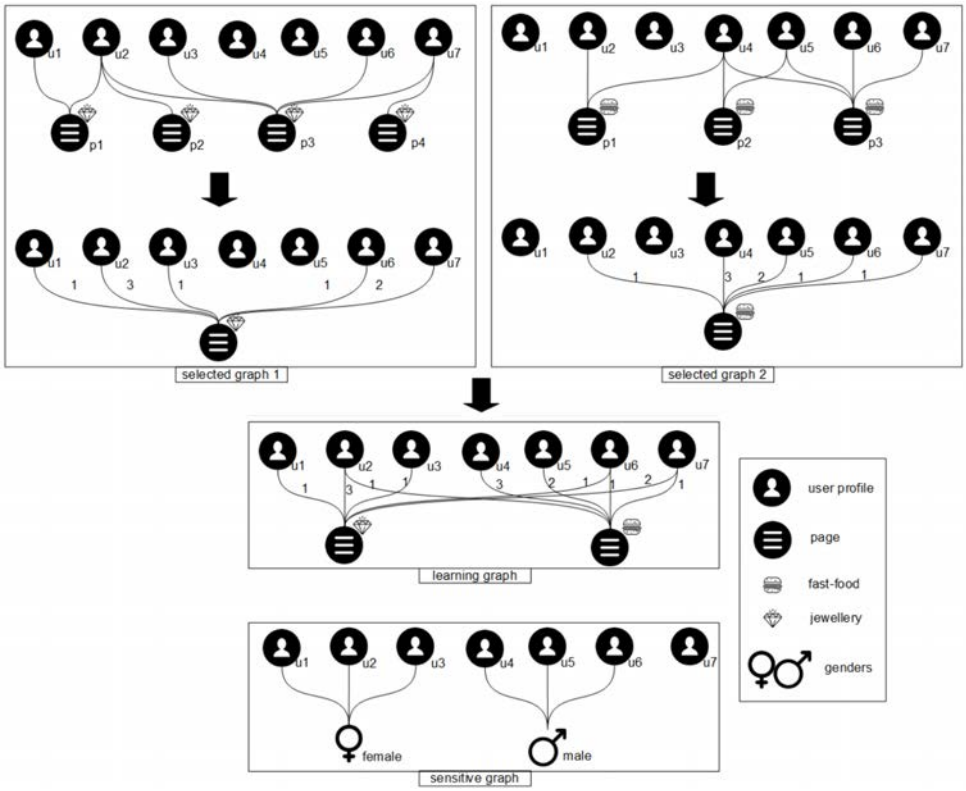


Figure 6.6: Example of creating a dense and discriminant graph for gender inference.

attribute. They are selected, abstracted and merged to create one dense and good graph. The gender of user profile u_7 can then be inferred based on this new graph.

6.6 Cleansing results

We analyse the distribution of learning graphs according to their learning rate lr, confidence rate cr, and Hamming distance rate hr (for relevant graphs) or discriminant rate dr (for discriminant graphs) with regard to a given sensitive graph. The dataset counts 1929 different graphs and 15012 totally crawled Facebook profiles. We give more details about this dataset in Annex A (dataset 1).

Relevant graphs

Among the 1929 analysed graphs we distinguish 4 graphs that are directly related to politics: the graph of politicians, the graph of political organizations, the graph of political parties and the graph of political ideologies. Table 6.3 details the sizes and the densities of these graphs. Since the graph of politicians contains a maximal number of nodes and holds more information about political views of the users, we compare the rest of the 3 graphs to it. Table 6.4 details the results of the comparisons according to the parameters lr, cr and hr.

We note that the learning rates lr of all graphs are low (below 3%), which means that these graphs do not provide extra information about users who do not publish their likes on the graph of politicians. In other words, users who mask their likes in the graph of politicians mask their

Graphs	♯ Connected user profiles	♯ Attribute values	Densities $\times 10^{-4}$
Politicians	2 554	4 589	9
Political Organizations	1 246	2 357	13
Political Parties	1 120	1 758	15
Political Ideologies	39	41	300

Table 6.3: Details about the graphs directly related to politics in the dataset.

Graphs	lr (in %)	cr (in %)	hr (in %)
Political Organizations	2.83	34.96	0.97
Political Parties	2.45	31.87	0.96
Political Ideologies	0.09	1.05	2.76

Table 6.4: Comparison of 3 graphs directly related to politics with the politicians graph.

likes in the three graphs of political organization, political parties and political ideologies too.

The confidence rates cr of the political organizations graph and the political parties graphs are high. About one third of users who publish their likes in the graph of the politicians publish their likes in those two graphs too. Moreover, the Hamming distance rate hr of those two graphs is low (below 1%), which means that these two graphs are very similar to the politicians graph. They are not useful for learning because they merely duplicate information.

The political ideologies graph has a very low learning rate (below 0.1%) and confidence rate (below 2%). Hence, this graph is not useful for learning. No conclusion can be drawn about its structural resemblance to the politicians graph.

Figure 6.7 gives the distribution of the 1928 learning graphs with regard to the Politicians graph. For this example, we notice the importance of the parameters lr, cr and hr in selecting the relevant graphs.

The variations of those parameters are large as detailed in Table 6.5. The distribution is 3 dimensional as shown by Figure 6.7. We particularly notice that the gender graph has high learning rate $lr = 72.31$ and confidence rate $cr = 83.47$ because 74.21% of users from the dataset publish their gender. However this graph has also a large Hamming distance rate $hr = 55.99$ and is not correlated to the politician graph.

We select the graphs that have a learning rate greater than $\theta_{lr} = 20\%$, a confidence rate greater than $\theta_{cr} = 60\%$ and a Hamming distance rate lower than $\theta_{hr} = 4\%$. Table 6.6 details the 23 selected graphs. The distribution of the selected relevant graphs according to the parameters lr, cr and hr is given in Figure 6.8.

Parameters (in %)	std	Mean	Max	Min
lr	4.43	1.13	88.37	0
cr	12.25	4.53	98.47	0
hr	7.65	2.35	100	0

Table 6.5: Distribution of the learning graphs parameters w.r.t. the sensitive graph Politicians.

We note that the communities graph has the second greatest learning rate $lr = 44.97\%$, which means that it holds much extra information about users who hide their likes in the sensitive graph. It also has the maximal confidence rate $cr = 98.47$ and the fifth lowest Hamming distance rate $hr = 1.75\%$ among the 23 selected relevant graphs, which means that its structure is very similar

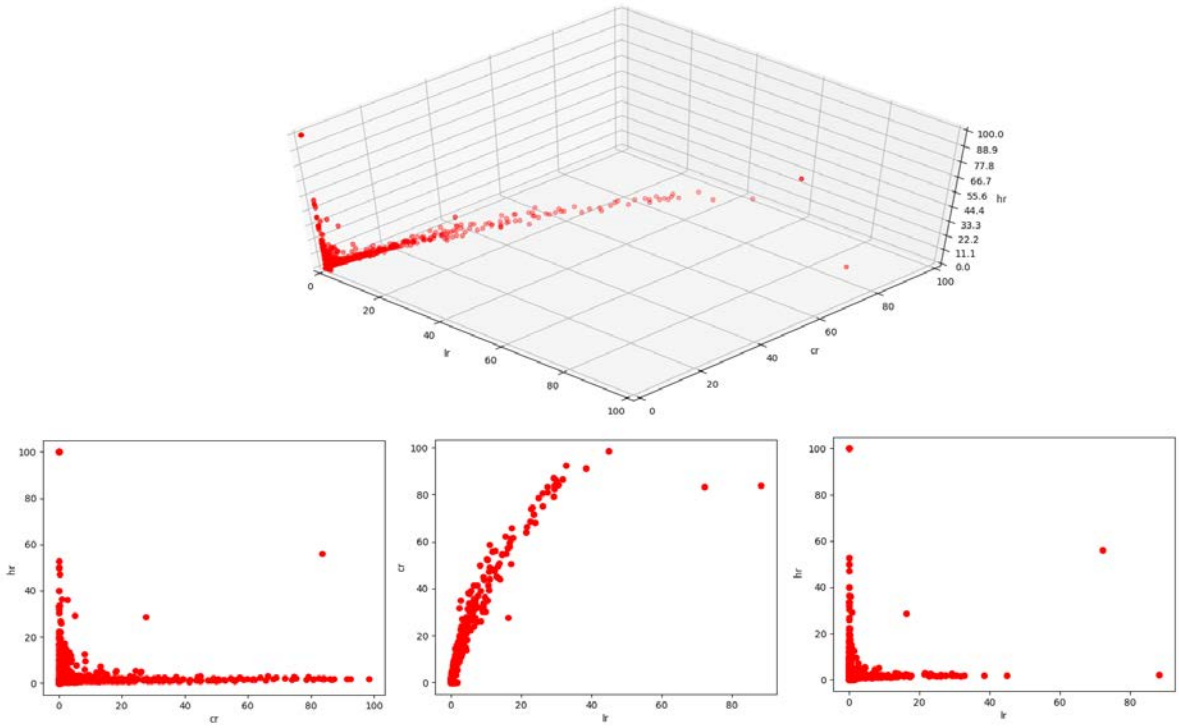


Figure 6.7: The distribution of the learning graph with regard to the sensitive graph Users-Politicians.

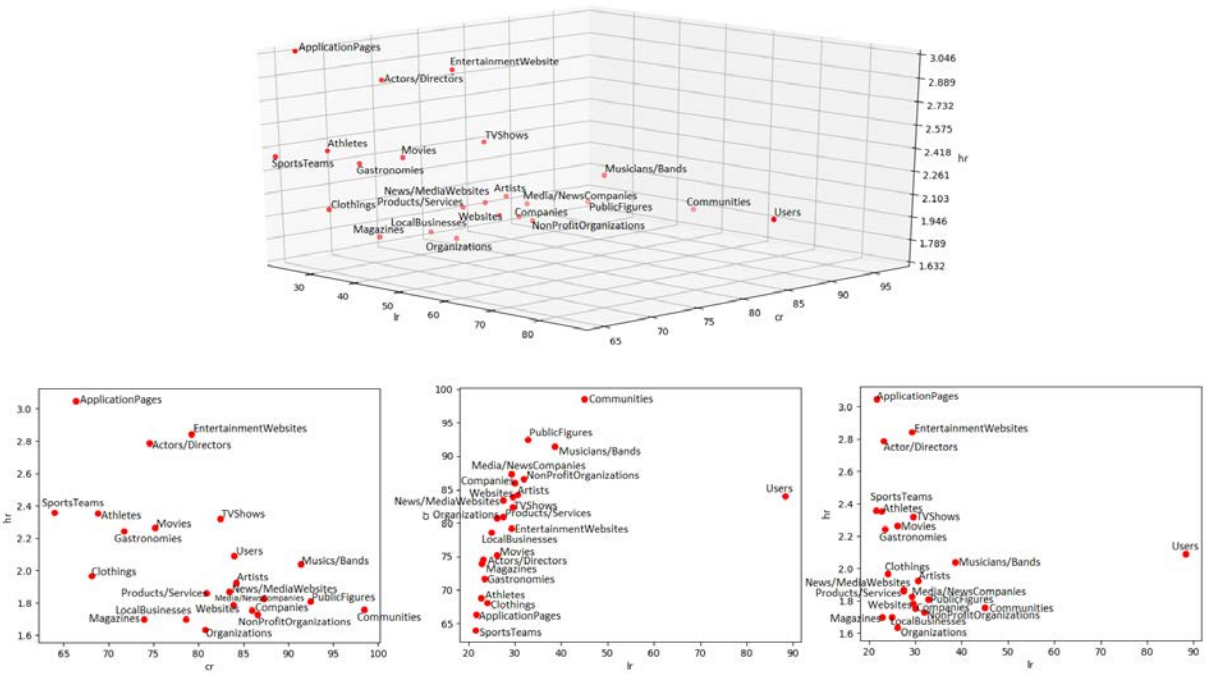


Figure 6.8: The distribution of the 23 selected relevant graphs with regard to the sensitive graph Users-Politicians.

Attribute graph	*lr* (in %)	*cr* (in %)	*hr* (in %)	♯ Connected user profiles	♯ Attributes values	Density $\times 10^{-4}$
Users	88.37	83.98	2.08	13 155	13 155	12
Communities	44.97	98.47	1.75	8 118	137 338	2
Musicians/Bands	38.58	91.38	2.03	7 141	84 762	4
Public Figures	32.86	92.44	1.80	6 455	28 289	3
Non Profit Organizations	31.85	86.57	1.72	6 180	25 847	3
Artists	30.65	84.22	1.92	5 970	31 681	3
Companies	30.05	85.94	1.75	5 939	20 750	3
Websites	29.57	83.94	1.78	5 829	17 931	3
TV Shows	29.48	82.41	2.31	5 778	11 876	6
Entertainment Websites	29.26	79.20	2.84	5 669	8 319	5
Media/News Companies	29.23	87.27	1.82	5 871	14 042	5
Products/Services	27.52	80.93	1.86	5 496	15 986	4
News/Media Websites	27.44	83.43	1.86	5 550	9 247	7
Organizations	26.20	80.77	1.63	5 328	14 738	3
Movies	26.09	75.17	2.26	5 171	16 282	6
Local Businesses	24.91	78.58	1.69	5 111	17 321	3
Clothings	23.99	68.12	1.96	4 729	16 090	4
Gastronomies	23.52	71.73	2.24	4 763	8 422	6
Actors/Directors	23.12	74.54	2.78	4 785	10 425	6
Magazines	22.82	73.96	1.69	4 733	9 955	6
Athletes	22.68	68.79	2.35	4 583	14 123	6
Application Pages	21.68	66.36	3.04	4 396	4 244	8
Sports Teams	21.48	63.93	2.35	4 309	10 433	5
std	13.42	8.58	0.38			
mean	30.71	80.09	2.07			
max	88.37	98.47	3.04			
min	21.48	63.93	1.63			
Total graph	93,66%			14 222	543 113	2

Table 6.6: Details about the distribution of the 23 selected graphs with regard to the sensitive graph Users-Politicians.

to the structure of the politicians graph. The friendship graph (Users-Users) has the maximal learning rate $lr = 88.37\%$ and a high confidence rate $cr = 83.98\%$ since 87.62% of users are connected to this graph. However, this graph has a Hamming distance rate $hr = 2.08\%$ greater than the mean hr of selected graphs which means that learned information from this graph is less reliable than the learned information from the communities graph.

Discriminant graphs

Gender. We analyse the distribution of learning graphs according to their learning rate lr, confidence rate cr, and discriminant rate dr with respect to the gender graph. The Dataset counts 1929 different graphs and 15012 totally crawled Facebook profiles. 74.21% of user profiles publish their gender and 59.69% of them are male. We give more details about this dataset in Annex A.

Figure 6.9 details the distribution of 1928 learning graphs with regard to the Users-Genders graph. For this example, we notice that the majority of attributes are not discriminant for the gender. A 57.52% proportion of the graphs have a discriminant rate below 60%. We recall that the discriminant gender rate of a given graph g, is zero when all user profiles that are connected

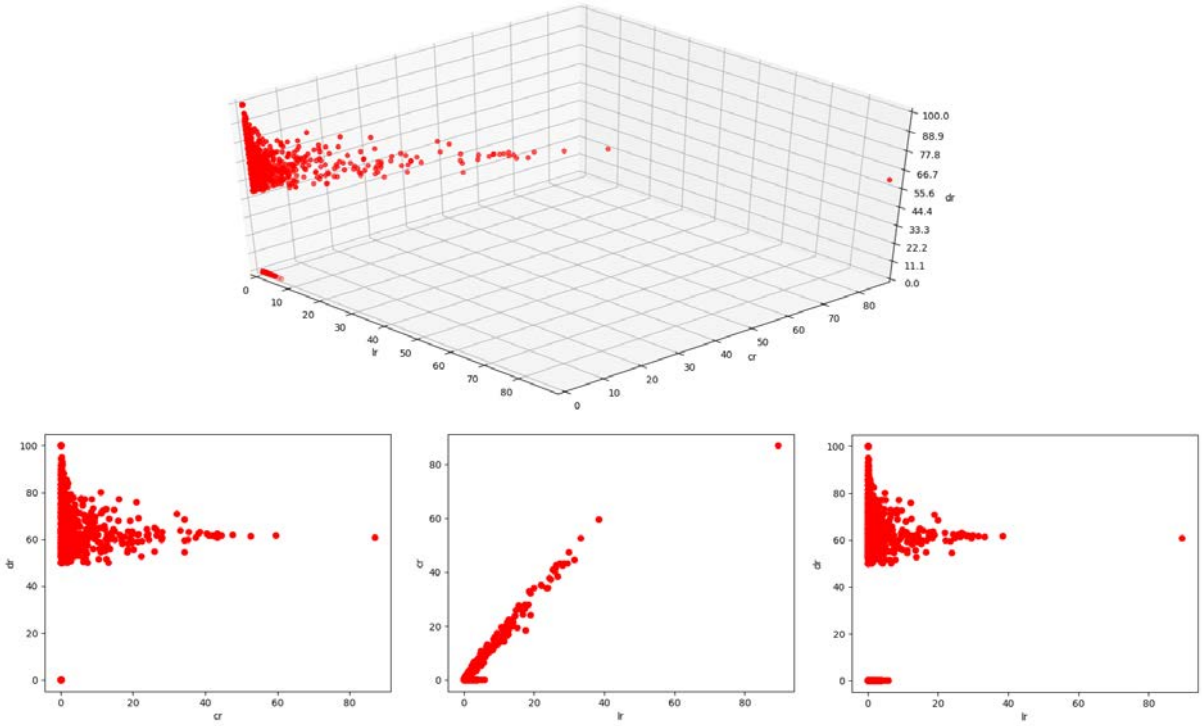


Figure 6.9: The distribution of the learning graph with regard to the sensitive graph Users-Genders.

to g do not publish their gender. Otherwise, its value is always above 50% as only two possible genders are possible on Facebook (male or female).

We note that the friendship graph (Users-Users) has the highest learning rate $lr = 89.53\%$ and the highest confidence rate $cr = 86.96\%$ since 87.62% of users are connected to this graph. However this graph has a low discriminant rate $dr = 60.71\%$ as 59.69% of users that publish their gender are male. Table 6.7 details the variation of the three parameters lr, cr and dr.

Parameters (in %)	std	Mean	Max	Min
lr	4.05	1.28	89.53	0
cr	5.85	1.81	86.96	0
dr	34.76	43.57	100	0

Table 6.7: Statistic details about the distribution of all the learning graphs with regard to the sensitive graph Users-Genders.

For this example, the graphs that have a high discriminant rate (above 75%) have a low learning and confidence rates (below 21%) as shown by Figure 6.9. Hence, we select the graphs that have a learning rate higher than $\theta_{lr} = 2\%$, a confidence rate higher than $\theta_{cr} = 5\%$ and a discriminant rate higher than $\theta_{dr} = 75\%$. Table 6.8 gives more details about the 8 selected graphs. The learning rate of the resulting dense graph is 22.93% and its confidence rate is 37.23%. The distribution of the selected discriminant graphs according to the parameters lr, cr and dr is detailed in Figure 6.10. We recall that our algorithms do not use any semantic information. They are first applied to anonymized labels and graph labels are de-anonymized at the end only for result presentation.

To increase the learning rate and to be able to infer the gender of more users, we decrease the discriminant rate threshold to generate more dense graphs with different criteria.

Attribute graph	*lr* (in %)	*cr* (in %)	*dr* (in %)	♯ Connected user profiles	♯ Attribute values	Density $\times 10^{-4}$
Sports Leagues	12.24	20.84	75.97 (M)	2 796	3 897	8
Recreation&Sports Websites	8.73	16.10	77.09 (M)	2 132	2 132	11
Software	4.96	8.55	77.23 (M)	1 145	1 247	18
Video Games	4.83	11.04	80.16 (M)	1 417	1 811	16
Outdoor&Sporting Goods	4.13	6.65	77.19 (M)	901	1 395	17
Women's Clothing Stores	3.69	5.41	77.28 (F)	746	1 066	18
Automotive	2.82	6.03	75.15 (M)	781	1 405	17
Vehicle Companies	2.27	5.64	77.39 (M)	716	1 300	19
std	3.15	5.30	1.34			
mean	5.46	10.03	77.18			
max	12.24	20.84	80.16			
min	2.27	5.41	75.14			
Created graph	22.93	37.23		5 036	14 253	2 639

Table 6.8: Details about the distribution of the 8 selected discriminant graphs with regard to the sensitive graph Users-Genders.

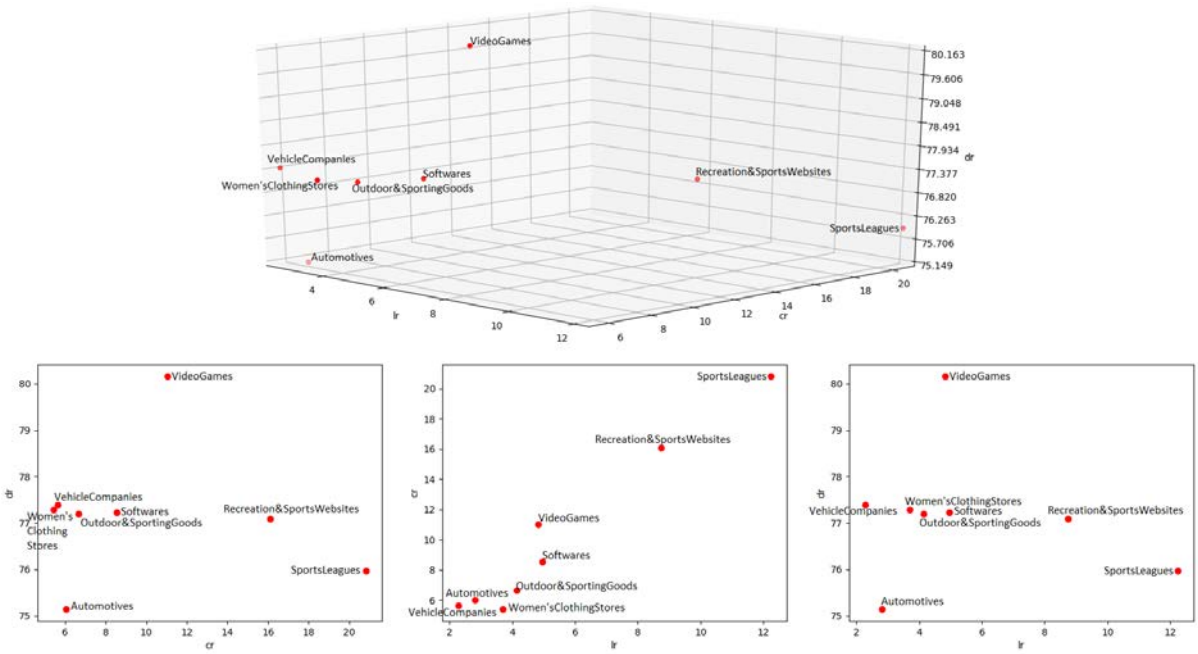


Figure 6.10: The distribution of the 8 selected discriminant graph with regard to the sensitive graph Users-Genders.

Relationship status. We analyse the distribution of learning graphs according to their learning rate *lr*, confidence rate *cr*, and discriminant rate *dr*, with regard to the relationship status graph. We conduct experiments on dataset 1 (see Annex A). Only 15.95% of user profiles publish their relationship status according to 11 possible options. Table 6.9 gives more details about the

relationship status of user profiles in the dataset. We notice that most users are single (42.83%) as the crawled profiles belong to students from University of Lorraine and few of their directly linked user profiles through groups memberships or friendships.

Relationship Status	Percentage of user profiles
Single	42.83
In a relationship	26.51
Married	19.12
Engaged	5.31
It's complicated	1.92
In an open relationship	1.42
Widowed	0.83
In a domestic partnership	0.72
Divorced	0.66
In a civil union	0.42
Separated	0.26
User profiles that publish their status	2 395 (15.95%)

Table 6.9: Distribution of relationship status of user that publish their status in the dataset.

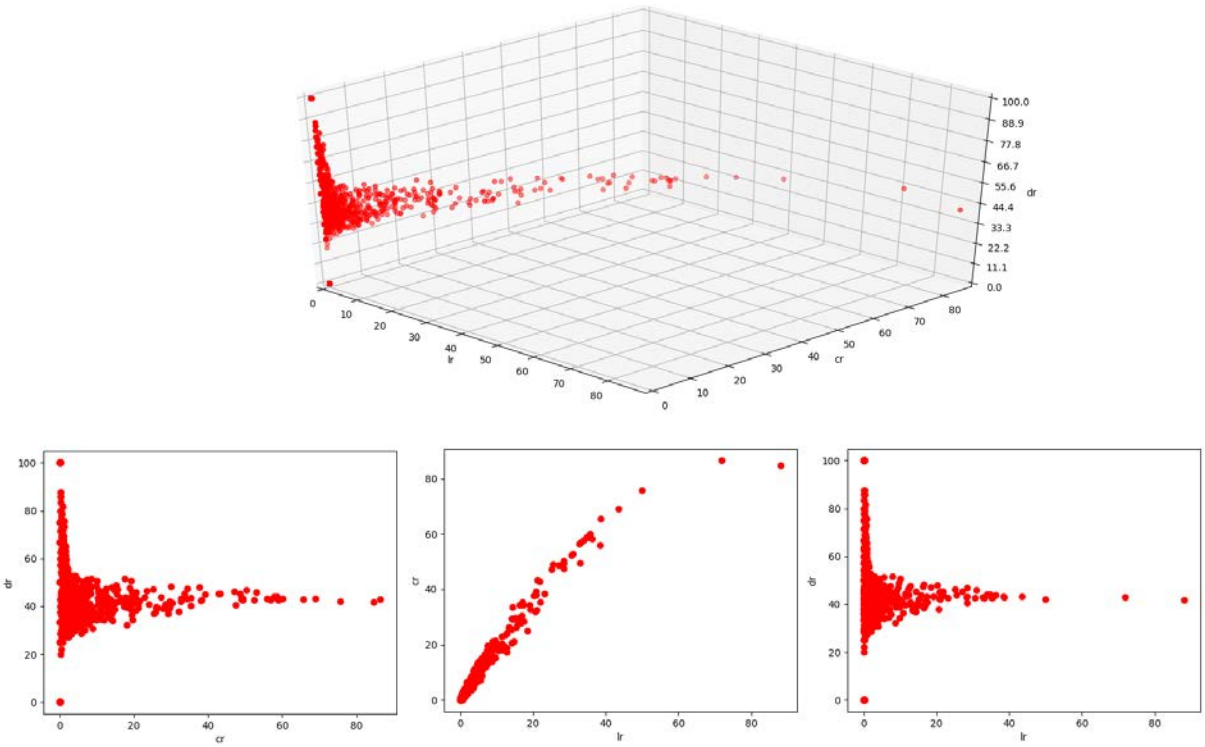


Figure 6.11: Distribution of the learning graph w.r.t. the sensitive graph Users - RelationshipsStatus.

Figure 6.11 gives the distribution of 1928 learning graphs with regard to the Users-RelationshipStatus graph. For this example we note that the majority of attributes are discriminant for the relationship status. 57.31% of the graphs have a discriminant rate above 40%. We recall

Parameters (in %)	std	Mean	Max	Min
lr	5.14	1.50	88.18	0
cr	8.22	2.80	86.47	0
dr	31.32	44.76	100	0

Table 6.10: Distribution of learning graph parameters w.r.t. the sensitive graph Users - RelationshipsStatus.

that the discriminant relationship status rate of a given graph g is zero when all user profiles that are connected to g do not publish their relationship status. Otherwise it is always above 9% as there are 11 possible genders on Facebook. Table 6.10 details the variation of the three parameters lr, cr and dr.

We note that the friendship graph (Users-Users) and the gender graph have the maximal learning rates, respectively $lr = 88.18\%$ and $lr = 71.89\%$. They also have the maximal confidence rates, respectively $cr = 84.72\%$ and $cr = 86.47\%$ since 87.62% of user profiles are connected to the friendship graph and 74.21% of user profiles publish their genders. However we note that their discriminant rates, respectively $dr = 41.74\%$ and $dr = 43.11\%$ are related to the value "single" and are very close to the percentage of single users in the dataset (42.83%). These results are specific to the sampled dataset. Moreover, as the cleanse algorithms do not rely on any semantic information, these graphs may not be discriminant for another sampled dataset where the percentages of relationship status of crawled user profiles are very similar. In fact in our case the cleanse algorithms detect that the majority of crawled users are single. As a consequence, the studied community contains many single users and this information is important for inferring the relationship status of users who hide their status in this community.

For this example we note that all graphs that have a learning rate greater than 10% are discriminant for the relationship status "single" as the majority of users in the dataset are single. However, many graphs with low learning rate (below 1%) are discriminant ($dr \geq 75\%$) for other relationships status. Table 6.11 gives more detail about selected graphs with two different criteria.

6.7 Conclusions

Cleansing the dataset is an important task. First, the number of graphs is large (1 929 graphs in dataset D1 in Annex A) and they contain many nodes (1 037 872 nodes in dataset D1 in Annex A). Consequently, training the inference algorithms on all graphs and nodes would be time consuming. Second, some graphs contain useless information about a particular targeted sensitive attribute.

We distinguish two types of attribute. The first type of attribute includes attributes that can have an arbitrary values. User profiles can be connected to several values of the same attribute such as book (literature) pages. The second type of attribute includes attributes that have predefined values: user profiles can be connected to at most one value of the same attribute such as the gender. The properties of the graphs that models those two types of attributes are different. Hence, we define two methods to cleanse the dataset with regard to the type of the sensitive attribute.

In the next chapter, we will analyse the selected and generated dense graphs in order to infer the sensitive values to which the target profile is most likely to be connected.

Criteria	Graphs	lr (in %)	cr (in %)	dr (in %)	♯ Connected user profiles	♯ Attribute values	Densities $\times 10^{-4}$
$\theta_{lr} = 20$ $\theta_{cr} = 35$ $\theta_{dr} = 45$	Movies	30.93	52.94	45.90 (Single)	5 171	16 282	6
	Actors/Directors	28.36	50.40	46.81 (Single)	4 785	1 0425	6
	Athletes	27.09	48.64	45.24 (Single)	4 583	14 123	6
	Sports Teams	25.21	47.10	46.37 (Single)	4 309	1 0433	5
	Animators	21.88	42.92	45.33 (Single)	3 788	6 826	5
	Just For Fun	21.15	43.22	45.22 (Single)	3 704	9 454	4
	Comedians	20.32	37.70	48.06 (Single)	3 467	3 580	8
7 selected graphs	std	3.73	4.81	0.97			
	mean	24.99	46.13	46.13			
	max	30.94	52.94	48.06			
	min	20.32	37.7	45.21			
	Created graph	47.20	74.27		7 735	71 123	5 505
$\theta_{lr} = 0.1$ $\theta_{cr} = 0.04$ $\theta_{dr} = 75$	Portuguese Restaurants	0.18	0.21	80 (Married)	28	32	368
	Watches	0.14	0.04	100 (Widowed)	19	13	891
	German Restaurants	0.13	0.29	75 (Married)	24	24	451
	Monarchs	0.13	0.13	100 (In a relationship)	19	16	658
	Internet Companies	0.12	0.04	100 (Widowed)	16	22	682
	Theme Parks	0.10	0.21	80 (In a relationship)	18	14	754
38 selected graphs	std	0.15	0.24	10.5			
	mean	0.21	0.29	87.01			
	max	0.86	1.21	100			
	min	0.1	0.04	75			
	Created graph	3.25	4.55		520	1 985	635

Table 6.11: Distribution of the selected discriminant graphs with regard to the sensitive graph Users-RelationshipsStatus.

7

Analysing cleansed data and inferring target sensitive values

Contents

7.1 Introduction **95**
7.2 Translating social attributed network to a text document **96**
7.3 Applying Word2Vec to compute node embeddings **98**
7.4 Inferring hidden sensitive values of the target user profile **102**
7.5 Measuring inference accuracy **104**
7.6 Experiments and results **105**
7.7 Conclusions **110**

7.1 Introduction

In the two previous chapters, we have modelled the social attributed network by graphs and selected the best graphs for learning.

In this chapter we aim to generate a text document that abstracts all information from the selected graphs. To that end, following [Perozzi et al., 2014] we perform random walks on all selected graphs and record the steps as words in a text document. We define two ways of performing random walks depending on the cleansed graphs (relevant graphs or discriminant graphs). Moreover, according to the importance (relevance/ or discrimination power) of graph, we quantify the amount of information used from it to produce the text document. Then, we assimilate the problem of inferring hidden sensitive links between the target user profile and the sensitive values to the problem of inferring missing words in a text document. Thus, we use a Word2Vec NLP model [Mikolov et al., 2013b, Goldberg and Levy, 2014, Rong, 2014] to analyse the text document in order to infer the closest sensitive words to the word that represent the target user profiles. Sensitive words represent the values of the sensitive attribute. Finally, we conducts experiments to infer the most probable pages of politicians to be liked by the target, his gender and relationship status. We use AUC, the Under the Receiver Operating Characteristic (ROC) Curve (AUC) as defined in [Gao et al., 2015] to measure the accuracy of the inferred links.

7.2 Translating social attributed network to a text document

We plan to translate latent information from the selected graphs to a document that will be processed as explained in the next section. The resulting document holds information about paths in the graphs and their frequencies.

Random walks for relevant graph approach

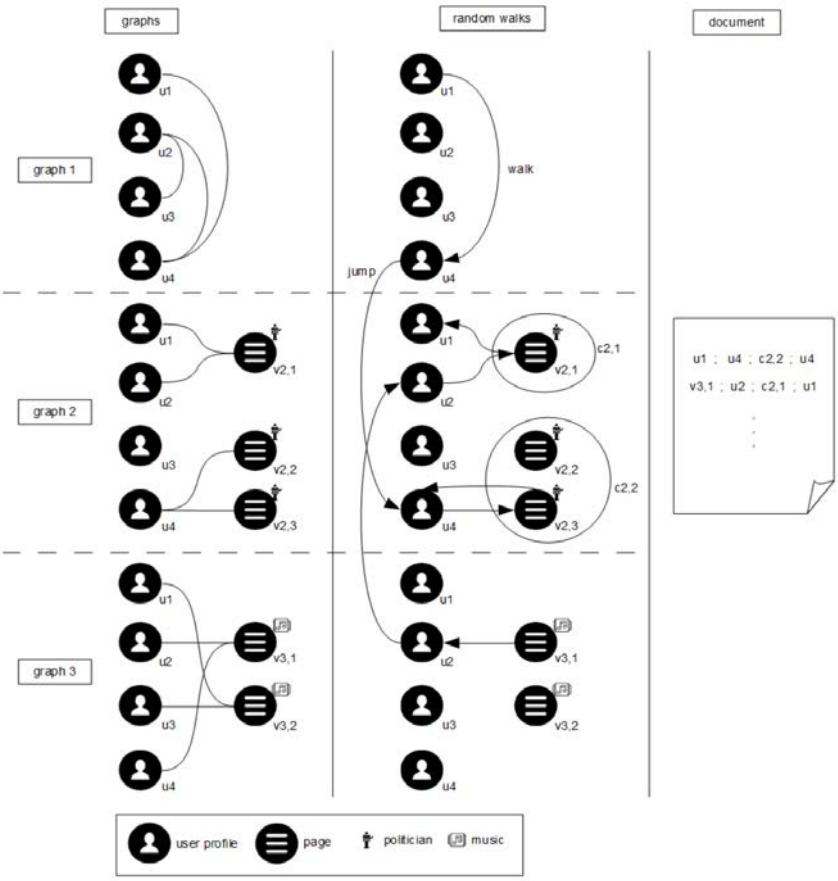


Figure 7.1: Example of multi graph random walk.

As illustrated in Figure 7.1 the document is constructed by connecting all graphs through random jumps between them and random steps between their nodes (see also [Perozzi et al., 2014]). For each step the walker state changes and a new word is written in the text document. A jump is only possible from a user node in a graph to the same user node (i.e. with the same label) in a different graph. A jump does not generate a word in the document. In this example we aim to infer the liked pages of politicians masked by user profile $u3$. The sensitive graph is Graph 2 and the learning graphs are Graph 1 and Graph 3. Since the values of the sensitive attributes (the pages of politicians in our example) are labelled (each value belongs to a unique cluster), they are represented by the label of their clusters in the final document. For instance the first walk depicted by Figure 7.1 is $[u_1, u_4, v_{2,3}, u_4]$. But for efficiency the walk $[u_1, u_4, c_{2,2}, u_4]$ is stored instead in the document since the value $v_{2,3}$ belongs to the cluster $c_{2,2}$.

Let n_g be the total number of selected graphs that model the social network. All graphs are

bipartite $G_x = (U, V_x, L_x)$ except the friendship graph if selected $G_1 = G_f = (U, F)$. Let U be the set of users in all graphs and n_u its cardinality. Jumps between two graphs G_x and G_y are possible if the current walker state is a user profile, say u_z, that has a non null degree in both graphs. The walker is allowed to jump from user node u_z to Graph G_y with a probability $p_{z,y}$. The probability $p_{z,y}$ is defined in Equation 7.1 where relevance is a parameter used to quantify the importance of each graph in inferring the secret values of the sensitive attribute of the target. It is computed as defined by equation 6.8. We also note $deg_x(u_z)$ the degree of user u_z in Graph G_x.

$$p_{z,y} = \begin{cases} \frac{relevance(G_y)}{\sum_{\{1 \leq x \leq n | deg_x(u_z) > 0\}} relevance(G_x)} & \text{if } deg_y(u_z) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

Jump and transition matrices. For each graph $G_y = (U, V_y, L_y)$ we define two line stochastic adjacency matrices, $T_{U \times V_y}$ and $T_{V_y \times U}$, and a jump matrix, J_y, that lead to G_y as detailed in 7.2.

$$J_y = diag(p_{z,y} | u_z \in U)$$

$$T_{U \times V_y}(i,j) = \begin{cases} \frac{1}{deg_y(u_i)} & \text{if}(u_i, v_j) \in L_y \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

$$T_{V_y \times U}(i,j) = \begin{cases} \frac{1}{deg(v_i)} & \text{if}(u_j, v_i) \in L_y \\ 0 & \text{otherwise} \end{cases}$$

where U is the set of all users in all graphs and $deg(v_i)$ is the degree of the value v_i.

For the friendship graph $G_1 = G_f = (U, F)$ we define a jump matrix J_1 in the same way as in Equation 7.2 but only one line stochastic adjacency matrix $T_{U \times U}$ as detailed in Equation 7.3.

$$T_{U \times U}(i,j) = \begin{cases} \frac{1}{deg_f(u_i)} & \text{if}(u_j, u_i) \in F \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

We define now a first order random walk where the next step probabilities depend only on the current location Given a source node S we perform a multi-graph random walk of fixed length l. Steps are generated by the distribution detailed in Expressions 7.4:

$$\forall k \in [2, l], P(s_k | s_{k-1}) = \begin{cases} p_{z,y} \times \frac{1}{deg_y(s_{k-1})} & \text{if } (s_{k-1}, s_k) \in L_y \\ & \text{and } s_{k-1} = u_z \text{ and } s_k \in V_y \\ p_{z,f} \times \frac{1}{deg_f(s_{k-1})} & \text{if } (s_{k-1}, s_k) \in F \\ & \text{and } s_{k-1} = u_z \text{ and } s_k \in U \\ \frac{1}{deg_y(s_{k-1})} & \text{if } (s_{k-1}, s_k) \in L_y \\ & \text{and } s_{k-1} \in V_y \text{ and } s_k \in U \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

The transition matrix is defined by blocks as follows:

$$T = \left[\begin{array}{c|ccccc} J_1 \times T_{U \times U} & J_2 \times T_{U \times V_2} & \cdots & J_i \times T_{U \times V_i} & \cdots & J_n \times T_{U \times V_n} \\ \hline T_{V_2 \times U} & & & & & \\ \cdots & & & & & \\ T_{V_i \times U} & & & 0 & & \\ \cdots & & & & & \\ T_{V_n \times U} & & & & & \end{array} \right]$$

For the example in Figure 7.1 the jump matrices and the right stochastic adjacency matrices are as following (assuming $relevance(G_1) = relevance(G_2) = relevance(G_3)$):

$J_1 = diag(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}, \frac{1}{3})$, $J_2 = J_3 = diag(\frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})$ and

$$T_{U\times U} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad T_{U\times V_2} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \qquad T_{U\times V_3} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$T_{V_2\times U} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad T_{V_3\times U} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Hence, the transition matrix is deduced as following:

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ v_{2,1} \\ v_{2,2} \\ v_{2,3} \\ v_{3,1} \\ v_{3,2} \end{array} \begin{array}{c} \begin{array}{ccccccccc} u_1 & u_2 & u_3 & u_4 & v_{2,1} & v_{2,2} & v_{2,3} & v_{3,1} & v_{3,2} \end{array} \\ \left[\begin{array}{cccc|ccc|cc} 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & 0 \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 0 & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 1 & & & 0 & & \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{2} & & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & & & & & \end{array}\right] \end{array}$$

Random walks for discriminant graph approach

We transform the newly created dense graph by the discriminant graph approach to cleanse data (Example: Figure 6.6) into an oriented and weighted social graph $G^{'}$. The weight of the link (u, v_i) from the user profile u to the super value v_i is $discriminant(l_i)$ and the weight of the link (v_i, u) is $\mathtt{deg}_i(u)$. $discriminant(l_i)$ is computed as defined by Equation 6.9 where l_i is a selected discriminant graph used to create the new dense graph and the values of the graph l_i are abstracted and represented by the value vi. Following [Perozzi et al., 2014], we perform random walks in $G^{'}$ in order to translate latent information from paths and their frequencies into a document.

For each step in the walk on $G^{'}$, the text document records the current node id, except when the node is a user profile whose sensitive value is published: in that case the sensitive value is recorded instead of the id.

7.3 Applying Word2Vec to compute node embeddings

In the previous section we have performed multi-graph random walks in the social network and generated a text document from these walks. Walks presented in the final document can be interpreted as sentences, where the words are network nodes. Hence, inferring a link between a user node and an attribute value node is similar to the natural language processing (NLP) problem of estimating the likelihood of words co-occurrence in a corpus.

We use a Word2Vec NLP model [Mikolov et al., 2013b, Goldberg and Levy, 2014] to encode the nodes in embeddings. Embeddings were first introduced in [Bengio et al., 2003]. The basic idea is to map one-hot encoded vectors that represent words in a high-dimensional vocabulary space to a continuous vector space with lower dimension. This approach has the virtue of storing the same information in a low-dimensional vector.

Word2Vec

Word2Vec relies on a secession of related models to create neural word embeddings [Rong, 2014]. Since learning word embeddings is unsupervised, different models are available to define the way Word2Vec learns the embedding of each word from the vocabulary. Models are selected according to the objectives. The skip-gram model aims to compute words embeddings in order to predict the context of a given word. However, the continuous bag of words (CBOW) model aims to compute words embeddings in order to predict a word given its context. A context of a given word is defined by the $c-1$ words surrounding it where c is the size of the window of the context. The order of context words is not important for the CBOW model. But it is important for the skip-gram model where less weight is given to the words that are distant from the target word [Mikolov et al., 2013a]. Hence, skip-gram model is slower than CBOW model but it is better for handling infrequent words.

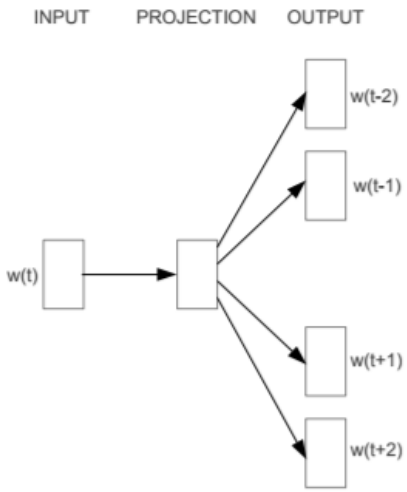


Figure 7.2: The skip-gram model [Mikolov et al., 2013a].

Skip-gram. Figure 7.2 details skip-gram model. For each target word $w(t)$ from the vocabulary taken as an input, this model returns a set of words of size c. This set of word is the context in which the target word passed as input is more likely to appear in the processed document. We note that each input word of the vocabulary is represented by a one-hot vector. This vector has $v-1$ zeros and a "1" in the position of the corresponding word, where v is the size of the vocabulary. The output of skip-gram model is actually one single vector of size v.

This vector represents a probability distribution of co-occurrence between the input word and each word from the vocabulary within a window of size c where $\lfloor \frac{c-1}{2} \rfloor$ is the maximum distance taken into consideration between the target input word and the other words from the document when computing co-occurrence probabilities.

The closer a word is to the target input word in the document, the higher is its corresponding probability in the output vector. The skip-gram model computes the best weight matrix between layers in the neural network in order to generate best probabilities distribution (as output vectors) that take into consideration all the computed co-occurrence probabilities. The objective function of the skip-gram model is given by Equation (4) in [Perozzi et al., 2014]. The skip-gram model has the advantage of generating good word representations [Mikolov et al., 2013b] and it shows good results when it comes to learning structural representations of nodes in a social network [Grover and Leskovec, 2016, Perozzi et al., 2014]. This model can be adapted to detect communities. It can infer a set of nodes as the output context (e.g. user profiles) that are more likely to be linked to a particular node given as input (e.g. value of an attribute).

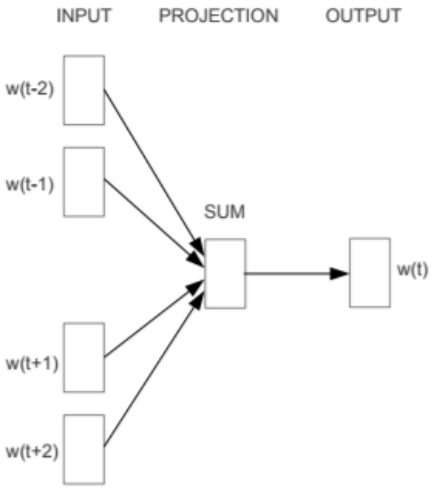


Figure 7.3: The CBOW model [Mikolov et al., 2013a].

CBOW. Figure 7.3 details the CBOW model. For each input (bag of words that are at distance c at most from each other), this model returns a single word $w(t)$. The bag of words passed as input is the context in which the target output word, $w(t)$, is more likely to appear in the processed document. We note that the input of the CBOW model is actually one single vector that averages all the vectors corresponding to the words in the context. This vector has $v - c$ zeros and $\frac{1}{c}$ in the position of the corresponding words of the context where v is the size of the vocabulary. The output of the CBOW model is also a single vector of size v. This vector represents a probability distribution of co-occurrence between all the words of the context and each word from the vocabulary within a window of size c.

The CBOW model can be adapted to infer the closest value of the sensitive attribute to a given user profile and his public values of attributes where the user profile and his public values of attributes represent the context and the inferred value is the output of the model.

Neural network. The neural network underlying Word2Vec is shallow for both models (skip-gram and CBOW). As depicted in Figure 7.4, it has one hidden layer. The weights between the input layer and the hidden layer are represented by a $v \times n$ matrix $W_{v \times n}$. The weights between the hidden layer and the output layer are represented by a $n \times v$ matrix $W'_{v \times n}$ where v is the size of the vocabulary and n is the number of neurons in the hidden layer. The main objective of the neural network is to learn the matrix $W_{v \times n}$. Each row of $W_{v \times n}$ is actually the embedding

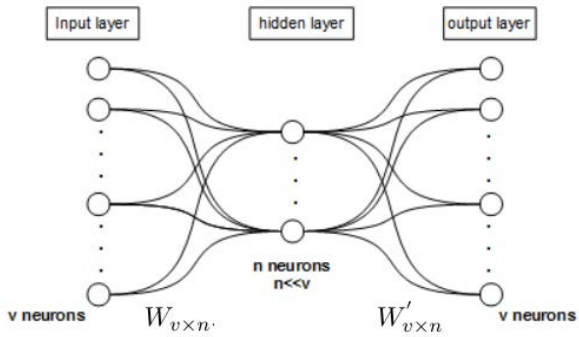


Figure 7.4: The neural network of Word2Vec.

of the corresponding word. The dimension of the embedding is n, the number of neurons in the hidden layer. So the neural network performs two tasks. First, it learns a good representation of words according to the chosen model (skip-gram or CBOW). Second, it reduces the dimension of the space in which the words are represented ($n << v$). Hence, the hidden layer can be seen as a projection layer.

Since running gradient descent to compute the embeddings of a huge vocabulary is very slow, the authors of [Schakel and Wilson, 2015] have proposed two training algorithms, hierarchical softmax and negative sampling.

Hierarchical softmax. The output of the neural network is a vector of dimension v. This vector represents a probability distribution. Probabilities are computed using the softmax function. However, the cost of computing the softmax for each entry in the output vector is huge. Besides many vectors must be computed along the learning process while changing the contexts and updating the weights.

The hierarchical softmax is an efficient way to compute the probability distribution. The hierarchical softmax algorithm decomposes the probabilities of observing each words into a sequence of probabilities. The sequence of probabilities determines to which group of words each word is more likely to belong. To organize these groups the algorithm uses a Huffman tree. The tree has v leafs and $v - 1$ inner units. The words of the vocabulary are the leafs. In this model the vectors of the output weights (hidden layer$\rightarrow$ output layer) represents the inner unit instead of the words.

Negative sampling. The negative sampling algorithm modifies the weight of only few selected words in the last layer each time. The weights of frequent words are more likely to be changed each time. Hence, hierarchical softmax gives better results for infrequent words. However, negative sampling gives better results for low dimensional embeddings.

Tuning the model and the algorithm for attribute2vec

In [Perozzi et al., 2014] and [Grover and Leskovec, 2016] the authors analysed friendships and used skip-gram model. In these works, the context is composed only of user profiles. Hence, user profiles that have similar friends will be mapped to similar embeddings. This model helps detect communities and it can be used to predict set of potential friends (output context) of a given user profile (input).

The main objective of our work is to infer the values of the sensitive attribute that are more likely to be the right values of the target user profile. Friends are considered as an attribute among others. The input of the model is the public values of attributes of the target user profile. Those values can be seen as the context in which the target user profile appears in the document text that traduces the social network. Moreover, there is no order between the attribute values of the user profile target. Which means that they can be seen as a bag of preferences. Hence, the CBOW model is more adequate for our goals.

Hyper-parameters The context window size is usually tuned between 5 and 10 for NLP tasks. However, unlike natural languages where word orders knuckle under grammar rules, users can befriend any other user on the social network without restrictions. Moreover, the degree of separation on social networks is low and it is equal to 3.5 on Facebook [Bhagat et al., 2016]. Hence, we limit the context window to only 3 in our case. In other words, the context of a given node in a social network is only composed of nodes that are directly linked to it.

The dimension of the embeddings is usually tuned between 100 and 300 for NLP tasks. However, the size of the vocabulary in social network (number of nodes) is much higher than the size of the vocabulary in natural languages. For instance English counts about 600k words (not roots). On the other hand, Facebook counts more than 2 billion monthly active users in 2018 [Noyes, 2018]. Hence, we use 512-dimensional Word2Vec embeddings to analyse social networks in the case of the relevance approach. On the other hand, we use 128-dimensional Word2Vec embedding in the case of discriminant approach since graphs are abstracted and the vocabulary is considerably reduced.

7.4 Inferring hidden sensitive values of the target user profile

Nodes (user profiles, clusters of values of attributes, values of attributes) are encoded by vectors (embeddings). The vectors of the sensitive nodes are ranked according to a distance measure to the node of the target node.

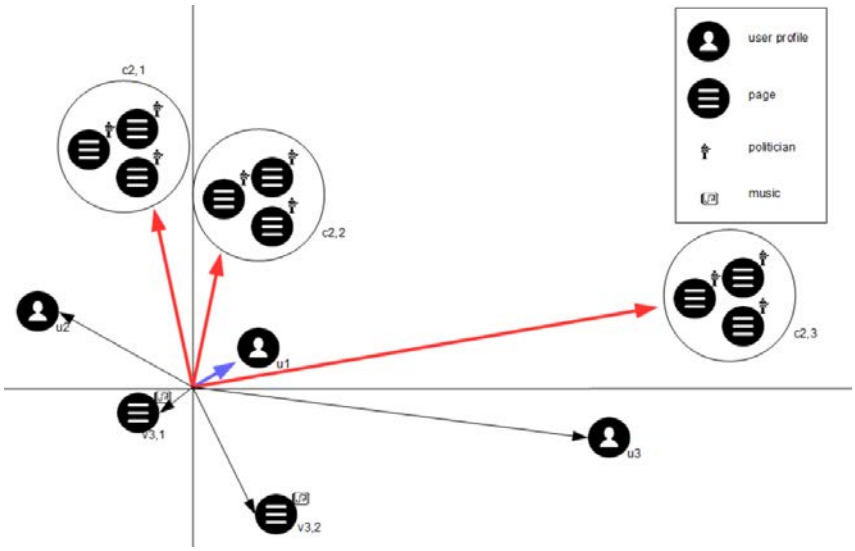


Figure 7.5: Example of 2-dimensional vectors that encode 8 nodes.

Figure 7.5 depicts an example of 2-dimensional vectors that encode 8 nodes: 3 user profiles

(u_1, u_2 and u_3), 2 pages of musics ($v_{3,1}$ and $v_{3,2}$) and 3 clusters of pages of politicians ($c_{2,1}$, $c_{2,2}$ and $c_{2,3}$). The clusters of the pages of politicians are the sensitive values and their vectors are red. The node u_1 is the target user profile and its vector is blue. The clusters of pages of politicians will be ranked according to their distances to u_1. The inference algorithm will infer as most probable pages of politicians to be liked by u_1, the pages of politicians of the cluster that has the smallest rank (the closest cluster to u_1).

In [Schakel and Wilson, 2015] Schakel et al. show that Word2Vec unsupervised learning algorithm encodes word semantics by affecting vectors in the same direction for co-occurrent words during training. Besides, the magnitude of a vector reflects both the frequency of appearance of related words in the corpus and the homogeneity of contexts where a context is a set of words that have high co-occurrence probability in the corpus.

In fact, the words that appear in the same contexts have small angular distances between them. The less overlapping the contexts are, the larger the angular distances between their different words are. However, words that appear in many contexts are represented by vectors that average vectors pointing in many contexts directions. Hence, the vectors magnitude generally decreases with respect to the number of contexts. Moreover, the higher the word frequency is, the higher the chance that it is used in different contexts is. Consequently, the vector magnitude also decreases with respect to frequency. From these remarks, we conclude that the euclidean distance is not a good measure for our inference purpose. Actually, words that appear in many contexts have low magnitude. As a result, their euclidean distances will be small and, using this criteria, they would be considered close even if they do not appear in any common context. For instance, the euclidean distance between the cluster of pages of the most popular politicians will be small even if they are rivals. In the example depicted by Figure 7.5 the politicians of the clusters $c_{2,1}$ and $c_{2,2}$ are rivals. The angular distance between those two clusters is big. However, the euclidean distance is small. Moreover, the euclidean distance between a user that has many friends, for instance the user u_1 in Figure 7.5, and a popular music like "despacito", for instance the page of music $v_{2,1}$ in Figure 7.5, will be small. But popular users do not necessarily like popular musics.

Ranking values/clusters of values

In this work we focus on angular distance between vectors since it holds information about contexts. To measure semantic similarity between nodes we apply cosine similarity which is widely used in NLP. This metric measures the cosine of the angle formed by two vectors which represent two different nodes. It yields values in the interval $[-1, 1]$ that quantifies the contextual similarity between nodes regardless their *centrality*. We recall that in a social network the centrality of a node quantifies its importance. In our case, the centrality of a node is its frequency in the generated text document by random walking.

We rank all the sensitive values (or clusters of values) by cosine similarity to the target user profile. The lowest the cosine similarity is, the lowest the rank of the corresponding values (or clusters of values) is. The values that have the lowest rank are more likely to be the secret values of the target user profile from all the values of the sensitive attribute. Secret values are actually the true values of the target user but are not published by him on the social network.

Detecting target interest in the sensitive attribute

To investigate the interest of the target user in the sensitive attribute, we first check if he has a particular interest in it in general. If the mean cosine similarity between the vector of the target

profile and all the vectors of the values (or clusters of values) of the sensitive attribute is positive, then we presume that the target is interested in that attribute. The highest the mean is, the highest the interest is. In the example depicted by Figure 7.5 the user u_1 has a particular interest in politics since the angles between his vector and all the vectors of the clusters of politicians are acute. Consequently, the mean cosine similarity is positive. However, he has no interest in musics since the angles between his vector and all the vectors of the musics are obtuse. Consequently, the mean cosine similarity is negative.

We stress that the measured interests are relative. For the example depicted by Figure 7.5, we conclude that the user u_1 has a high interest in politics when the algorithm is trained on graphs of friendships, liked pages of musics and liked pages of politicians. However, u_1 may have higher interest in movies. When training the algorithm about liked pages of politicians and liked pages of movies, his vector may point to movies contexts. In that case, the mean cosine similarity to clusters of politicians may be negative.

Secondly, we check if the target user has a particular interest in some values of attributes among all the values of the sensitive attributes. All the values (or clusters of values) that have a cosine similarity higher than 0.5 are considered as particular values that interest the target.

For some specific attributes such as gender and relationship status we do not measure the interest of the target since all users have one gender and one relationship status in the social network, and those kind of attributes are not subject of interest (i.e. the network do not provide means to express it).

7.5 Measuring inference accuracy

We use the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) as defined in [Gao et al., 2015] to measure the accuracy of the inferred links. The amount that AUC exceeds 0.5 tells how much the inference algorithm is better than random guessing. The AUC for link prediction problem is computed as following:

$$\frac{nr_{(nel>esl)} + 0.5 \times nr_{(nel=esl)}}{n_{nel} \times n_{esl}}$$

where n_{nel} is the number of not existing links, n_{esl} is the number of existing but secret links, $nr_{(nel>sl)}$ is the number of couples of a not existing link and a secret link of smaller rank, $nr_{(nel=esl)}$ is the number couples of a not existing link and a secret link of the same rank. Note that AUC value will be 0.5 if the ranks are independent and identically distributed.

In our model, links between the targeted user profile and all the values of the sensitive attribute which belong to the same cluster will have the same rank. Assuming that all clusters have different ranks (the cosines are coded on 2 bytes in an euclidean space of dimension 512 where vectors are coded on 1024 bytes) the AUC can be computed as following:

$$AUC = AUC_1 + AUC_2 \times \frac{nr_{(nel=esl)}}{n_{nel} \times n_{esl}}$$
$$AUC_1 = \frac{nr_{(nel>esl)}}{n_{nel} \times n_{esl}}$$

where AUC_1 is the accuracy of ranking clusters and AUC_2 is the accuracy of ranking values inside the selected cluster c_s (that should contain the secretly preferred value). We make the following approximations when the goal is to predict one given secret link at a time ($n_{esl} = 1$).

$$n_{nel} \times n_{sl} \simeq m - 1$$
$$nr_{(nel=esl)} \simeq |c_s| - 1$$
$$AUC \simeq AUC_1 + AUC_2 \times \frac{|c_s|-1}{m-1}$$

For instance the number of pages of politicians in the dataset D1 detailed in Annex A is $m = |Politicians| = 4589$ and $\sqrt{m} - 1 \leq |c_s| - 1 \leq 2\sqrt{m} - 1$ we have

$$0.014 = \frac{1}{\sqrt{m}+1} = \frac{\sqrt{m}-1}{m-1} \leq \frac{|c_s|-1}{m-1} \leq 2\frac{\sqrt{m}-1}{m-1} = 2\frac{1}{\sqrt{m}+1} = 0.029$$

For all the results in next section the rank inside clusters is generated by independent and identical distribution ($AUC_2 = 0.5$).Therefore $AUC_2 \times \frac{|c_s|-1}{m-1}$ is negligible w.r.t. AUC_1 in that case and does not affect the global accuracy of the prediction.

7.6 Experiments and results

Experiments

We have conducted several experiments on two datasets D1 and D2 detailed in Annex A. For each experiment we generate a new social graph from the dataset by selecting the user profiles (targets) that publish their preferences concerning the sensitive attribute and at least one other attribute (friends are considered as an attribute too). Then we remove all the links in the sensitive graph of 10% of the selected user profiles (targets). The algorithm makes sure that all the nodes in the resulting social graph remain connected. The experiments have consisted then in inferring the hidden links based on information from the learning graphs.

The clusters of sensitive values of attributes are computed from the new social network after deleting sensitive links.

Results

Politicians. We conduct an experiment on Dataset D1. The sensitive graph models the links between 2 554 user profiles and 4 589 pages of politicians. We have used relevant graph approach to cleanse the data as detailed in the previous chapter. From the 1 928 learning graphs, we selected the graphs that have a learning rate greater than $\theta_{lr} = 20\%$, a confidence rate greater than $\theta_{cr} = 60\%$ and a Hamming distance rate lower than $\theta_{hr} = 4\%$. Table 6.6 details the 23 selected graphs relevance measures. For the defined thresholds ($\theta_{lr} = 20\%$, $\theta_{cr} = 60\%$, $\theta_{hr} = 4\%$) the precision is equal to 0.79. However, inference accuracy when the 23 relevant graphs are selected randomly is only 0.41. We conducted more tests by selecting manually 3 graphs that are semantically close to politics as follows. Graph G_1 models the links between 1 246 user profiles and 2 357 political organizations, G_2 models the links between 1 120 user profiles and 1 758 political parties and G_3 models the links between 39 user profiles and 41 political ideologies. Although the selected graphs seem promising, the inference accuracy is only 0.46. This can be explained by the fact that the selected graphs are very sparse and users are vigilant when publishing their preferences about those attributes. Consequently, the algorithm cannot learn well from them.

As music and politics are empirically known to be correlated [Street, 2012], we check the ability of our algorithms to infer the preferred politicians of Facebook users based only on the music graphs. We selected only two graphs for learning. G_1 models the links between 802 user profiles and 477 musical genres, and G_2 models the links between 7 141 user profiles and 84 762

Selection method	accuracy	♯ targets	♯ deleted links	♯ nodes
Relevance-based selection (23 graphs)	0.79	252	409	558125
Music selection (2 graphs)	0.62	233	379	100251
Politic selection (3 graphs)	0.46	123	297	19168
Random selection (23 graphs)	0.41	204 (average)	351 (average)	11200 (average)

Table 7.1: Experimental results.

musicians/bands. The inference accuracy is equal to 0.62 for this experiment. We note that the musicians/bands graph was automatically selected by our relevance-based selection method. Table 7.1 summarizes the results of the conducted experiments.

Relationship status. The sensitive graph models the relationship status of user profiles. We have used the discriminant graph approach to cleanse the data as detailed in the previous chapter. We notice that most users that publish their relationship status are single, married or in relationship. For instance, only 10 users publish that they are in civil union in Dataset D1 (see Annex A). Only 3 users publish that they are in domestic partnership in Dataset D2 (see Annex A). To simplify the presentation we define two meta-relationship status as follows:

$$\begin{aligned} R1 = & \{Single, Divorced, Separated, Widowed, Complicated\} \\ R2 = & \{Domestic\ partnership, Married, Engaged, \\ & Relationship, Civil\ union, Open\ relation\} \end{aligned}$$

Meta-relationship status		R1	R2
♯ user profiles	D 1	1 114	1 281
	D 2	208	783

Table 7.2: Relationship status of users in the datasets.

We aim to infer the meta-relationship status of users. Table 7.3 gives more details about the selected attributes from dataset D2.

We notice that discriminant attributes toward $R1$ are focused around educations and leisures. On the other hand, discriminant attributes toward $R2$ are focused around businesses. The accuracy in AUC of inferring the meta-relationship status is higher than 0.7 in both datasets D1 and D2 as soon as the target publishes values concerning at least 4 selected attributes by the cleanser.

Genders. The sensitive graph models the gender of user profiles. We have used discriminant graph approach to cleanse the data as detailed in the previous chapter. Table 7.4 gives details about the sensitive graph in the two crawled datasets. Table 7.5 gives details about the selected attributes in Dataset D1.

We notice that discriminant attributes toward male are focused around sports, games and software. On the other hand, discriminant attributes toward female are focused around health, home and luxury. The accuracy in AUC of inferring the gender is higher than 0.83 in datasets

Attribute	Discrimination	Meta-relationship status
Education	2.75	88.41 % R1
Community College	2.74	90.02 % R1
Consulting Agency	2.71	90.70 % R2
Sports & Recreation	2.56	91.18 % R2
Home & Garden Website	2.49	91.89 % R2
Automotive, Aircraft & Boat	2.48	92.86 % R2
Locality	2.47	92.59 % R2
Corporate Office	2.46	91.18 % R2
News & Media Website	2.42	90.32 % R2
Financial Service	2.41	90.00 % R2
Industrial Company	2.40	89.29 % R2
Educational Consultant	2.02	75.00 % R1
Playground	1.80	66.67 % R1
Phone/Tablet	1.70	63.64 % R1
Plastic Surgeon	1.60	60.00 % R1
Consulate & Embassy	1.60	60.00 % R2
School Sports Team	1.53	52.00 % R1
Dive Bar	1.45	54.55 % R1
Video	1.44	51.00 % R1
Playlist	1.41	53.04 % R1

Table 7.3: Selected attributes in D2 for relationship status inference.

Genders		Female	Male
♯ user profiles	D 1	4 491	6 650
	D 2	1 606	2 991

Table 7.4: Genders of users in the datasets.

D1 and higher than 0.67 for the dataset D2 as soon as the target publishes values concerning at least 2 selected attributes by the cleanser.

Processing times Table 7.6 displays the processing times. The clock speed of the processor is 2.3 GHz. Cleansing and random walk algorithms are not paralleled Cleansing takes more time than the other processes in the case of discriminant approach. In fact, it handles hundreds of thousands of nodes, compares hundreds of graphs to the sensitive graph and compute their importance. The random walk, in the case of discriminant approach, is performed on a small graph containing only few tens of super-values and few thousands of user profiles. On the other hand, in the case of relevant approach, it is performed on bigger graphs containing tens of thousands of values. The machine dispose only of 8GB of RAM memory. Each chunk of 5k steps, about 25MB, is stored separately in a text file. Those files are then parsed by Word2Vec. Word2Vec speed depends on the size of vocabulary. It is fast in the case of gender inference since the vocabulary in the document is limited to only user profiles, super-values and sensitive values.

Ranking has to compute cosine similarity of only few vectors that represent the sensitive values to the vector of the target user. Cleansing tasks allow to select only important attributes and reduce the vocabulary. Consequently, it speeds up the inference tasks (random walk and

Attribute	Discrimination	Genders
Sports League	4.22	75.97 % Male
Recreation & Sports Website	3.80	77.09 % Male
Video Game	3.66	80.16 % Male
Cars	3.25	73.15 % Male
Amateur Sports Team	3.03	72.86 % Male
Sport	2.80	73.07 % Male
Jewelry/Watches	2.72	56.26 % Female
Electronics	2.68	73.19 % Male
Software	2.52	77.23 % Male
Outdoor & Sporting Goods	2.35	77.19 % Male
Women's Clothing Store	2.35	77.28 % Female
Home Decor	2.29	54.60 % Female
Stadium, Arena & Sports Venue	2.28	74.45 % Male
Baby Goods/Kids Goods	2.14	66.61 % Female
Kitchen/Cooking	2.08	55.93 % Female
Bags/Luggage	2.04	59.16 % Female
Beauty, Cosmetic & Personal Care	2.03	60.59 % Female
Cosmetics Store	1.98	66.25 % Female
Hair Salon	1.92	61.44 % Female
Home & Garden Website	1.72	55.18 % Female

Table 7.5: Selected attributes in D1 for gender inference.

Process			Cleansing	Random walk	Word2Vec	Ranking
Time (in seconds)	Discriminant approach	D1	423	34	50	0.12
		D2	243	25	30	0.12
	Relevant approach	D1	782	523	924	1
		D2	574	451	733	1

Table 7.6: Processing times.

Word2Vec). Moreover, it increases the accuracy by discarding irrelevant information.

Parameter sensitivity analysis

Let us investigate the impact of the cleansing parameters (lr, cr and hr). All experiments detailed in this section are conducted on dataset D1 to infer the political orientation of users.

Table 7.7 shows that only 3 graphs among the 1928 available graphs have a learning rate lr higher than 30%. Based on those graphs, inference accuracy can be very low. For instance, inference accuracy based on gender attribute is only 0.36. Based only on the users (i.e. friendship) graph accuracy is getting better: 0.64. The communities graph gives high accuracy for inferring political views: 0.74. However, we notice that the best accuracy is obtained when selecting graphs with learning rate between 10% and 40%. Table 7.7 shows that the learning rate parameter lr is important to select the best graphs for inference. However, accuracy does not depend only on this parameter since some graphs such as gender graph that have high learning rate may lead to very low accuracy results.

lr (in %)	Inference accuracy in AUC	♯ selected graphs	♯ attacked targets	♯ masked links
[0, 10[	0.61	1873	252	409
[10, 20[	0.80	31	254	411
[20, 30[	0.86	16	254	418
[30, 40[	0.80	5	253	410
[40, 50[	0.74	1 (Communities)	251	408
[70, 80[	0.36	1 (Genders)	213	353
[80, 90]	0.64	1 (Users)	214	350

Table 7.7: Impact of *lr* on inference accuracy.

hr (in %)	Inference accuracy in AUC	♯ selected graphs	♯ attacked targets	♯ masked links
[0, 5[	0.68	1744	253	410
[5, 10[	0.59	87	177	304
[10, 20[	0.53	58	87	167
[20, 30[	0.45	11	83	129
[30, 40[	0.42	13	11	21
[40, 50[	0.42	5	2	3
[50, 100]	0.41	10	211	351

Table 7.8: Impact of *hr* on inference accuracy.

Table 7.8 shows that when the Hamming distance rate *hr* decreases, accuracy increases. However, most graphs have a low Hamming distance rate because only small part of them can be compared to the sensitive graph, as few users publish their preferences in both graphs. Hence, their structure is not fairly comparable to the politicians' graph structure. To cope with this problem we compute a third parameter: the confidence rate, *cr*, that indicates how reliable the structure comparison is.

cr (in %)	Inference accuracy in AUC	♯ selected graphs	♯ attacked targets	♯ masked links
[0, 10[	0.63	1711	245	409
[10, 20[	0.43	94	246	407
[20, 30[	0.74	37	245	405
[30, 40[	0.54	28	248	404
[40, 50[	0.72	16	247	410
[50, 60[	0.38	14	250	407
[60, 70[	0.63	8	248	403
[70, 80[	0.60	6	248	393
[80, 90[	0.65	11	255	419
[90, 100]	0.82	3	253	410

Table 7.9: Impact of *cr* on inference accuracy.

Table 7.9 shows that the confidence rate, *cr*, does not give information about the best graph to select when it is considered alone. It must be coupled with other parameters. For instance, if

a given graph g has a high confidence rate but a low Hamming distance rate that means that it is a good graph for inference. However, if a given graph g has a high confidence rate and high Hamming distance rate that means that g is probably a bad graph for inferring the sensitive attribute.

7.7 Conclusions

In this chapter, we have translated the latent information from the selected graph by the cleanser to a document text. Then we have used NLP techniques (Word2Vec algorithms) in order to infer the hidden links between the target user profile and the sensitive values.

Sensitive data inferences are fast and accurate (AUC > 0.67). Moreover, the algorithms are able to automatically generate rules about correlated attributes and quantify their importances in learning tasks. Rules are generated depending only on the structure of the social network of the target himself. Hence, we avoid generating general rules that may not be available for all different communities of users.

For the conducted experiments, we note that the friendship graph was not selected among important ones to deduce both the gender and the relationship status of users. This probably means that alternative techniques based on homophily would be inaccurate in this context.

We have observed that the privacy of users toward a given sensitive attribute, s, is threatened (inference accuracy AUC > 0.67) as soon as they start publishing at least three important correlated attributes to s from a set of 20 selected important attributes.

8

Conclusions and perspectives

> You have to fight for your privacy or you lose it.
>
> *Eric Schmidt*

In this work we have analysed privacy leakage in on-line social networks. First, we have introduced a measure of sensitivity of discussed subjects on social media. The most sensitive subjects according to the behaviours of french participants in our study are *Religion, Money, Politics, Dating, Shopping and Health.* Then, we have studied information leakage. In order to infer sensitive information about a given target, we first disclose his local network (1-hop nodes from the target). To that end, we have designed and tested on-line link disclosure attacks with certainty. We have carried out several attacks on real Facebook profiles. We conclude that adversary can easily and rapidly disclose hidden links (friendship and group membership) with certainty taking advantage of social network APIs. We have also exploited interest groups in order to carry out link disclosure attacks. In order to start the attack, the target must publish his membership to at least one group. However, it is possible to extend this starting condition and use interactions (comments, tags, share, likes ...) as starting line of search. We have also designed a sampling algorithm to collect the most important data around a given target. The collected data are then processed to infer the values of sensitive attributes. Our cleansing algorithms show that it is possible to detect and quantify the correlation between subjects based only on structural information. Our experiments show that fast and simple algorithms for comparing graph structures can detect hidden correlation in the behaviour of users about semantically different attributes such as politics, musics and gastronomies. We have used Jaccard index to tune a fast algorithm of graph comparison. However, it would be interesting to try other indexes (Katz, Adamic/Adar, Common Neighbours...). We have also designed a clustering algorithm in order to group values of attributes by similarity of preferences. Our clustering algorithm uses Jaccard index to compute similarities. Again, it would be interesting to investigate the performance of alternative indexes.

Our work does not exploit semantics information when processing data. We rely on Facebook algorithms to check the page types. However, we have noticed that some pages are mislabelled. Moreover, we have noticed that some types of pages are very close and can be clustered. For instance, a health cluster may includes medical equipment shops, acupuncturists and medical services. To this end, in future work we plan to introduce natural language processing techniques to help clustering attributes and checking their values. Taking into consideration the correlation between attributes considerably speed up inference and increases accuracy. We have used Maha-

lanobis distance to weight the correlation between attributes. However, our parameter sensitivity analysis shows that correlation measure can be improved to get better accuracy. We envisage to investigate the impact of other weighting methods on inference accuracy.

The algorithms we propose are embedded in a user friendly system called SONSAI. SONSAI can be installed on any commercial laptop with Windows OS. It contains very few parameters to set and is designed to be used with basic IT knowledge. It permits users to audit their local networks and rapidly detect potential privacy leakage with good accuracy. The architecture of SONSAI can be extended to 3-tier architecture. The first tier would collect data and anonymise it (data layer). The second tier process anonymized data (process layer). The third tier is a client (presentation layer). It communicates with the two other tiers in order to generate rights for data collection, processing and de-anonymization through security communication protocols. Only personal data that are related to the user can be de-anonymized. De-anonymization must be only performed by the third tier (presentation layer). Furthermore, functionalities of SONSAI can be extended to perform identity prediction attacks based on inferred attribute values.

Finally, we notice that our proposed system SONSAI is close under some aspects to a recommendation system: an item suggestion can be viewed as an attribute value prediction. Hence, it can be used in wide range fields such as financial services, marketing, professional collaborations... However, unlike most recommendation systems our tool also provides explanations for the predicted values, namely an ordered list of attributes that have played a significant role in the computation.

Our findings can be exploited to design effective countermeasures in order to combat privacy leakage on social network. Two main techniques can be investigated. The first technique consists of deleting information and links in order to vacillate inferences due to lack of data. The second technique consists of adding information and links in order to alter the accuracy of inference due to data disagreement. The main challenges in both techniques are to balance social networks' utility, self privacy and neighbour privacy.

Our findings also turn attention toward further privacy issues concerning data anonymization. In fact, correlation between attributes can be exploited to de-anonymize the label of sensitive attributes. For instance, the adversary may disclose the graph of political affiliations by searching a graph with structure similar to the graph of musicians and different from the graph of relationship status.

A

Datasets

A.1 Dataset 1 (D1)

We have simultaneously targeted 100 Facebook profiles of users that live in North-East France. We have crawled at distance 2 from each profile to generate the first dataset. Data are collected in 2016. The Table A.1 gives more details about the dataset 1.

♯ Crawled profiles	15 012	♯ Discovered profiles	3 353 590
♯ Pages	1 022 847	♯ Types of pages	1 926
♯ Groups	135 381	♯ Relationship status	11

Table A.1: Details about the dataset 1.

74.21% of user profiles publish their gender and 59.69% of them are male. Only 15.95% of user profiles publish their relationship status according to 11 possible options. Table A.2 gives more details about the published attributes of crawled profiles. Figure A.1 depicts the frequencies of published attributes per user in dataset 2.

A.2 Dataset 2 (D2)

To generate a second dataset, we have simultaneously targeted 17 Facebook profiles of users that live in Île-de-France. We have crawled at distance 2 from each profile in 2017. The Table A.3 gives more details about the dataset 2. Table A.4 gives more details about the published attributes of crawled profiles.

Figure A.2 depicts the friendship graph. We have used OpenOrd algorithm [Martin et al., 2011] to draw the graph. The graph contains only crawled nodes. It contains 6 550 nodes and 101 581 undirected links between them. The average degree of node is 31.01. The diameter of the graph is 7. The radius of the graph is 4. The average path length is 3.56. 45.66% of nodes are male while 24.52% are female. 29.82% of nodes did not publish their gender. The sizes of nodes in Figure A.2 are proportional to the number of published visited places by users. The visited places are published as tags on photos. We notice that bigger nodes are central in Figure A.2. In other words, the number of published recent places is correlated to the centrality of users. Users that publish their geolocation information have more chances to effectively develop their friendship network and become hub. We notice that some big nodes do not have many friends but they are central as they are close to other central nodes. On the other hands several small nodes have high degrees. But, they are far from central nodes. We also notice that there are no

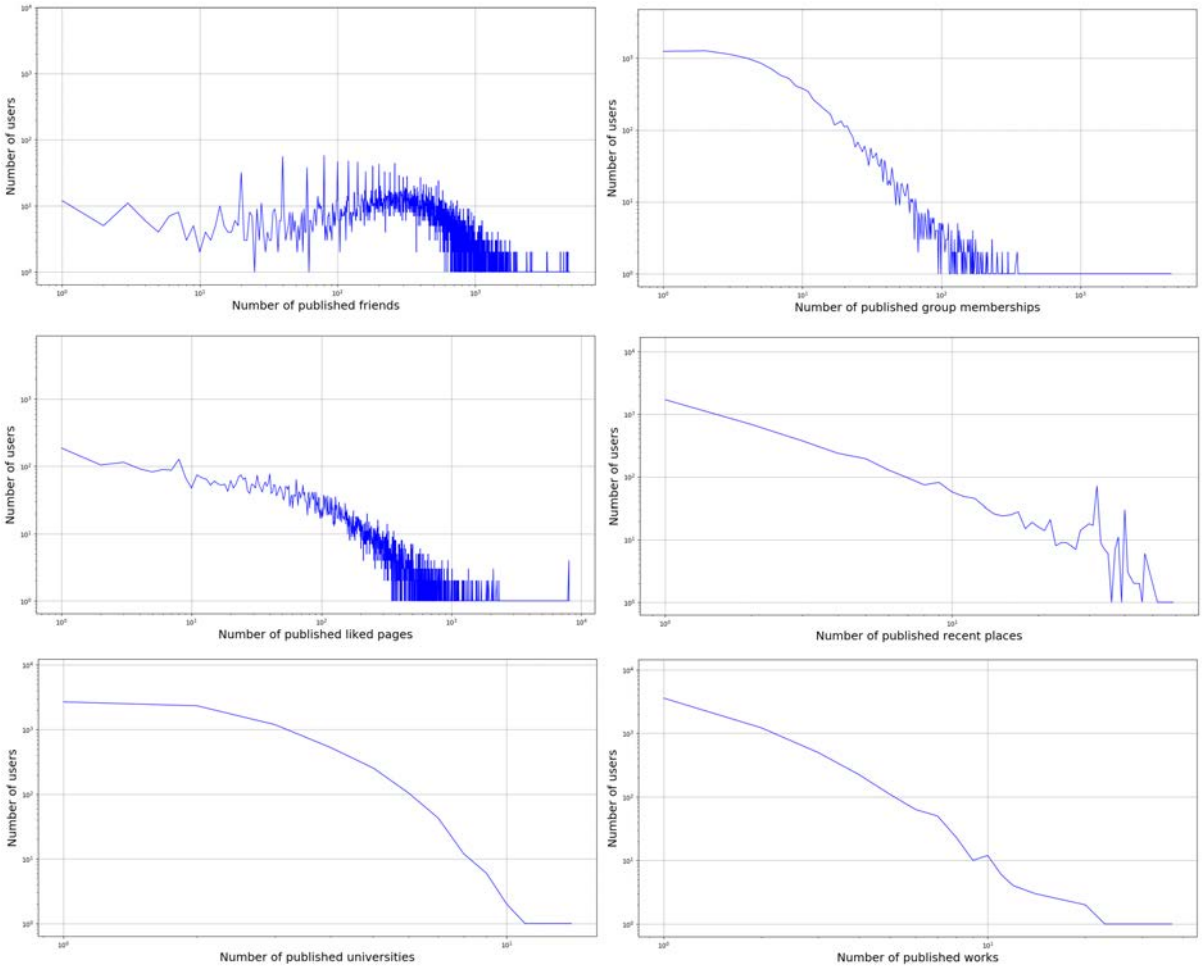


Figure A.1: Frequencies of published attributes per user in dataset 1.

correlation between centrality and gender. That is, both male and female can be influencer in a social network. Users who publish their geolocation information have more chances to become influencer even if they have relatively small number of friends.

Figure A.3 depicts the frequencies of published attributes per user in dataset 2.

Attributes	rate of published attribute (in%)	♯ values
friends	56.61	3 353 590
pages	61.87	1 022 847
groups	78.62	135 381
visited places	24.41	3 384
recent places	28.12	15 533
works	38.72	8 647
universities	47.66	9 383
home town	45.92	2 618
current city	53.15	2 259
living towns	7.5	1 112
telephone numbers	0.49	83
address street	0.77	106
address city	0.15	22
social media accounts	0.21	98
web pages	4.90	1 128
emails	0.59	97
birth date	5.99	755
birth celebration day	0.25	35
gender	74.21	2
interested in	11.68	2
spoken languages	15.62	493
religion orientation	1.94	193
religion opinion	0.25	37
politic orientation	1.33	188
politic opinion	0.27	40
civil state type	18.29	887
civil state partner	5	751
family links	20.14	8 989
biography	13.61	1 650
nicknames	11.05	1 846
quotes	10.96	1 255

Table A.2: Details about the published attributes of crawled profiles in dataset 1.

♯ Crawled profiles	6 550	♯ Discovered profiles	1 010 966
♯ Pages	298 604	♯ Types of pages	1 293
♯ Groups	29 062	♯ Relationship status	11

Table A.3: Details about the dataset 2.

Attributes	rate of published attribute (in%)	♯ values
friends	49.86	1 010 966
pages	58.56	305 158
groups	73.52	29 062
visited places	32.80	1 864
recent places	36.61	13 107
works	45.81	6 016
universities	38.67	3 457
home town	41.65	807
current city	41.65	807
living towns	8.09	607
birth date	5.95	336
birth celebration day	1.46	66
gender	70.18	2
interested in	4.38	2
spoken languages	13.43	249
religion orientation	1.24	72
religion opinion	0.15	10
politic orientation	1.01	63
politic opinion	0.12	8
civil state type	11.95	300
civil state partner	3.51	230
biography	12.27	804
nicknames	7.35	518
quotes	8.42	552

Table A.4: Details about the published attributes of crawled profiles in dataset 2.

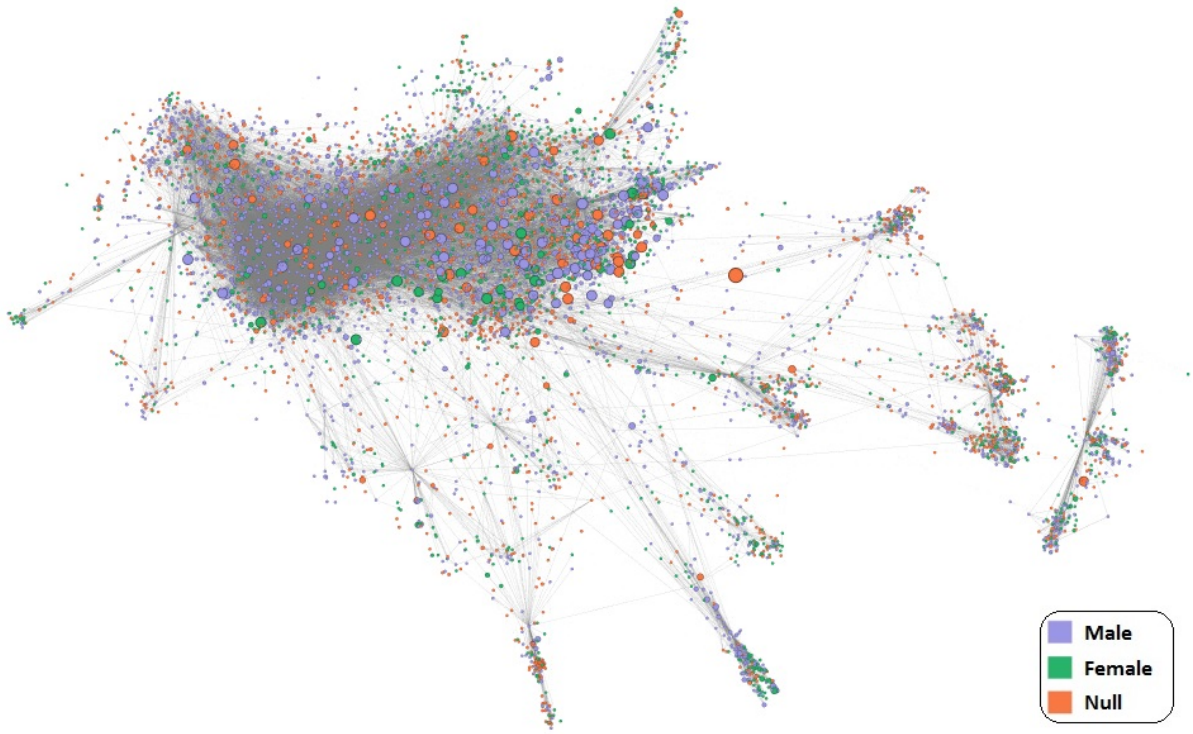


Figure A.2: Friendship graph of dataset 2.

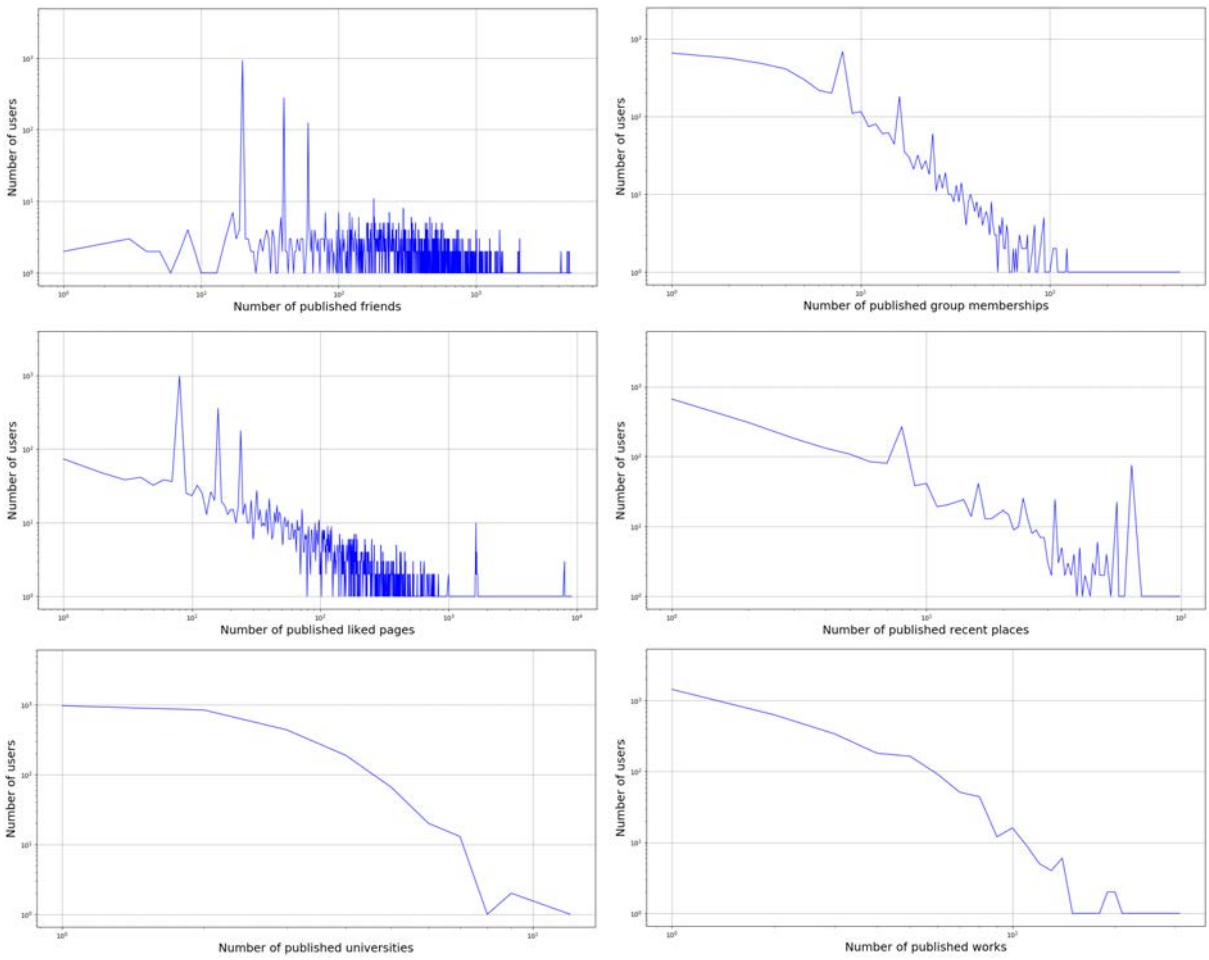


Figure A.3: Frequencies of published attributes per user in dataset 2.

B

Questionnaire

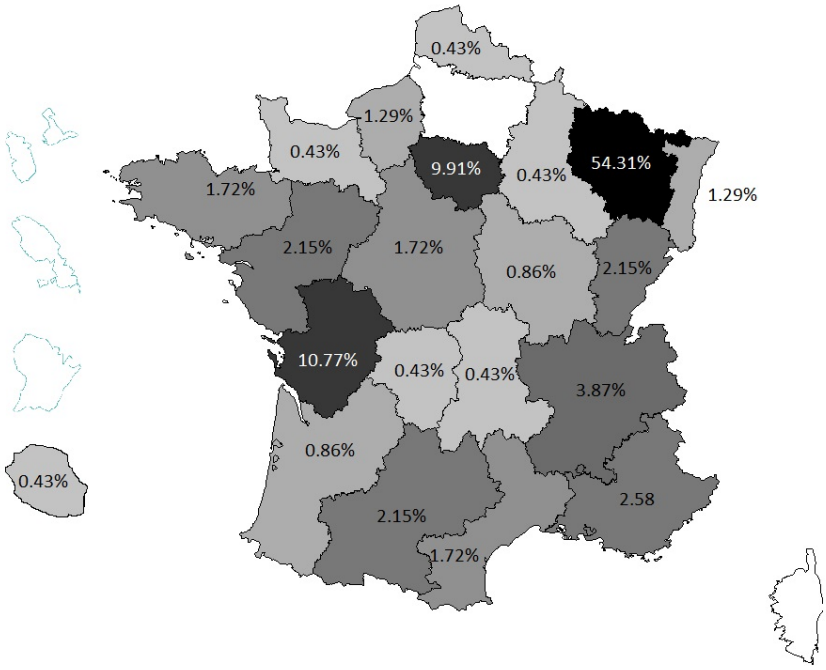
1- Mon âge est ... ans
232 réponses
L'âge moyen est 40 ans

2- J'ai des enfants qui utilisent les réseaux sociaux
139 répondants ont plus de 34 ans
- Oui 59.71 %
- Non 38.85 %
- Je ne sais pas 1.44 %

3- Je suis retraité(e)
20 répondants ont plus de 64 ans
- Oui 60 %
- Non 40 %

4- Je suis
232 réponses
- femme 63.36 %
- homme 36.64 %

5- Ma région est
232 réponses



6- Mon niveau d'étude est
232 réponses
- Secondaire 10.35 %
- Supérieur 89.65 %

7- Ma discipline d'étude est
232 réponses
- Sciences formelles et naturelles 56.96 %
- Sciences humaines et sociales 43.04 %

8- Le nombre de réseaux sociaux que j'utilise est
232 réponses
- 0 8.19 %
- 1 31.47 %
- 2 ou plus 60.34 %

9- Le nombre de sites/forums sur lesquels j'ai un profil est 232 réponses
- 0 7.33 %
- 1 16.81 %
- 2 ou plus 70.69 %
- Je ne sais pas 5.17 %

10- J'ai créé un profil sur un site/forum que je n'utilise plus et je ne sais pas si ce profil existe encore
176 répondants actifs sur plusieurs forums ou sites internet

- Oui 72.16 %
- Non 27.84 %

11- Ma fréquence d'utilisation des réseaux sociaux
213 répondants actifs sur les réseaux sociaux

Réseaux	jamais	un peu	beaucoup
Facebook	7.98 %	27.70 %	64.32 %
Twitter	59.62 %	24.88 %	15.49 %
Instagram	80.75 %	12.68 %	6.57 %
Google+	50.70 %	41.78 %	7.51 %
Linkedin	46.01 %	40.38 %	13.61 %
Viadeo	67.14 %	26.76 %	6.10 %
Youtube	22.06 %	52.11 %	25.82 %
Tumblr	90.14 %	8.45 %	1.41 %

12- Sur les forums et sites suivants
215 répondants actifs sur les forums ou sites internet

forums et sites internet	Je ne suis pas actif	Je cherche des réponses sans publier	Je fais des publications avec mon identifiant	Je fais des publications anonymes
Informatique, Technologie	33.02 %	53.02 %	13.49 %	0.47 %
Jeux, Musique, Film, Humour, Art, Livre	40.93 %	40.00 %	17.21 %	1.86 %
Santé, Achats, Cuisine, Maison, Astuce	33.95 %	55.81 %	9.30 %	0.93 %
Economie, Politique, Actualité, Infos	45.58 %	41.39 %	9.77 %	3.26 %
Sortie, Rencontre, Chat	75.81 %	14.88 %	9.30 %	0 %
Travail, Etude, Affaire	48.84 %	34.88 %	15.81 %	0.46 %
Activisme, Evénement	60.46 %	27.44 %	11.16 %	0.93 %
Voyage, Transport, Vacances, Assurance	41.39 %	44.19 %	13.49 %	0.93 %
Philosophie, Religion, Libre pensée	73.95 %	19.53 %	5.58 %	0.93 %

13- Mon forum/site informatique préféré est
30 répondants font des publications
- Stackoverflow 20.833 %
- Github 16.66 %
- Clubic 10.00 %
- Autre 13.33 %

14- Mon forum/site santé préféré est
32 répondants font des publications
- Doctissimo 45.45 %
- Autre 54.54 %

15- Mon forum/site politique, économie et infos préféré est
28 répondants font des publications
- Actu-politique 10.71 %
- Autre 89.29 %

16- Mon forum/site de rencontre/sortie préféré est
20 répondants font des publications
- On Va Sortir 26.31 %
- Blablacar 10.53 %
- Autre 63.16 %

17- Mon forum/site assurance, transport et voyage préféré est
31 répondants font des publications
- Tripadvisor 29.03 %
- Blablacar 22.58 %
- MAIF Social Club 6.45 %
- Autre 41.94 %

18- J'utilise des e-mails différents pour créer mes profils sur des réseaux/sites/forums différents.
219 répondants actifs sur plusieurs réseaux sociaux, sites internet ou forums
- Oui 47.94 %
- Non 52.05 %

19- J'ai les mêmes pseudos ou des pseudos qui se ressemblent sur des réseaux/sites/forums différents.
219 répondants actifs sur plusieurs réseaux sociaux, sites internet ou forums
- Oui 65.75%
- Non 34.25%

20- Mes pseudos/identifiants sont généralement
232 réponses (choix multiples)
- ☐ Mon nom ou prénom 35.34 %
- ☐ Un nom anonyme 60.78 %
- ☐ Une photo anonyme 14.22 %
- ☐ Ma photo 11.64 %
- ☐ Autre 12.93 %

21- J'ai plusieurs profils sur le même réseau social.
213 répondants actifs sur les réseaux sociaux
- ○ Oui 16.43 %
- ○ Non 83.57 %

22- Sur deux réseaux différents
140 répondants actifs sur plusieurs réseaux sociaux
- ○ Je n'ai aucun ami/lien commun 10.00 %
- ○ Je ne sais pas si j'ai un ami/lien commun 12.14 %
- ○ La majorité de mes amis/liens sont les mêmes 35.00 %
- ○ La majorité de mes amis/liens ne sont pas les mêmes 42.86 %

23- Mes amis/liens sont amis/ont des liens entre eux
213 répondants actifs sur les réseaux sociaux
- ○ Oui 67.14 %
- ○ Non 12.67 %
- ○ Je ne sais pas 20.19 %

24- Sur le même réseau
213 répondants actifs sur les réseaux sociaux (choix multiples)
- ☐ Je ne sépare pas entre ma vie professionnelle et personnelle en ajoutant mes collègues, camarade, membre de famille et voisin sur le même profil 56.34 %
- ☐ Je ne sais pas exactement qui figure dans ma liste d'ami 6.10 %
- ☐ Je n'ai pas supprimé un ami 7.51 %
- ☐ Je fais le tri dans mes amis 61.03 %
- ☐ J'ai des ex dans mes amis 15.96 %

25- Le nombre total de mes amis/liens sur tous mes réseaux est
140 répondants actifs sur plusieurs réseaux sociaux
- ○ Inférieur à 200 54.29 %
- ○ Entre 200 et 500 33.57 %
- ○ Entre 500 et 1 000 7.86 %
- ○ Supérieur à1 000 4.28 %

26- Le nombre de mes amis/liens sur mon réseau préféré est
213 répondants actifs sur les réseaux sociaux
- ○ Inférieur à 100 52.58 %
- ○ Entre 100 et 200 28.17 %
- ○ Entre 200 et 500 15.49 %
- ○ Supérieur à 500 3.76 %

27- J'ai plusieurs amis/lien en commun avec une personne que je ne connais pas. Alors,
213 répondants actifs sur les réseaux sociaux (choix multiples)

☐ Je ne sais pas si cette personne peut voir mes publications	10.33 %
☐ Je jette un coup d'œil sur le profil de cette personne	48.83 %
☐ J'accepte une demande d'ajout venant de sa part	8.45 %
☐ Je demande à mes amis, qui est cette personne	8.45 %
☐ Je lui envoie une demande d'ajout	0 %
☐ Je ne fais rien	53.52 %

28- Je publie des photos sans demander l'accord des personnes figurant sur ces photos.

213 répondants actifs sur les réseaux sociaux

○ Oui	15.96 %
○ Non	84.04 %

29- Mon ami(e)/lien a publié une superbe photo

213 répondants actifs sur les réseaux sociaux (choix multiples)

☐ Je partage la photo en sélectionnant les personnes qui peuvent la voir	21.13 %
☐ Je partage la photo sans faire attention à qui peut la voir	8.92 %
☐ Je "tague" une personne que j'ai identifiée sur la photo	5.16 %
☐ Je télécharge la photo	13.14 %
☐ Je n'interagis pas	18.78 %
☐ Je fais j'aime	65.26 %

30-Mes enfants peuvent visualiser toutes mes publications

76 répondants actifs sur les réseaux sociaux avec leurs enfants

○ Oui	71.05 %
○ Non	23.68 %
○ Je ne sais pas	5.26 %

31- Je peux visualiser toutes les publications de mes enfants

76 répondants actifs sur les réseaux sociaux avec leurs enfants

○ Oui	51.31 %
○ Non	35.53 %
○ Je ne sais pas	13.16 %

32- J'ai déjà utilisé ces technologies ou j'ai déjà vu ces émissions

232 réponses (choix multiples)

☐ Le Wifi dans un aéroport, centre commercial ou autres lieux publics	83.19 %
☐ Sacrée soirée avec Jean-Pierre Foucault	38.36 %
☐ GSM, Nokia 5110, 3210, 3310 ...	55.60 %
☐ Jour de foot avec Karim Bennani	5.60 %
☐ Self scanning au supermarché	32.33 %
☐ C'est mon choix (France 3)	41.81 %
☐ VHS (video home system)	73.27 %
☐ Wii, XboX ou PS	69.40 %
☐ Galaxy s5 ou s4	32.76 %
☐ DVD ou Blu-ray	84.48 %
☐ Iphone 6 ou 5	34.48 %
☐ Atari 2600	13.79 %
☐ disquette	84.48 %
☐ 7 sur 7	50.43 %
☐ Minitel	76.72 %
☐ Aucun	1.29 %

33- Le nombre d'heure que je passe sur les réseaux sociaux est

213 répondants actifs sur les réseaux sociaux

○ Moins de 7 heures par semaine	58.68 %
○ Entre 7 et 14 heures par semaine	29.58 %
○ Plus de 14 heures par semaine	11.74 %

34- Mon ami(e)/lien a publié/partagé un contenu

213 répondants actifs sur les réseaux sociaux (choix multiples)

☐ Je fais j'aime rien que pour faire plaisir à mon ami(e)/lien	8.45 %
☐ Je ne fais pas un commentaire sauf si le contenu me plait	52.58 %
☐ Je critique le contenu en commentaire si je ne l'aime pas	14.08 %
☐ Je ne partage pas le contenu sauf s'il me plait vraiment	39.44 %
☐ Je ne fais pas "j'aime" sauf si le contenu me plait	64.79 %
☐ Je peux le partager même s il ne me plait pas	0.47 %
☐ Je masque la publication si je ne l'aime pas	23.94 %
☐ Je n'interagis pas	19.72 %

35- Les sujets desquels je parle sur les réseaux sociaux sont

213 répondants actifs sur les réseaux sociaux (choix multiples)

☐ Actualité et infos	35.52 %
☐ Argent	0.93 %
☐ Courses	1.88 %
☐ Cuisine	25.82 %
☐ Emissions de télévision	15.02 %
☐ Etudes	21.60 %
☐ Famille	31. 45 %
☐ Mode	10.80 %
☐ Politique	25.82 %
☐ Religion	5.63 %
☐ Rencontres	5.16 %
☐ Santé	17.37 %
☐ Sorties	37.56 %
☐ Sport	21.13 %
☐ Technologie	27.70 %
☐ Travail	30.52 %
☐ Voyage	36.62 %
☐ Aucun	13.61 %
☐ Autre	17.37 %

36- Autres sujets que j'évite
69 réponses

Thèmes	**Sujets mentionnés et mots fréquents dans les réponses**	**Participant en %**
Politique	Politique, Guerre, Conflit, Complotisme	50.72
Religion	Religion	33.33
Vie personnelle et familiale	Famille, Vie privée, Numéros de téléphone et contact	21.74
Vie sentimentale et sexuelle	Sexe, Sentiment, Amour, Intime, pornographie infantile	17.39
Vie financière	Impôts, Argent	10.14
Actualité	Polémique, Grossièreté, Débats, Problèmes, Infos	10.14
Vie professionnelle	Travail	7.24
Santé	Santé, Sport, Nutrition	5.80
Art	Goûts, Couleurs, Poésie	2.90
Vacances et Voyages	Départ en vacances	1.45

37-Les publicités qui me sont les plus suggérées sur les réseaux sociaux sont

213 répondants actifs sur les réseaux sociaux (choix multiples)

☐ Auto, Moto, GPS	7.51 %
☐ Bijouteries	6.10 %
☐ Bricolage	7.51 %
☐ Des groupes que mes amis rejoignent	20.18 %
☐ Électroménager	6.57 %
☐ Immobilier	15.96 %
☐ Magazines	7.98 %
☐ Montre	3.28 %
☐ Multimédia	14.08 %
☐ Vacances et voyages	31.45 %
☐ Vêtements et chaussures	29.58 %
☐ J'ai un bloqueur de publicités	30.98 %
☐ Je ne sais pas	23.47 %
☐ Autre	4.22 %

Bibliography

[2803media.fr, 2010] 2803media.fr (2010). Répartition de l'âge des utilisateurs des réseaux sociaux en France.

[Abid et al., 2016a] Abid, Y., Imine, A., Napoli, A., Raïssi, C., Rigolot, M., and Rusinowitch, M. (2016a). Analyse d'activité et exposition de la vie privée sur les médias sociaux. In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, pages 545–546.

[Abid et al., 2016b] Abid, Y., Imine, A., Napoli, A., Raïssi, C., and Rusinowitch, M. (2016b). Online link disclosure strategies for social networks. In *Risks and Security of Internet and Systems - 11th International Conference, CRiSIS 2016, Roscoff, France, September 5-7, 2016, Revised Selected Papers*, pages 153–168.

[Abid et al., 2016c] Abid, Y., Imine, A., Napoli, A., Raïssi, C., and Rusinowitch, M. (2016c). Stratégies de divulgation de lien en ligne pour les réseaux sociaux. In *32eme Conférence sur la Gestion de Données - Principes, Technologies et Applications, Poitiers*.

[Abid et al., 2017] Abid, Y., Imine, A., Napoli, A., Raïssi, C., and Rusinowitch, M. (2017). Two-phase preference disclosure in attributed social networks. In *Database and Expert Systems Applications - 28th International Conference, DEXA 2017, Lyon, France, August 28-31, 2017, Proceedings, Part I*, pages 249–263.

[Abid et al., 2018] Abid, Y., Imine, A., and Rusinowitch, M. (2018). Sensitive attribute prediction for social networks users. In *2nd International workshop on Data Analytics solutions for Real-LIfe APplications March 26th, 2018 Vienna, Austria*.

[Adamic and Adar, 2003] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

[Adomavicius and Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6):734–749.

[Arrington, 2006] Arrington, M. (2006). Aol proudly releases massive amounts of private data.

[Backstrom et al., 2011] Backstrom, L., Dwork, C., and Kleinberg, J. M. (2011). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. *Commun. ACM*, 54(12):133–141.

[Bakhshandeh et al., 2011] Bakhshandeh, R., Samadi, M., Azimifar, Z., and Schaeffer, J. (2011). Degrees of separation in social networks. In *Proceedings of the Fourth Annual Symposium on Combinatorial Search, SOCS 2011, Castell de Cardona, Barcelona, Spain, July 15.16, 2011*.

[Barbaro and Zeller, 2006] Barbaro, M. and Zeller, T. (2006). A face is exposed for aol searcher no. 4417749.

[Barbieri et al., 2014] Barbieri, N., Bonchi, F., and Manco, G. (2014). Who to follow and why: link prediction with explanations. In *The 20th ACM SIGKDD, New York, USA - August 24 - 27*, pages 1266–1275.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[Bennett and Lanning, 2007] Bennett, J. and Lanning, S. (2007). The netflix prize. *Proceedings of KDD Cup and Workshop 2007*.

[Bhagat et al., 2016] Bhagat, S., Burke, M., Diuk, C., Filiz, I. O., and Edunov, S. (February 4, 2016). Three and a half degrees of separation. In *Facebook Research*.

[Bilge et al., 2009] Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. (2009). All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 551–560.

[buffer.com, 2018] buffer.com (2018). Daily posting limits.

[Chen et al., 2012] Chen, T., Kâafar, M. A., Friedman, A., and Boreli, R. (2012). Is more always merrier?: a deep dive into online social footprints. In *Proceedings of the 2012 ACM workshop on Workshop on Online Social Networks, WOSN 2012, Helsinki, Finland, August 17, 2012*, pages 67–72.

[Chester and Srivastava, 2011] Chester, S. and Srivastava, G. (2011). Social network privacy for attribute disclosure attacks. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, Kaohsiung, Taiwan, 25-27 July 2011*, pages 445–449.

[CNIL, 2010] CNIL (2010). Guide la sécurité des données personnelles.

[CNIL, 2016] CNIL (2016). Droit au déréférencement : la formation restreinte de la cnil prononce une sanction de 100.000 euros a l'encontre de google.

[Conover et al., 2011] Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*, pages 192–199.

[Conti et al., 2012] Conti, M., Poovendran, R., and Secchiero, M. (2012). Fakebook: Detecting fake profiles in on-line social networks. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*, pages 1071–1078.

[Cukierski et al., 2011] Cukierski, W., Hamner, B., and Yang, B. (2011). Graph-based features for supervised link prediction. In *The 2011 International Joint Conference on Neural Networks, IJCNN 2011, San Jose, California, USA, July 31 - August 5, 2011*, pages 1237–1244.

[CURIA, 2014] CURIA (2014). An internet search engine operator is responsible for the processing that it carries out of personal data which appear on web pages published by third parties.

[Dougnon et al., 2015] Dougnon, R. Y., Fournier-Viger, P., and Nkambou, R. (2015). Inferring user profiles in online social networks using a partial social graph. In *28th Canadian Conference on Artificial Intelligence, Halifax, Canada, June 2-5*, pages 84–99.

[Dow et al., 2013] Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*

[Edgar, 2004] Edgar, D. (2004). Data sanitization techniques. *A Net 2000 Ltd. White Paper*, 8:1–8.

[Elahi et al., 2016] Elahi, M., Ricci, F., and Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50.

[Elkabani and Khachfeh, 2015] Elkabani, I. and Khachfeh, R. A. A. (2015). Homophily-based link prediction in the facebook online social network: A rough sets approach. *J. Intelligent Systems*, 24(4):491–503.

[Estivill-Castro et al., 2014] Estivill-Castro, V., Hough, P., and Islam, M. Z. (2014). Empowering users of social networks to assess their privacy risks. In *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27-30, 2014*, pages 644–649.

[Facebook, 2015] Facebook (2015). Facebook terms.

[Feinleib, 2014] Feinleib, D. (2014). *Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data Revolution.* Apress.

[Gao et al., 2015] Gao, F., Musial, K., Cooper, C., and Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015:172879:1–172879:13.

[Garrett Brown and Borders, 2008] Garrett Brown, Travis Howe, M. I. A. P. and Borders, K. (2008). Social networks and context-aware spam. *In ACM conference on computer supported collaborative*, 10.

[Ge et al., 2014] Ge, J., Peng, J., and Chen, Z. (2014). Your privacy information are leaking when you surfing on the social networks: A survey of the degree of online self-disclosure (DOSD). In *IEEE 13th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2014, London, UK, August 18-20, 2014*, pages 329–336.

[Gerville-Réache and Couallier, 2011] Gerville-Réache, L. and Couallier, V. (2011). Échantillon représentatif (d'une population finie) définitions statistiques et propriétés. Échantillon représentatif, Sondage, Quotas, Probabilités d'inclusion.

[Gimenes et al., 2014] Gimenes, G. P., Gualdron, H., Raddo, T. R., and Jr., J. F. R. (2014). Supervised-learning link recommendation in the DBLP co-authoring network. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom 2014 Workshops, Budapest, Hungary, March 24-28, 2014*, pages 563–568.

[Gjoka et al., 2011] Gjoka, M., Butts, C. T., Kurant, M., and Markopoulou, A. (2011). Multigraph sampling of online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1893–1905.

[Golbeck and Rothstein, 2008] Golbeck, J. and Rothstein, M. (2008). Linking social networks on the web with FOAF: A semantic web case study. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1138–1143.

[Goldberg and Levy, 2014] Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

[Gong and Liu, 2018] Gong, N. Z. and Liu, B. (2018). Attribute inference attacks in online social networks. *ACM Trans. Priv. Secur.*, 21(1):3:1–3:30.

[Gong et al., 2014] Gong, N. Z., Talwalkar, A., Mackey, L. W., Huang, L., Shin, E. C. R., Stefanov, E., Shi, E., and Song, D. (2014). Joint link prediction and attribute inference using a social-attribute network. *ACM TIST*, 5(2):27:1–27:20.

[Google, 2018] Google (2018). Google privacy and terms.

[Griffith, 2008] Griffith, V. (2008). Books that make you dumb.

[Griffith, 2010] Griffith, V. (2010). Musics that make you dumb.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 855–864.

[Guns and Rousseau, 2014] Guns, R. and Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2):1461–1473.

[Guo and Chen, 2012] Guo, S. and Chen, K. (2012). Mining privacy settings to find optimal privacy-utility tradeoffs for social network services. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*, pages 656–665.

[Heatherly et al., 2013a] Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. (2013a). Preventing private information inference attacks on social networks. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1849–1862.

[Heatherly et al., 2013b] Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. M. (2013b). Preventing private information inference attacks on social networks. *IEEE Trans. Knowl. Data Eng.*, 25(8):1849–1862.

[Hu et al., 2017] Hu, P., Chan, K. C. C., and He, T. (2017). Deep graph clustering in social network. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1425–1426.

[Ingram, 2017] Ingram, D. (2017). Facebook said it has over 5 million advertisers.

[Irvine, 2017] Irvine, M. (2017). Facebook ad benchmarks for your industry [new data].

[Jain et al., 2013] Jain, P., Kumaraguru, P., and Joshi, A. (2013). @i seek 'fb.me': identifying users across multiple online social networks. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 1259–1268.

[Jin et al., 2013] Jin, L., Joshi, J. B. D., and Anwar, M. (2013). Mutual-friend based attacks in social network systems. *Computers & Security*, 37:15–30.

[Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 5:39–43.

[Kayes and Iamnitchi, 2015] Kayes, I. and Iamnitchi, A. (2015). A survey on privacy and security in online social networks. *CoRR*, abs/1504.03342.

[Kong et al., 2013] Kong, X., Zhang, J., and Yu, P. S. (2013). Inferring anchor links across multiple heterogeneous social networks. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 179–188.

[Kontaxis et al., 2011] Kontaxis, G., Polakis, I., Ioannidis, S., and Markatos, E. P. (2011). Detecting social network profile cloning. In *Ninth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2011, 21-25 March 2011, Seattle, WA, USA, Workshop Proceedings*, pages 295–300.

[Korolova et al., 2008] Korolova, A., Motwani, R., Nabar, S. U., and Xu, Y. (2008). Link privacy in social networks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 289–298.

[Lee et al., 2010] Lee, V. E., Ruan, N., Jin, R., and Aggarwal, C. C. (2010). A survey of algorithms for dense subgraph discovery. In *Managing and Mining Graph Data*, pages 303–336.

[Leskovec et al., 2007] Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *TWEB*, 1(1):5.

[Li et al., 2007] Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115.

[Li et al., 2015] Li, R., Yu, J. X., Qin, L., Mao, R., and Jin, T. (2015). On random walk based graph sampling. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 927–938.

[Li et al., 2017] Li, Y., Luo, P., Fan, Z., Chen, K., and Liu, J. (2017). A utility-based link prediction method in social networks. *European Journal of Operational Research*, 260(2):693–705.

[Liben-Nowell and Kleinberg, 2007] Liben-Nowell, D. and Kleinberg, J. M. (2007). The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031.

[Lichtenwalter et al., 2010] Lichtenwalter, R., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 243–252.

[Liu et al., 2013] Liu, J., Zhang, F., Song, X., Song, Y., Lin, C., and Hon, H. (2013). What's in a name?: an unsupervised approach to link users across communities. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 495–504.

[Lorenz, 2017] Lorenz, T. (2017). Going "instagram official" is the new way to declare your relationship status.

[Ma et al., 2017] Ma, J., Qiao, Y., Hu, G., Huang, Y., Wang, M., Sangaiah, A. K., Zhang, C., and Wang, Y. (2017). Balancing user profile and social network structure for anchor link inferring across multiple online social networks. *IEEE Access*, 5:12031–12040.

[Machanavajjhala et al., 2007] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1):3.

[Man et al., 2016] Man, T., Shen, H., Liu, S., Jin, X., and Cheng, X. (2016). Predict anchor links across social networks via an embedding approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1823–1829.

[Martin et al., 2011] Martin, S., Brown, W. M., Klavans, R., and Boyack, K. W. (2011). Openord: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011, San Francisco Airport, CA, USA, January 24-25, 2011*, page 786806.

[Mason, 2014] Mason, W. (2014). Politics and culture on facebook in the 2014 midterm elections.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Mislove et al., 2010] Mislove, A., Viswanath, B., Gummadi, P. K., and Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 251–260.

[Narayanan and Shmatikov, 2006] Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105.

[Newman, 2001] Newman, M. (2001). Clustering and preferential attachment in growing networks. In *Physical Review E*, pages 1237–1244.

[Nilizadeh et al., 2014] Nilizadeh, S., Kapadia, A., and Ahn, Y. (2014). Community-enhanced de-anonymization of online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 537–548.

[Noyes, 2018] Noyes, D. (2018). The top 20 valuable facebook statistics.

[Perozzi et al., 2014] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 701–710.

[Perozzi and Skiena, 2015] Perozzi, B. and Skiena, S. (2015). Exact age prediction in social networks. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 91–92.

[Price, 2016] Price, L. (2016). 20 tales of employees who were fired because of social media posts.

[Ricci et al., 2011] Ricci, F., Rokach, L., and Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. Springer.

[Rong, 2014] Rong, X. (2014). word2vec parameter learning explained. *CoRR*, abs/1411.2738.

[Ryu et al., 2013] Ryu, E., Rong, Y., Li, J., and Machanavajjhala, A. (2013). curso: protect yourself from curse of attribute inference: a social network privacy-analyzer. In *Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks, DBSocial 2013, New York, NY, USA, June, 23, 2013*, pages 13–18.

[Schakel and Wilson, 2015] Schakel, A. M. J. and Wilson, B. J. (2015). Measuring word significance using distributed representations of words. *CoRR*, abs/1508.02297.

[Shullich, 2012] Shullich, R. (2012). Risk assessment of social media of social-media utilization in an enterprise. In *SANS Institute,InfoSec Reading Room*, pages 1–46.

[Singh et al., 2016] Singh, A., Bansal, D., and Sofat, S. (2016). Preventing identity disclosure in social networks using intersected node. *IJISP*, 10(3):25–41.

[Smith and Linden, 2017] Smith, B. and Linden, G. (2017). Two decades of recommender systems at amazon.com. *IEEE Internet Computing*, 21(3):12–18.

[Snoek et al., 2012] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 2960–2968.

[Son and Kim, 2017] Son, J. and Kim, S. B. (2017). Content-based filtering for recommendation systems using multiattribute networks. *Expert Syst. Appl.*, 89:404–412.

[Statista, 2018] Statista (2018). Number of monthly active facebook users worldwide as of 4th quarter 2017 (in millions).

[Street, 2012] Street, J. (2012). *Music & Politics*. Polity Press.

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

[Tanimoto et al., 2015] Tanimoto, S., Ohata, K., Yoneda, S., Iwashita, M., Sato, H., Seki, Y., and Kanai, A. (2015). Risk assessment of social-media utilization in an enterprise. In *16th*

IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015, Takamatsu, Japan, June 1-3, 2015, pages 577–580.

[Tapellini, 2010] Tapellini, D. (2010). Consumer reports survey: Social network users post risky information.

[Tauzin, 2018] Tauzin, A. (2018). Chiffres des utilisateurs des réseaux sociaux en france et dans le monde en 2018.

[Telikani and Shahbahrami, 2018] Telikani, A. and Shahbahrami, A. (2018). Data sanitization in association rule mining: An analytical review. *Expert Syst. Appl.*, 96:406–426.

[Thompson, 2010] Thompson, D. (2010). Google's ceo: 'the laws are written by lobbyists'.

[Ugander et al., 2011] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *CoRR*, abs/1111.4503.

[van Schaik et al., 2018] van Schaik, P., Jansen, J., Onibokun, J. A., Camp, J., and Kusev, P. (2018). Security and privacy in online social networking: Risk perceptions and precautionary behaviour. *Computers in Human Behavior*, 78:283–297.

[Vidyalakshmi et al., 2016] Vidyalakshmi, B. S., Wong, R. K., and Chi, C. (2016). User attribute inference in directed social networks as a service. In *IEEE International Conference on Services Computing, SCC 2016, San Francisco, CA, USA, June 27 - July 2, 2016*, pages 9–16.

[Wang et al., 2015a] Wang, G.-N., Gao, H., Mensah, L. C. D. N. A., and Fu, Y. (2015a). Predicting positive and negative relationships in large social networks. *PloS One*, pages 1–14.

[Wang et al., 2015b] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015b). Link prediction in social networks: the state-of-the-art. *SCIENCE CHINA Information Sciences*, 58(1):1–38.

[Whittaker, 2010] Whittaker, Z. (2010). Facebook does not erase user-deleted content.

[Wits, 2015] Wits, B. (2015). 25 fascinating facebook facts and figures.

[Wu et al., 2014] Wu, S., Chien, H., Lin, K., and Yu, P. S. (2014). Learning the consistent behavior of common users for target node prediction across social networks. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 298–306.

[Yin et al., 2010] Yin, Z., Gupta, M., Weninger, T., and Han, J. (2010). A unified framework for link recommendation using random walks. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010, Odense, Denmark, August 9-11, 2010*, pages 152–159.

[Zakharya and Benslimane, 2018] Zakharya, S. and Benslimane, A. (2018). On location-privacy in opportunistic mobile networks, a survey. *Journal of Network and Computer Applications*, 103:157–170.

[Zhan et al., 2016] Zhan, Q., Zhang, J., Yu, P. S., Emery, S., and Xie, J. (2016). Discover tipping users for cross network influencing (invited paper). In *17th IEEE International Conference on Information Reuse and Integration, IRI 2016, Pittsburgh, PA, USA, July 28-30, 2016*, pages 67–76.

[Zhang et al., 2016] Zhang, J., Wang, S., Zhan, Q., and Yu, P. S. (2016). Intertwined viral marketing in social networks. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 239–246.

[Zhang and Zhang, 2012] Zhang, L. and Zhang, W. (2012). An information extraction attack against on-line social networks. In *2012 International Conference on Social Informatics (SocialInformatics), Washington, D.C., USA, December 14-16, 2012*, pages 49–55.

[Zheleva and Getoor, 2009] Zheleva, E. and Getoor, L. (2009). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th WWW 2009, Madrid, Spain*, pages 531–540.

[Zheleva et al., 2012] Zheleva, E., Terzi, E., and Getoor, L. (2012). *Privacy in Social Networks*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

Résumé

Cette thèse vise à comprendre le risque de fuite d'informations personnelles sur un réseau social. Nous étudions les violations potentielles de la vie privée, concevons des attaques, prouvons leur faisabilité et analysons leur précision. Cette approche nous aide à identifier l'origine des menaces et constitue un premier pas vers la conception de contre-mesures efficaces. Nous avons d'abord introduit une mesure de sensibilité des sujets à travers une enquête par questionnaire. Puis, nous avons conçu des attaques de divulgation (avec certitude) des liens d'amitié et des liens d'appartenance aux groupes sur "Facebook". Ces attaques permettent de découvrir le réseau local d'une cible en utilisant uniquement des requêtes légitimes. Nous avons également conçu une technique d'échantillonnage pour collecter rapidement des données utiles autour d'une cible. Les données collectées sont ensuite représentées par des graphes et utilisées pour effectuer des inférences d'attributs (avec incertitude). Pour augmenter la précision des attaques, nous avons conçu des algorithmes de nettoyage. Ces algorithmes quantifient la corrélation entre les sujets, sélectionnent les plus pertinents et permettent de gèrer la rareté (sparsity) des données. Enfin, nous avons utilisé un réseau de neurones pour classer les données et déduire les valeurs secrètes d'un attribut sensible d'une cible donnée avec une précision élevée mesurée par AUC sur des données réelles. Les algorithmes proposés dans ce travail sont inclus dans un système appelé SONSAI qui aide les utilisateurs finaux à contrôler la collecte d'informations sur leur vie privée.

Mots-clés: réseaux sociaux, sujets sensibles, divulgation de liens, inférence d'attributs, vie privée

Abstract

In this thesis we shed the light on the danger of privacy leakage on social network. We investigate privacy breaches, design attacks, show their feasibility and study their accuracies. This approach helps us to track the origin of threats and is a first step toward designing effective countermeasures. We have first introduced a subject sensitivity measure through a questionnaire survey. Then, we have designed on-line friendship and group membership link disclosure (with certainty) attacks on the largest social network "Facebook". These attacks successfully uncover the local network of a target using only legitimate queries. We have also designed sampling techniques to rapidly collect useful data around a target. The collected data are represented by social-attribute networks and used to perform attribute inference (with uncertainty) attacks. To increase the accuracy of attacks, we have designed cleansing algorithms. These algorithms quantify the correlation between subjects, select the most relevant ones and combat data sparsity. Finally, we have used a shallow neural network to classify the data and infer the secret values of a sensitive attribute of a given target with high accuracy measured by AUC on real datasets. The proposed algorithms in this work are included in a system called SONSAI that can help end users analysing their local network to take the hand over their privacy.

Keywords: social networks, sensitive subjects, link disclosure, attribute inference, privacy leak

Rapport de synthèse sur le projet :

Protection de la vie privée sur les réseaux sociaux

Membres du projet :

M. Younes ABID	Doctorant
M. Abdessamad Imine	Inria PESTO
M. Amedeo Napoli	Inria ORPAILLEUR
M. Chedy Raïssi	Inria ORPAILLEUR
M. Michaël Rusinowitch	Inria PESTO

June 18, 2018

1 Introduction

Afin de bénéficier du pouvoir social des réseaux sociaux, les utilisateurs ont tendance à être plus actifs et à partager plus de contenus dans une quête de renommée, de richesse, d'emploi ou simplement d'interactions sociales. Cependant, ils sont incapables d'évaluer les risques que des informations sensibles soient déduites sur eux-mêmes. Même les utilisateurs avertis qui se soucient de leur vie privée peuvent être exposés au risque de divulgation des informations sensibles personnelles telles que leur opinions politique et leur orientation sexuelle en se basant sur des informations inoffensives mais corrélées comme les couleurs, les musiques et les auteurs préférés.

Ce rapport est un rapport de synthèse sur les techniques développées au cours de notre projet afin d'aider les utilisateurs des réseaux sociaux à évaluer le risque qu'un tiers découvre leur réseau d'amis et déduise des valeurs de leurs attributs sensibles. La sensibilité d'un attribut est une notion subjective qui peut différer d'un utilisateur à l'autre. Certains utilisateurs peuvent considérer les opinions politiques ou l'origine ethnique comme sensibles alors que d'autres les considèrent comme inoffensives. Nous avons identifié à travers une enquête les attributs les plus sensibles pour un échantillon situé en France.

Les utilisateurs doivent manipuler avec soin leurs publications concernant les attributs corrélés à l'attribut sensible afin de protéger leurs vie privée. Par exemple, si la musique est très corrélée à la politique dans le réseau social, les utilisateurs doivent prêter attention aux musiques qu'ils publient afin de préserver le secret de leurs opinions politiques. La corrélation peut être complexe à comprendre et/ou inattendue pour les utilisateurs standards. C'est pourquoi nous proposons un outil pour les aider à gérer leurs publications: une fois qu'un utilisateur a défini l'attribut sensible qu'il souhaite cacher autant que possible, l'outil vérifie si d'autres attributs publiés donnent des indications sur cet attribut sensible et quels attributs sont les plus révélateurs. Si tel est le cas, l'utilisateur peut modifier ou supprimer ses préférences concernant cet attribut afin de diminuer ou annuler la corrélation. L'outil est conçu pour fonctionner raisonnablement avec les ressources limitées d'un ordinateur personnel, en collectant et traitant une partie relativement petite des données sociales.

Dans la Section 2, nous proposons une définition des sujets sensibles. Cette définition est basée sur le comportement des utilisateurs des réseaux sociaux qui ont participé à notre enquête par questionnaire en 2015. Dans la Section 3, nous détaillons l'architecture et les fonctionnalités de notre système d'audit de la vie privée sur les réseaux sociaux: SONSAI. SONSAI évalue la vulnérabilité des utilisateurs face aux attaques de prédiction des liens d'amitié et aux attaque d'inférence de valeur d'attributs sensibles. Il est constitué de deux outils : Un collecteur qui explore le réseau social et échantillonne des données pour les collecter. Un analyseur qui analyse les données collectées et affiche les résultats des attaques tout en indiquant les informations qui ont joué un rôle important dans l'analyse pour aider l'utilisateur à se protéger contre ces attaques.

2 Définition des sujets sensibles

Afin de lutter contre les fuites des informations sensibles, il est important de définir quelles informations personnelles sont sensibles. Certains chercheurs considèrent que toutes les informations non publiées par un utilisateur donné sont sensibles pour lui [Ryu et al., 2013, Vidyalakshmi et al., 2016]. Alors que d'autres choisissent quelques informations et les considèrent comme sensibles, comme l'affiliation politique [Heatherly et al., 2013, Conover et al., 2011], l'age [Perozzi and Skiena, 2015] et l'orientation sexuelle [Heatherly et al., 2013]. Il est également possible de s'appuyer sur la définition d'informations sensibles données par la loi relative à l'informatique, aux fichiers et aux libertés. Cependant, les réseaux sociaux évoluent plus vite que la loi. Par exemple, les données de santé n'ont pas été considérées sensibles par la loi française du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (version 1978). Il a été jugé sensible beaucoup plus tard.

Il est également possible de s'appuyer sur une définition de sujet sensible donnée par les médias sociaux eux-mêmes. Par exemple, selon Google, les données sensibles sont «*relatives à des faits médicaux confidentiels, à des origines raciales ou ethniques, à des croyances politiques ou religieuses ou à la sexualité*» [Google, 2018]. Mais comment pouvons-nous faire confiance aux réseaux sociaux dans la définition de ce qui est sensible ou non, sachant qu'ils tirent le meilleur parti de leurs profits en utilisant des informations personnelles pour une publicité ciblée?

Nous avons mené une enquête par questionnaire pour définir des sujets sensibles en fonction du comportement des internautes français. Cette méthode a l'avantage d'être rapide, précise et peut facilement être mise à jour. Les sujets sensibles sont définis par les utilisateurs eux-mêmes au lieu d'être imposés par les réseaux sociaux ou les lois. En outre, il est possible de prendre en compte les résultats statistiques de nouvelles enquêtes plus récentes afin de mettre à jour la définition sans répéter l'ensemble du processus. De plus, l'enquête nous a permis d'évaluer la vulnérabilité des internautes à certaines attaques de confidentialité.

Notre échantillon compte 232 utilisateurs de médias sociaux qui ont fourni des réponses valides et cohérentes. Ces utilisateurs sont situés dans 21 régions françaises et ont suivi plus de 18 disciplines différentes d'études.

Nous avons classé les sujets discutés sur les médias sociaux selon quatre critères : le taux de discussion sur les réseaux sociaux, le taux de discussion sur les forums et les sites web, le taux de publication anonyme et les sujets évités. Sur la base de ces critères, nous avons proposé une définition des sujets sensibles. Ensuite, nous avons calculé le coefficient de sensibilité des sujets étudiés dans notre enquête. Parmi les 25 sujets analysés dans l'enquête, nous avons défini 6 sujets sensibles comme représenté par le Tableau 1. Les sujets délicats sont évités **ou** [1] dont le taux de publication anonyme sur les forums et sites web est supérieur à la moyenne. Les sujets épineux sont délicats **et** dont le taux de discussion sur les forums et sites web **ou** les réseaux sociaux sont en dessous du seuil de la moyenne de toutes les discussions moins l'écart type sur ce média. Les sujets controversés sont les sujets évités **et** dont le taux de publication anonyme sur les forums et sites web est supérieur à la moyenne. Les sujets sensibles sont épineux **ou** controversés.

[1] Ou indique une disjonction inclusive.

Sujets	**délicat**	**épineux**	**controversé**	**sensible**
Argent	×	×	×	×
Religion, Libre-pensée	×	×	×	×
Achats	×	×	×	×
Rencontre	×	×		×
Santé	×		×	×
Politique	×		×	×
Famille	×			
Actualité	×			
Travail	×			
Voyages	×			
Sport	×			
Art	×			
Jeux	×			
Cuisine	×			
Mode				
Émission de télévision				
Études				
Technologie				
Musique	×	**Information insuffisante**		
Film	×			
Humour	×			
Livre	×			
Maison	×			
Astuce	×			
Sexe	×			

Table 1: Sujets sensibles

Notre méthode basée sur la sensibilité peut être enrichie par d'autres études statistiques pour analyser la sensibilité de plus de sujets (tels que la sexualité). Le coefficient de sensibilité varie de 0 à 4. Plus le coefficient de sensibilité est élevé, plus le sujet correspondant est sensible. Plus le taux de discussion d'un sujet donné est élevé, plus son coefficient de sensibilité est faible. Cependant, plus le taux de publication anonyme d'un sujet donné est élevé et plus il est évité, plus son coefficient de sensibilité est élevé. Le Tableau 2 trie les sujets sensibles par ordre décroissant du plus sensible au moins sensible sur les réseaux sociaux. Nous notons que cette enquête peut être répétée ou enrichi par les résultats d'autres enquêtes pour mettre à jour la liste des sujets sensibles ainsi que les coefficients de sensibilité.

Enfin, nous avons analysé le comportement des internautes afin d'identifier certaines vulnérabilités sur la vie privée. Environ 76% des parents participants confirment qu'ils ne contrôlent pas ce que leurs enfants peuvent découvrir sur les réseaux sociaux. De plus, environ 77,63% des internautes participants sont vulnérables aux attaques de croisement de profils entre différents médias sociaux car ils utilisent des e-mails ou des pseudos similaires. Par ailleurs, environ 65,25% des internautes participants sont exposés au risque de fuite d'informations sensibles sur le même réseau social. Enfin, notre étude montre que plus de 70% des utilisateurs de médias sociaux sont exposés au risque de fuite d'informations sensibles, principalement dû à une utilisation maladroite

Sujet x	Coefficient de sensibilité $C(x)$
Religion	2.25
Argent	2.18
Politique	2.08
Rencontre	2.00
Achats	1.85
Santé	1.63

Table 2: Ordre décroissant des sujets sensibles

des médias sociaux et à une méconnaissance des problèmes de la vie privée.

3 SONSAI: Outil de sensibilisation et d'aide à la protection de la vie privée

Dans ce travail, nous visons à fournir aux utilisateurs des réseaux sociaux un outil pour protéger leurs vies privées. À cette fin, nous étudions les attaques potentielles à la vie privée. Nous étudions leurs faisabilités et analysons leurs impacts. Cette approche nous permet d'identifier le périmètre des menaces pour ensuite concevoir des contre-mesures efficaces dans un travail future. Concrètement, nous concevons des attaques en ligne sur le plus grand réseau social du monde, «Facebook». Les attaques sont testées en ligne sur plusieurs profils de volontaires.

Afin de lutter efficacement contre les fuites de vie privée, il est très important de prendre en compte la combinaison d'attaques (prédiction de lien et prédiction d'attribut). En fait, ces attaques sont étroitement liées et lorsqu'elles sont combinées, elles présentent des menaces plus importantes pour la vie privée. Par exemple, un adversaire peut effectuer des attaques de prédiction de lien afin de dévoiler le réseau local de sa cible (les amis et groupes de la cible). Ensuite, il peut effectuer une attaque de prédiction d'attribut basée sur le réseau local découvert.

3.1 Attaque de prédiction de lien en ligne

Un réseau social peut être défini comme un site Web qui permet aux utilisateurs de créer des pages personnelles afin de partager des informations avec leurs amis et connaissances. Ces pages sont généralement appelées profils et contiennent des informations personnelles. Les profils sont connectés les uns aux autres par le biais de liens d'amitié qui peuvent être symétriques ou asymétriques, selon la politique du réseau. Pour imiter les interactions sociétales réelles (c'est-à-dire non cybernétiques), certains réseaux sociaux tels que Facebook, Linkedin et Viadeo permettent la création de groupes en plus de la création de profils. Les profils sont connectés à des groupes via des liens d'adhésion.

Afin d'effectuer des attaques de prédiction de liens en ligne, il est important de prendre en compte sa faisabilité. Par exemple, afin d'effectuer des attaques de prédiction de liens en ligne sur Facebook, un adversaire peut être tenté de vérifier les listes d'amis publics des utilisateurs de Facebook dans l'espoir

de trouver la cible dans ces listes. Cependant, Facebook compte environ 2 milliards d'utilisateurs actifs par mois et une approche aléatoire peut durer des années. De plus, Facebook est très dynamique. Par exemple, chaque seconde 5 nouveaux profils sont créés et 8 500 commentaires sont affichés [Noyes, 2018]. Ainsi, les attaques basées sur la reconnaissance de motif de réseau sont assez difficiles à réaliser. Le but de ces attaques est d'identifier une partie du réseau où la structure des connexions entre les utilisateurs est connue, par exemple une structure de connexions similaire au réseau de connaissance dans la vraie vie.

Nous avons conçu une stratégie d'attaque de divulgation de lien en ligne (avec certitude). La stratégie proposée est passive: l'adversaire n'a pas besoin d'interagir avec sa cible pour éviter d'attirer son attention. Notre attaque est réalisée sur un réseau social supportant plusieurs niveau de visibilité (secret, amis seulement, amis et leurs amis, tout le monde) et a été testée en ligne sur des profils de volontaires contrairement à l'inférence de liens hors ligne (avec incertitude) proposée dans [Wang et al., 2015, Gao et al., 2015] et l'attaque active par divulgation de lien dans [Jin et al., 2013] effectuée sur un réseau social supportant deux niveaux de visibilité (secret et amis directs). L'attaque décrite dans [Jin et al., 2013] divulgue l'amitié par le biais d'une requête d'ami commun mais elle n'a pas été testée en ligne. En outre, en explorant efficacement le réseau du groupe cible, notre stratégie d'attaque est capable d'effectuer des attaques de révélation de groupes, d'amitié et d'amis communs selon une stratégie qui minimise le nombre de requêtes. Seules les requêtes légitimes sont utilisées pour effectuer des attaques (c'est-à-dire des requêtes et des outils fournis par le réseau social ciblé). Notre étude exploite la relation intrinsèque entre les communautés (habituellement représentées en tant que groupes) et les amitiés entre individus. Pour développer une attaque efficace, nous avons analysé les distributions de groupes, les densités et les paramètres de visibilité d'un échantillon d'utilisateurs d'un réseau social (Facebook).

Le résultat de nos tests effectués sur 14 517 profils Facebook montre que la probabilité pour un utilisateur de Facebook de rejoindre au moins un groupe réunissant moins de 50 membres et de publier son adhésion à celui-ci est de 0,49. Ainsi, environ la moitié des profils Facebook analysés sont exposés au risque de divulgation de liens d'amitié par des groupes auxquels ils adhèrent et qui regroupent moins de 50 membres. Le nombre espéré de liens d'amitié publiés entre un membre donné et tous les autres membres du même groupe est $|g| \times PD(g)$. L'analyse de 1 100 groupes Facebook de tailles comprises entre 2 et 80 montre que le nombre espéré de liens divulgués entre les membres cibles et les groupes augmente de 2 lorsque la taille des groupes augmente de 10. Les groupes et les membres peuvent choisir de publier ou cacher la relation d'adhésion. Nous avons conçu une attaque pour dévoiler des groupes autour des cibles. Le réseau de groupe autour de la cible est ensuite utilisé pour divulguer les liens d'amitié et d'appartenance à des groupes qu'il cache. Les résultats des attaques effectuées sur les profils Facebook actifs montrent que 5 liens d'amitié différents sont divulgués en moyenne pour chaque requête.

3.2 Attaque de prédiction d'attribut

L'attaque de prédiction d'attribut comprend deux étapes: (i) la collecte des données et (ii) l'analyse des données. La collecte des données doit être rapide, sélective, passive et indétectable. En effet, les réseaux sociaux sont très dy-

namiques et contiennent de gros volumes de données. La collecte aléatoire peut créer des données inutiles. D'autre part, la collecte massive prend beaucoup de temps. Nous avons conçu un algorithme d'échantillonnage rapide et sélectif afin de guider le collecteur vers les données les plus importantes et d'accélérer le processus.

L'algorithme d'échantillonnage conçu prend en considération trois paramètres: (i) la proximité des nœuds échantillonnés au profil de l'utilisateur cible, (ii) la centralité des nœuds échantillonnés et (iii) le type des nœuds échantillonnés. La proximité d'un nœud échantillonné donné fait référence à la longueur du chemin le plus court entre celui-ci et le nœud cible. La centralité d'un nœud échantillonné donné fait référence au nombre de chemins entre celui-ci et le nœud cible. Nous distinguons trois principaux types de nœuds sur Facebook: les profils d'utilisateurs, les pages et les groupes. En outre, les pages ont différents types tels que des pages de musiques, de livres, de politiciens etc. Le collecteur tire parti des attaques de prédiction de lien détaillées dans la section 3.1. Ces attaques sont passives et n'utilisent que des requêtes permises par le réseau social pour collecter des données afin de rester indétectables par la cible et le réseau social.

Nous avons conçu un algorithme d'analyse de données rapide et précis qui peut analyser des informations incomplètes à cause des contraintes de la collecte. En effet, le collecteur échantillonne les données à collecter pour limiter le temps de collecte et diminuer le nombre de requête de collecte. Un grand trafic de collecte peut facilement être signaler par le réseau comme une attaque. L'ensemble du processus d'analyse ne dépasse pas quelques minutes afin de rapidement identifier le périmètre des menaces et pouvoir ensuite concevoir des contre-mesures dans un travail future. Le processus d'analyse comprend deux étapes: (i) quantifier l'importance de chaque attribut collecté et (ii) les utiliser pour déduire les valeurs secrètes de l'attribut sensible de la cible.

Pour quantifier l'importance des attributs collectés, nous avons conçu un algorithme pour évaluer et trier les attributs. Cet algorithme peut rapidement détecter et quantifier la corrélation entre les attributs. Par exemple, ils peut détecter la corrélation entre les préférences politiques et musicales. De plus, l'amitié est considérée comme un attribut parmi d'autres. Par conséquent, l'influence de l'amitié est également prise en compte lors de l'évaluation de l'importance des attributs. D'autre part, nous avons conçu un algorithme pour regrouper des valeurs similaires d'attributs afin d'accélérer le processus d'inférence et traiter des informations incomplètes. Nous distinguons deux types d'attributs. Le premier type comprend des attributs pouvant avoir plusieurs valeurs arbitraires. Les profils utilisateur peuvent être connectés à plusieurs valeurs du même attribut telles que des pages de livre. Le deuxième type d'attribut inclut des attributs qui ont des valeurs prédéfinies : les profils d'utilisateur peuvent être connectés à au plus une valeur du même attribut, comme le genre. Par conséquent, nous avons défini deux méthodes pour analyser les données selon le type de l'attribut sensible. Lorsque l'attribut sensible est un attribut qui admet des valeurs arbitraires, l'analyseur quantifie la corrélation entre les attributs en fonction de la similarité des valeurs préférées par les utilisateurs. Par exemple, si les utilisateurs qui aiment le même genre de musique aiment le même groupe de politiciens, alors la corrélation entre politicien et musique est élevée. Cependant, lorsque l'attribut sensible est un attribut qui possède un petit ensemble de valeurs prédéfinies, l'analyseur quantifie la corrélation entre

les attributs en fonction du pouvoir de discrimination des valeurs préférées des utilisateurs. Par exemple, si la plupart des utilisateurs qui aiment les pages de football sont des hommes, alors la discrimination entre le genre et le football est élevée.

Les attributs les plus corrélés à l'attribut sensible sont ensuite utilisés pour inférer les préférences de la cible concernant l'attribut sensible. L'analyse consiste à calculer les probabilités de préférences de la cible concernant les valeurs de l'attribut sensible en comparent ses autres préférences aux préférences des autres utilisateurs qui ont publiés leurs valeurs sensibles.

Nous avons mené plusieurs expériences sur de grands ensembles de données pour tester nos algorithmes. Pour générer des ensembles de données, nous avons collecté des profils Facebook. Pour chaque profil collecté, nous avons collecté la liste des pages qu' il aime, la liste de ses amis, son genre et son état civil. Pour générer le premier ensemble de données (D1), nous avons exploré le réseau d'amitié de 100 profils Facebook des utilisateurs du Nord-Est de la France. Les données sont collectées en 2016. D1 contient 1 926 types de pages différents, 1 022 847 pages différentes et 15 012 profils Facebook différents collectés. Le Tableau 3 détaille l'ensemble de données D1.

♯ Profils collectés	15 012	♯ Pages	1 022 847
♯ état civil	11	♯ Types de pages	1 926
♯Pages de politiciens	4 589	♯Profil qui publient leurs politiciens préférés	2 554
♯Profil qui publient leur genre	11 141	♯Profil qui publient leur état civil	2 395

Table 3: Détails sur l'ensemble de données D1.

Pour générer le deuxième ensemble de données (D2), nous avons exploré le réseau d'amitié de 17 profils Facebook d'utilisateurs résidant en Île-de-France. Les données sont collectées en 2017. D2 contient 1 296 types de pages différents, 298 604 pages différentes et 6 550 profils Facebook différents collectés.

Le Tableau 4 détaille l'ensemble de données D1.

♯ Profils collectés	6 550	♯ Pages	298 604
♯ état civil	11	♯ Types de pages	1 296
♯Profil qui publient leur genre	4 597	♯Profil qui publient leur état civil	991

Table 4: Détails sur l'ensemble de donnée D2.

Le Tableau 5 détaille les 23 attributs les plus corrélés à l'attribut sensible "pages des politiciens". L'expérience a été menée sur l'ensemble de données 1 (D1) qui contient 1 929 attributs. Facebook définit 11 état civils différents. Pour simplifier la présentation, nous définissons deux classes d'état civils comme suit:

$E1=$ {Celibataire, Divorcé, Sépare, Veuf, Compliqué}
$E2=$ {Partenariat domestique, Marié, Engage,
Relation, Union civile, Relation ouverte}

Le Tableau 6 donne des détails sur les 20 attributs les plus corrélés à l'attribut sensible «état civil». L'expérience a été menée sur l'ensemble de données 2 (D2) qui contient 1 299 attributs. Nous rappelons que le score de corrélation prend en compte le pourcentage de statut de classe d'état civil des utilisateurs ainsi que le taux d'utilisateurs qui publient à la fois leur état civil et leurs préférences. Nous remarquons que la plupart des attributs corrélés à la classe $E1$ sont axés sur les formations et les loisirs. D'autre part, la plupart des attributs corrélés à la classe $E2$ sont axés sur les entreprises. Le Tableau 7 donne des détails sur les 20 attributs les plus corrélés à l'attribut sensible " genre ". L'expérience a été menée sur l'ensemble de données (D1). Nous rappelons que le score de corrélation prend en compte le pourcentage de genre des utilisateurs ainsi que le taux d'utilisateurs qui publient à la fois leur genre et leurs préférences. Nous remarquons que la plupart des attributs corrélés aux hommes sont axés sur les sports, les jeux et les logiciels. D'autre part, la plupart des attributs corrélés aux femmes sont axés sur la santé, la maison et le luxe.

Attributs	♯ Profiles qui publient leurs valeurs	♯ Valeurs d'attribut
Utilisateurs	13 155	15 012
Communautés	8 118	137 338
Musiciens/Bande de musiciens	7 141	84 762
Figures publiques	6 455	28 289
Associations à but non lucratif	6 180	25 847
Artistes	5 970	31 681
Entreprises	5 939	20 750
Sites Internet	5 829	17 931
Émissions de télévision	5 778	11 876
Sites Web de divertissement	5 669	8 319
Médias/Nouvelles	5 871	14 042
Produits/Services	5 496	15 986
Sites Web d'actualités/médias	5 550	9 247
Organisations	5 328	14 738
Films	5 171	16 282
Entreprises locales	5 111	17 321
Vêtements	4 729	16 090
Gastronomies	4 763	8 422
Acteurs/Réalisateurs	4 785	10 425
Magazines	4 733	9 955
Athlètes	4 583	14 123
Pages d'application	4 396	4 244
Équipes sportives	4 309	10 433

Table 5: Les 23 attributs les plus corrélés à l'attribut sensible «pages des politiciens» dans l'ensemble de données D1.

Nous avons mené plusieurs expériences sur les deux ensembles de données D1 et D2. Pour chaque expérience, nous générons un nouvel ensemble de données auxiliaires à partir de l'ensemble de données original en sélectionnant

Attributs	Corrélations	Discrimination
Éducation	2.75	88.41 % E1
Collège communautaire	2.74	90.02 % E1
Agence de consultation	2.71	90.70 % E2
Site web de loisirs & sports	2.56	91.18 % E2
Site web de Maison & jardin	2.49	91.89 % E2
Automobile, Avion & Bateau	2.48	92.86 % E2
Localité	2.47	92.59 % E2
Siège social	2.46	91.18 % E2
Sites Web d'actualités/médias	2.42	90.32 % E2
Service financier	2.41	90.00 % E2
Société industrielle	2.40	89.29 % E2
Conseillère pédagogique	2.02	75.00 % E1
Cour de récréation	1.80	66.67 % E1
Téléphone/Tablette	1.70	63.64 % E1
Chirurgien plastique	1.60	60.00 % E1
Consulat & Ambassade	1.60	60.00 % E2
Équipe sportive scolaire	1.53	52.00 % E1
Bar de plongée	1.45	54.55 % E1
Vidéo	1.44	51.00 % E1
Playlist (musique)	1.41	53.04 % E1

Table 6: Les 20 attributs les plus corrélés à l'attribut sensible «État civil» dans l'ensemble de données D2

Attributs	Corrélations	Discrimination
Ligue sportive	4.22	75.97 % Mâle
Site de loisirs & sports	3.80	77.09 % Mâle
Jeux vidéo	3.66	80.16 % Mâle
Voitures	3.25	73.15 % Mâle
Équipes de sport amateur	3.03	72.86 % Mâle
Sport	2.80	73.07 % Mâle
Bijoux & Montres	2.72	56.26 % Femelle
Électronique	2.68	73.19 % Mâle
Logiciels	2.52	77.23 % Mâle
Produits de plein air et de sport	2.35	77.19 % Mâle
Magasin de vêtements féminins	2.35	77.28 % Femelle
Décoration de maison	2.29	54.60 % Femelle
Stade & Arena	2.28	74.45 % Mâle
Articles pour bébés / articles pour enfants	2.14	66.61 % Femelle
Cuisine	2.08	55.93 % Femelle
Sacs/Bagages	2.04	59.16 % Femelle
Beauté, Cosmétique et Soins Personnels	2.03	60.59 % Femelle
Magasin de cosmétiques	1.98	66.25 % Femelle
Salon de coiffure	1.92	61.44 % Femelle
Site web de Maison & jardin	1.72	55.18 % Femelle

Table 7: Les 20 attributs les plus corrélés à l'attribut sensible «genre» dans l'ensemble de données D2

tous les profils utilisateurs (cibles) qui publient leurs préférences concernant l'attribut sensible et au moins un autre attribut (les amis sont également considérés comme un attribut). Ensuite, nous supprimons toutes les préférences

concernant l'attribut sensible de 10 % des profils utilisateur sélectionnés (cibles). Les expériences ont consisté ensuite à inférer les préférences supprimées en analysant l'ensemble de données auxiliaires. L'analyseur utilise des techniques d'intelligence artificielle pour trier les valeurs de l'attribut sensible en fonction de leur probabilité d'être les vraies valeurs de la cible. Les valeurs suggérées de l'attribut sensible généré par l'analyseur sont ensuite comparées aux vraies valeurs de la cible pour calculer la précision de l'inférence.

Politiciens. Nous avons réalisé une expérience sur l'ensemble de données D1. L'analyseur a sélectionné les 23 attributs les plus corrélés aux attributs «pages des politiciens», comme détaillé dans le tableau 5. Seules les préférences concernant les attributs sélectionnés sont ensuite analysées pour déduire les préférences des cibles concernant l'attribut sensible «pages de politiciens». La précision de l'inférence est égale à 79 %. En d'autres termes, en moyenne, l'ensemble inféré de pages de politiciens par l'analyseur est 79% semblable à celui réellement aimé par la cible. Cependant, la précision d'inférence lorsque les 23 attributs corrélés sont sélectionnés de façon aléatoire est seulement de 41%. Nous avons effectué plus de tests en sélectionnant manuellement 3 attributs sémantiquement proches de la politique: organisations politiques, partis politiques et idéologies politiques. Bien que les attributs sélectionnés semblent prometteurs, la précision de l'inférence n'est que de 46%. Cela peut s'expliquer par le fait que de nombreux utilisateurs sont vigilants et ne publient pas leurs préférences concernant ces attributs. Par conséquent, l'algorithme d'apprentissage ne peut pas les exploiter correctement car il ne dispose pas d'informations suffisantes sur les préférences.

Sélection d'attributs corrélés	Précisions en %	♯ Cibles	♯ Préférences supprimées
Sélection automatique par algorithme (23 attributs)	79%	252	409
Sélection des attributs de musiques (2 attributs)	62%	233	379
Sélection des attributs politiques (3 attributs)	46%	123	297
Sélection aléatoire (23 attributs)	41%	204 (moyenne)	351 (moyenne)

Table 8: Résultats expérimentaux d'inférence des pages des politiciens.

Comme la musique et la politique sont empiriquement connues pour être corrélées [Street, 2012], nous vérifions la capacité de nos algorithmes à déduire les politiciens préférés des utilisateurs de Facebook en se basant uniquement sur les attributs musical. Nous avons sélectionné seulement deux attributs: genres musicaux et musiciens/bandes de musiciens. La précision de l'inférence dans cette expérience est égale à 62 % donc significative. Nous notons que l'attribut musiciens/bandes de musiciens a été automatiquement sélectionné par l'analyseur. Le Tableau 8 résume les résultats des expériences conduites.

État civil. Nous menons cette expérience sur les deux ensembles de données D1 et D2. Le Tableau 9 donne plus de détails sur les utilisateurs qui publient leur état civil dans les deux ensembles de données.

Classe d'état civil		E1	E2
♯ profils qui publient	D 1 (15 012 profils)	1 114	1 281
	D 2 (6 550 profils)	208	783

Table 9: État civil des utilisateurs dans l 'ensemble des données D1 et D2.

La précision de l'inférence de l'état civil est supérieure à 70% dans les deux ensembles de données D1 et D2 dès que la cible publie ses préférences concernant au moins 4 attributs parmi les 20 attributs les plus corrélés au état civil.

Genres. Nous avons mené cette expérience sur les deux ensembles de données D1 et D2. La Table 10 donne plus de détails sur les utilisateurs qui publient leur genre dans les deux ensembles de données.

Genres		Femelle	Mâle
♯ profils qui publient	D 1 (15 012 profils)	4 491	6 650
	D 2 (6 550 profils)	1 606	2 991

Table 10: Genres des utilisateurs dans l 'ensemble des données D1 et D2.

La précision de l'inférence du genre est supérieure à 83% dans les ensembles de données D1 et supérieure à 67% dans l'ensemble de données D2 dès que la cible publie ses préférences concernant au moins deux attributs parmi les 20 principaux attributs les plus corrélés au genre.

Le temps de traitement Le Tableau 11 affiche les temps de traitement. La vitesse d'horloge du processeur est de 2,3 GHz. La machine ne dispose que de 8 Go de mémoire RAM. Grâce aux algorithmes de détection de corrélation, seules les préférences concernant les attributs importants sont analysées. Par conséquent, les tâches d'inférence sont accélérées et la précision est améliorée en écartant les informations non pertinentes.

	Attributs sensibles	Données	Corrélation	Analyse
Temps	État civil	D1	7m3s	1m24s
	Genre	D2	4m30s	55s
	Politiciens	D1	13m2s	24m7s
		D2	9m34s	19m44s

Table 11: Le temps de traitement.

3.3 Mode d'emploi

Dans le système développé (SONSAI), que nous détaillons dans cette partie, un attribut sensible est initialement spécifié par l'utilisateur du système et choisi dans la liste des attributs découverts par le collecteur dans le réseau social de l'utilisateur. L'attribut sensible peut être des pages de politiciens ou tout autre attribut jugé sensible par l'utilisateur. SONSAI aide les utilisateurs à simuler une attaque consistant à déduire des informations sensibles les concernant afin

d'évaluer leur niveau de protection. De plus, cela aide les utilisateurs à comprendre d'où vient la menace en générant une liste triée en fonction de l'importance des attributs corrélés. SONSAI est composé de deux outils principaux: un collecteur et un analyseur qui collectent et analysent les données séparément. Toutes les conclusions faites par l'analyseur ne dépendent que des données collectées autour de l'utilisateur. Ainsi, les règles de confidentialité générées sont personnelles et spécifiques à chaque utilisateur. En fait, en évitant d'utiliser les règles générales de confidentialité, nous évitons de tirer des conclusions erronées sur la vie privée des utilisateurs de différentes communautés. Les conclusions sont plus adaptées à chaque contexte d'utilisateur spécifique.

Collecteur. L'interface graphique de l'outil Collecteur est représentée dans la figure1. Pour se connecter au réseau Facebook et collecter des données, l'utilisateur doit fournir son identifiant et son mot de passe. Le collecteur explorera ensuite le réseau Facebook via le compte d'utilisateur. L'utilisateur peut également choisir d'explorer ses données à l'aide d'un compte d'adversaire utilisé par défaut par l'application et d'effectuer une attaque de divulgation de lien ciblant son propre profil en cochant l'option «attaque de divulgation de lien». L'utilisateur doit ensuite définir la durée de la collection sur une valeur non nulle. Il peut continuez la collection qu'il a précédemment commencée en cliquant sur le bouton " Collecter ". Il peut mettre à jour les données déjà collectées en cliquant sur le bouton "Mettre à jour". Les données les plus anciennes sont mises à jour en premier. Enfin, il peut arrêter la collection en cliquant sur le bouton «Stop».

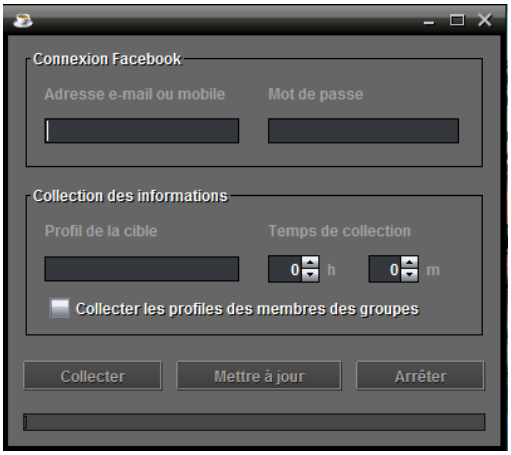


Figure 1: Collector GUI.

Analyseur. L'interface graphique de paramétrage de l'outil Analyseur est représentée dans la Figure 2. Afin de prendre en compte les données collectées récemment par le collecteur, l'utilisateur doit cliquer sur le bouton «Mettre à jour les données». Il peut ensuite sélectionner l'attribut sensible dans la liste de tous les attributs découverts autour de son profil. Il peut choisir la précision de l'analyse. La précision est en fait donnée comme le pourcentage d'attributs

corrélés sélectionnés pour l'analyse de tous les attributs disponibles dans le réseau de l'utilisateur; ces attributs sélectionnés sont ceux utilisés pour déduire les valeurs les plus proches de l'attribut sensible pour l'utilisateur. Lorsque l'utilisateur clique sur le bouton «Analyser», la page de résultats s'affiche. L'IHM des résultats de l'outil Analyseur est représentée dans les Figures 3 et 4.

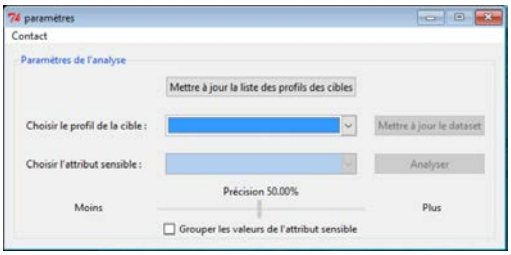


Figure 2: Analyser settings GUI.

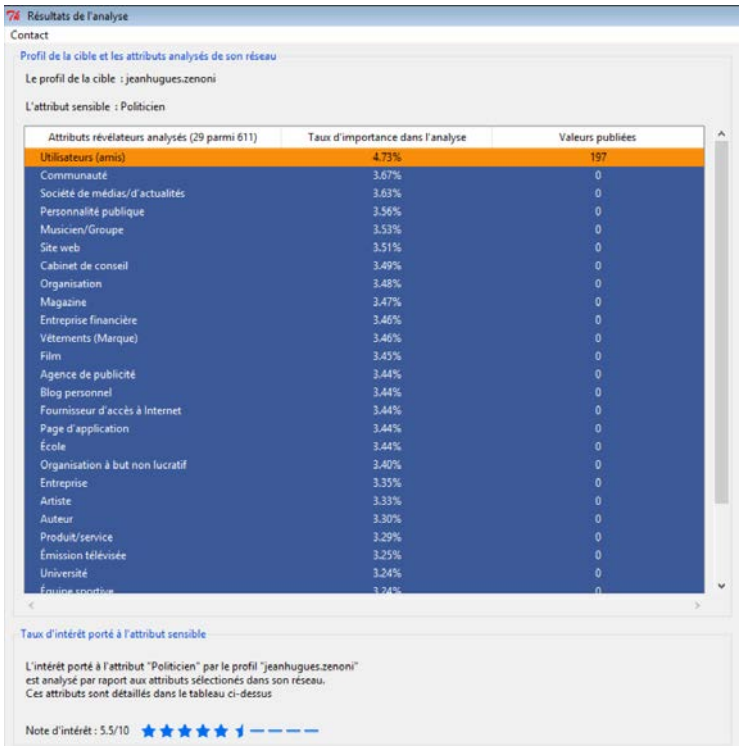


Figure 3: Analyser results GUI - left screen.

Le tableau en haut à gauche de l'écran (Figure 3) résume la liste des attributs sélectionnés et leur importance dans l'analyse.Les lignes correspondant aux attributs dont l'utilisateur cible a publié certaines valeurs sont de couleur orange.

Dans la partie inférieure gauche de l'écran (Figure 3) est l'évaluation de l'intérêt de l'utilisateur pour l'attribut sensible. Le score est calculé par rapport

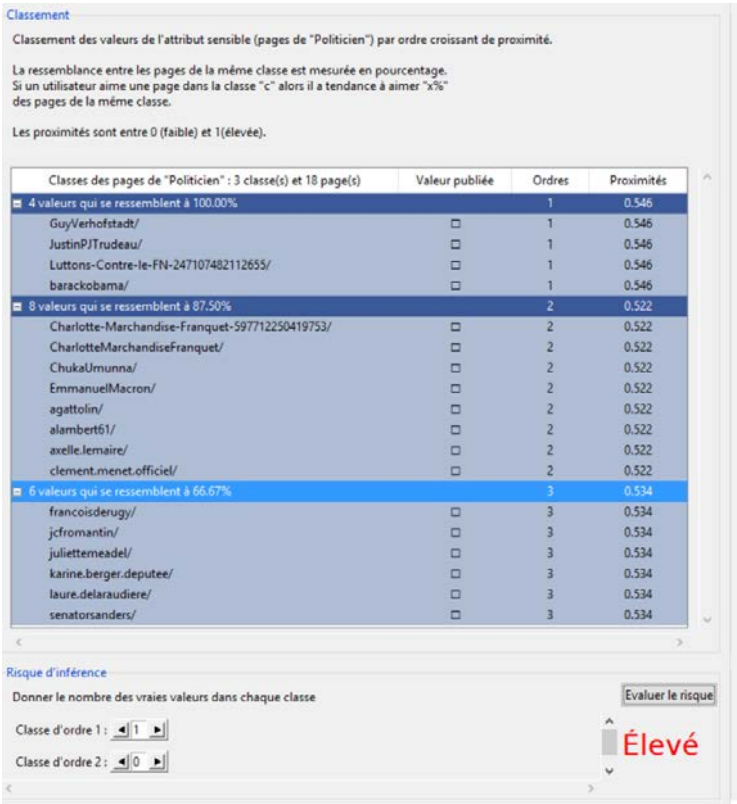


Figure 4: Analyser results GUI - right screen.

aux d'attributs sélectionnés

Le deuxième Tableau affiché dans la partie supérieure droite de l'écran (Figure 4) trie les valeurs de l'attribut sensible en fonction de leurs proximités avec l'utilisateur. Plus la proximité est élevée, plus la probabilité que la valeur soit la valeur réelle de l'utilisateur est élevée. Les valeurs sont regroupées dans des classes de tailles similaires d'une manière qui maximise la similitude entre les valeurs à l'intérieur d'une classe et la minimise entre les valeurs de différentes classes. La similitude est mesurée en pourcentage. Si un utilisateur aime une valeur dans une classe «c», il a tendance à aimer «x %» des valeurs de la même classe «c». L'utilisateur peut ouvrir les classes pour afficher les valeurs qu'elles contiennent. Il peut double-cliquer sur la valeur pour ouvrir sa page Facebook et modifier ses paramètre de confidentialité. Lorsque l'utilisateur publie ses valeurs, les cases correspondantes dans la colonne "Valeurs publiées" sont cochées. Dans la partie inférieure droite de l'écran (Figure 4), l'utilisateur peut évaluer le risque que ses valeurs sensibles soient déduites (par un tiers). Il doit d'abord spécifier le nombre de ses vraies valeurs dans chaque classe même s'il ne les a pas publiées sur Facebook. Puis il clique sur le bouton «Évaluer le risque». L'algorithme mesure la précision du classement. Nous définissons trois niveaux de risque d'inférence comme suit:

Si la précision est supérieure à 0,65, le risque d'inférence est élevé. En revanche, si elle est inférieur à 0,5, le risque d'inférence est faible. Si elle est

comprise entre 0,5 et 0,65, le risque d'inférence est considéré comme modéré.

4 Conclusion

Dans ce travail, nous avons analysé les risques de fuites des informations sensibles sur les réseaux sociaux. Premièrement, nous avons introduit une mesure de la sensibilité des sujets discutés sur les réseaux sociaux. Les sujets les plus sensibles selon les comportements des participants français à notre étude sont *Religion, Argent, Politique, Rencontres, Achats et Santé.* Afin de déduire des informations sensibles sur une cible donnée, nous dévoilons d'abord son réseau local (à 1 saut de la cible). À cette fin, nous avons conçu et testé des attaques de divulgation de liens en lignes avec certitude. Nous avons réalisé plusieurs attaques sur de vrais profils Facebook. Nous concluons que l'adversaire peut facilement et rapidement divulguer des liens cachés (amitié et appartenance à un groupe) avec certitude en profitant des API des réseaux sociaux.

Les données collectées autour de la cible sont ensuite traitées pour déduire les valeurs des attributs sensibles. Nos algorithmes montrent qu'il est possible de détecter et de quantifier la corrélation entre les attributs.

Nous avons remarqué que certains types d'attribut sont très proches et peuvent être regroupés. Par exemple, une classe de santé peut inclure des magasins d'équipement médical, des acupuncteurs et des services médicaux. À cette fin, dans les travaux futurs, nous prévoyons d'introduire des techniques de traitement du langage naturel pour aider à la classification des attributs.

Les algorithmes que nous proposons sont intégrés dans un système appelé SONSAI. SONSAI peut être installé sur n'importe quel ordinateur commercial avec Windows OS. Il contient très peu de paramètres à définir et est conçu pour être utilisé avec des connaissances informatiques de base. Il permet aux utilisateurs d'auditer leurs réseaux locaux et détecter rapidement les fuites potentielles d'information sensible avec une bonne précision.

Dans les travaux futurs, nous prévoyons d'auditer l'utilisation de SONSAI en recueillant les commentaires des utilisateurs sur l'amélioration possible et la fonctionnalité qu'ils souhaitent ajouter à l'outil.

Nos résultats peuvent être exploités pour concevoir des contre-mesures efficaces dans un travail futur afin de lutter contre les fuites d'informations sensibles sur les réseaux sociaux. Deux techniques principales peuvent être étudiées. La première technique consiste à supprimer des informations et des liens afin d'éviter les inférences dues au manque de données. La deuxième technique consiste à ajouter de l'information et des liens afin de modifier la précision de l'inférence en raison du désaccord sur les données. Les principaux défis dans les deux techniques sont d'équilibrer l'utilité des réseaux sociaux, la vie privée et la vie privée des voisins.

References

[Conover et al., 2011] Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third*

International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011, pages 192–199.

[Gao et al., 2015] Gao, F., Musial, K., Cooper, C., and Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015:172879:1–172879:13.

[Google, 2018] Google (2018). Google privacy and terms.

[Heatherly et al., 2013] Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. M. (2013). Preventing private information inference attacks on social networks. *IEEE Trans. Knowl. Data Eng.*, 25(8):1849–1862.

[Jin et al., 2013] Jin, L., Joshi, J. B. D., and Anwar, M. (2013). Mutual-friend based attacks in social network systems. *Computers & Security*, 37:15–30.

[Noyes, 2018] Noyes, D. (2018). The top 20 valuable facebook statistics.

[Perozzi and Skiena, 2015] Perozzi, B. and Skiena, S. (2015). Exact age prediction in social networks. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 91–92.

[Ryu et al., 2013] Ryu, E., Rong, Y., Li, J., and Machanavajjhala, A. (2013). curso: protect yourself from curse of attribute inference: a social network privacy-analyzer. In *Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks, DBSocial 2013, New York, NY, USA, June, 23, 2013*, pages 13–18.

[Street, 2012] Street, J. (2012). *Music & Politics*. Polity Press.

[Vidyalakshmi et al., 2016] Vidyalakshmi, B. S., Wong, R. K., and Chi, C. (2016). User attribute inference in directed social networks as a service. In *IEEE International Conference on Services Computing, SCC 2016, San Francisco, CA, USA, June 27 - July 2, 2016*, pages 9–16.

[Wang et al., 2015] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *SCIENCE CHINA Information Sciences*, 58(1):1–38.