Poster: The impact of public data during de-anonymization: a case study

Kevin De Boeck, Jenno Verdonck, Michiel Willocx, Jorn Lapon, Vincent Naessens
imec-DistriNet
KU Leuven
Ghent, Belgium
firstname.lastname@kuleuven.be

***Abstract*—Many companies, non-profit organizations and governmental bodies collect personal information during service interactions. However, releasing sensitive personal data may impose huge privacy risks. First, an increasing amount of sensitive personal information becomes publicly available online after user consent. Moreover, data breaches may result in huge data dumps that can contain personal records of millions of individuals. Hence, malicious entities are able to scrape, collect and combine personal data from multiple sources in order to compile detailed profiles of many individuals. This paper demonstrates the impact of publicly available data during de-anonymization by means of a concrete case study. Journalists are often reluctant or even prohibited to release the identity of suspects or victims in criminal cases. They do, however, often release initials and background (such as their age and residential location). Through a large scale study of over 132.000 news articles, this paper demonstrates that currently applied privacy measures are often insufficient and straightforward re-identification strategies can de-anonymize individuals.**

***Index Terms*—Privacy, re-identification, data breaches**

1. Introduction

Today, many companies and services we interact with in our daily lives collect data about individuals. Based on the acquired data, companies are able to offer personalized advertisements to customers and to build models to improve their service. Online service providers store preferences and personal interests for recommendations. While this is very convenient to end-users, huge data sources pose serious risks to our privacy. Besides data sources that are already publicly available online (e.g. social media and the whitepages), huge data breaches occur frequently. These breaches often result in huge amounts of personal data collected by services and companies being dumped on the internet.

Entities with doubtful or malicious intents can compile these dumps and scrape publicly available data in order to gain information on a large set of individuals. These data sources can in their turn be applied as background knowledge to perform large-scale re-identification attacks. For example, an anonymized dataset can be linked to the aforementioned publicly available datasets to support de-anonymization.

This work presents a practical use-case in which the negative impact of public data on the privacy of individuals is demonstrated. In news articles about crime-related cases, journalists are often reluctant or even prohibited to release the exact name and details of potential subjects or victims. They do however often opt to provide the reader with contextual information about these individuals. Examples are (partial) initials, the age and their residential location. If only the information in the newspaper was publicly available, only relatives would be able to de-anonymize individuals based on the info included in the article. However, due to the multiple publicly available data sources, interested readers can often quite easily re-identify a large set of individuals.

Contribution. This paper assesses the impact of public data for re-identification purposes. Firstly, this work categorizes publicly available data sources. Next, the negative impact of these data sources is assessed by means of a large-scale case study, namely the automatic re-identification anonymized individuals in news articles. Experimental results are presented and evaluated. The remainder of this work is structured as follows. Section 2 points to related work. Section 3 provides a general overview of public data sources. Next, our case study is presented in more detail. Preliminary results are discussed in section 5. This paper ends with conclusions and hints at future work.

2. Related work

Many research can be found on de-anonymization attacks. Pontes et al. [1] identify individuals based on the location information contained in reviews on the Foursquare platform. No data from external sources was required for this attack. Other papers, however, do rely on external data. Narayanan et al. [2] were able to link anonymized Netflix profiles to public IMDB profiles. Douriez et al. [3] demonstrated that drivers can be re-identified in an anonymized dataset by linking them to their unique cab ID. These works demonstrate that background knowledge and external data should be taken into consideration during anonymization. Several papers attempt to model the background knowledge of attackers [4]–[6] in order to improve the anonymization process. These papers approach background knowledge from a theoretical point of view, in which the knowledge of an attacker is fixed and pre-defined. In many real-life scenarios however, attackers have access to numerous data sources (e.g. data breaches and publicly available data). This work targets a real-life practical scenario in which the background knowledge of an adversary is not limited to a narrow, fixed amount of knowledge.

3. Public datasets

This work distinguishes three ways in which data can be acquired: data sharing, data scraping and data breaches.

Data sharing. Companies increasingly opt to share data with third parties. Data release can either occur in a bilateral agreement with a third party, or by publishing the data online.

Data scraping. This implies the collection of data beyond its intended use. Well-known examples of online data sources are the whitepages and social networks. Exploiting these online data sources by large-scale retrieval mechanisms allows malicious actors to optimize searches in the collected data, and additionally enables more complex and in-depth analysis of the data. Moreover, repeated crawling enables the construction of an historical view of an individual.

Data breaches. These occur when non-public data is accessed by an unauthorized person or party. Nowadays, this occurs frequently, as demonstrated by UpGuard [7] with their collection of the largest publicly-disclosed data breaches. Many well-known companies such as Facebook and Twitter have previously experienced multiple data breaches. The Facebook data breach [8] from 2019 is an example of a data breach impacting more than 500 million individuals, containing people's full names, phone numbers, locations, email addresses, dates of birth and more. The Facebook dataset is an example of a data breach that was dumped online for free. Other, more exclusive and harmful data is sold for premium prices.

4. Case study: News articles

The remainder of this work focuses on the data leakage by news articles. The goal is to investigate up to what extent individuals can be re-identified based on information given in news articles. News articles expose dates, locations and information about the persons involved such as their age, gender by means of pronouns, and surname. Journalists often replace full names with initials to protect an individual's privacy. The focus of this case-study is to reconstruct the full name of de-identified individuals based on the information in the news articles combined with publicly available data. This Section first presents the attacker model, after which the research methodology is outlined.

4.1. Attacker model

This study assumes a curious or malicious actor attempting to learn the full name of an individual based on the information included in a news article. The attacker attempts to single out one individual (or a small set of possible individuals) by searching through public datasets with key attributes contained in the article. Depending on the resources of the attacker, different capabilities are assumed. Weaker adversaries solely have access to freely available datasets (publicly shared data, data breach dumps and querying online platforms such as search engines), while stronger adversaries are able to scrape data and have access to a collection of datasets after a paywall.

Small scale attacks in which the identity of a few individuals in one of a limited numbers of articles must be exposed can be executed manually by an attacker. However, for the purpose of this paper, the whole process is automated and executed on a large set of news articles in order to correctly assess the broader privacy impact.

4.2. Research methodology

This research was executed in three steps. Firstly, a large set of news articles was crawled. Next, all relevant information was extracted from the crawled articles. Lastly, a re-identification attempt is made by mapping the extracted information to two public datasets.

Crawling news articles. A script was developed that automatically retrieves the title, abstract, content, author(s) and the date of all articles from an archive of news articles. Firstly, all articles present in the archive were crawled, starting from the year 1998 until January 2022. Next, articles were filtered based on two conditions: They had to contain keywords matching lawsuits (e.g. lawyer, trial, arrested) and keywords such as *theft*, *murder* or *extortion* in the abstract. Four prevalent news sites in Belgium were crawled for nearly 7 million articles. Based on the prerequisites, ±132.000 articles were selected. Afterwards, the data contained in the articles was extracted.

Data extraction. For the data extraction, a hybrid approach is employed, combining natural language processing and regular expressions. Two NLP models were used, namely Spacy [9] and Flert [10]. The Spacy model was used for sentence extraction while the Flert model allowed named entity recognition. This means identification of names, organizations, locations,... The data extraction process works as follows. Firstly, sentences were extracted from the article. Next, names and locations were identified. Afterwards, using regular expressions, ages were extracted. Finally, all found information per individual was merged. This process resulted in approximately 292.000 individuals. For our research, only news articles from 2017 and later are considered since these are more likely to match on our public data sources. This research focuses on reconstructing full names from (partial) initials or a partial name (only first or last name). Hence, full names are omitted from the test dataset. This leaves a total of 13.865 individuals in the test dataset, of which 2712 and 2904 include a location and age respectively. For 1128 individuals, both a location and age are known.

Re-identification. During the re-identification phase, the known attributes (initials, age, location) are matched to two different datasets, namely the aforementioned publicly available Facebook data breach dataset and the whitepages. The whitepages dataset is an extract from the Belgian whitepages[1]. It consists of 1.912.055 entries, and contains a name and a location (full address) for each record. The Facebook dataset (Belgian extract) consists of 3.183.529 entries, of which 1.546.421 entries have a location and 96.649 records have an age associated. Only a small subset of 71.512 records contain both a location and an age. Naturally, a more complete dataset (ideally the full population with all attributes), would inevitably result in more accurate re-identifications.

[1] Belgian PhoneBook: https://www.whitepages.be

5. Preliminary results

While this research is ongoing, it is possible to share some preliminary results. In the results presented below, a distinction is made between matches solely based on the names of individuals and matches based on the combination of name with another attribute (location, age). The more matching attributes, the higher the certainty of a correct match. Often, one individual in the test dataset matches to multiple individuals in the matching dataset (false positive results). A large amount of matches defines a low certainty for the attacker. Therefore, a second distinction is made between individuals with less than five matches and individuals with five or more matches. Note that the correct individual is not necessarily in the matching dataset.

The results from the experiment with the Belgian whitepages are presented in Table 1. For over half of the individuals in our test dataset, at least one match was found. Approximately a quarter of them had less than five matches. When matching a combination of name and location, only 787 individuals were matched, however half of them had less than five matches. When only considering the partial initials (complete first name and abbreviated last name or vice versa), 3666 individuals were matched (444 when matching name + location). The majority of these individuals (77%) had less than five matches.

TABLE 1. White Pages matches with our test dataset

Attribute	Matches	Initials	Partial name	Partial initials
Name	< 5	0	1198	778
	≥ 5	830	2567	2888
Name + Location	< 5	26	53	344
	≥ 5	245	19	100

The results from the experiment with the Facebook dataset are displayed in Table 2. 9600 individuals matched but only 19% had fewer than five matches. When also matching on location, 991 individuals were matched with half of them having fewer than five matches. Matching on age gives similar results. Matching on age and location often resulted in unique matches. However, due to the small number of records in the matching dataset that contain both an age and a location, the total amount of matched individuals for this category is low.

TABLE 2. Facebook matches with our test dataset

Attribute	Matches	Initials	Partial name	Partial initials
Name	< 5	0	1249	618
	≥ 5	830	3240	3663
Name + Location	< 5	36	52	433
	≥ 5	238	65	167
Name + Age	< 5	25	58	481
	≥ 5	234	60	96
Name + Age + Location	< 5	20	2	7
	≥ 5	2	0	0

6. Conclusions and future work

This poster paper presented preliminary research results of a case study on the impact of public data for re-identification purposes. While the current results support our hypothesis that re-identification by employing publicly available data is a realistic threat when publishing anonymized data, it should be noted that the accuracy of the attack strongly depends on the quality of the datasets available for the attacker. For example, because neither matching datasets used in the experiments are complete, our current results display a large number of false positives. However, even in its current form, our demonstrator eases the re-identification of individuals significantly by narrowing the search-space to a small set of individuals. Moreover, additional improvements to our current work can be made. Firstly, when other more complete matching datasets (more attributes and/or more individuals) are available, the quality of our matches will undoubtedly increase. Secondly, the experiments with different matching datasets are currently executed separately. Combining all public data sources into one combined matching dataset could also significantly improve our results.

After the offensive line of research presented in this paper, a next step could be to apply our research results to create guidelines for taking into account public data when leaking anonymized information on individuals. In this case specifically, a software tool could warn journalists when they unintentionally compromise the privacy of an individual.

References

[1] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida, "We know where you live: Privacy characterization of foursquare behavior," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 898–905.

[2] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 111–125.

[3] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?" in *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2016, pp. 140–148.

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

[5] T. Li, N. Li, and J. Zhang, "Modeling and integrating background knowledge in data anonymization," in *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 2009, pp. 6–17.

[6] W. Du, Z. Teng, and Z. Zhu, "Privacy-maxent: integrating background knowledge in privacy quantification," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 459–472.

[7] "The 63 biggest data breaches (updated for february 2022)," Feb 2022. [Online]. Available: https://www.upguard.com/blog/biggest-data-breaches

[8] "Facebook data on 533 million users reemerges online for free," Apr 2021. [Online]. Available: https://www.bloomberg.com/news/articles/2021-04-03/facebook-data-on-533-million-users-leaked-business-insider

[9] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.

[10] S. Schweter and A. Akbik, "Flert: Document-level features for named entity recognition," 2020.

The impact of public data during de-anonymization: a case study

Kevin De Boeck, Jenno Verdonck, Michiel Willocx, Jorn Lapon, Vincent Naessens
imec-DistriNet, KU Leuven

Abstract

Many companies, non-profit organizations and governmental bodies collect personal information during service interactions. However, releasing sensitive personal data may impose huge privacy risks. First, an increasing amount of sensitive personal information becomes publicly available online after user consent. Moreover, data breaches may result in huge data dumps that can contain personal records of millions of individuals. Hence, malicious entities are able to scrape, collect and combine personal data from multiple sources in order to compile detailed profiles of many individuals. This paper demonstrates the impact of publicly available data during de-anonymization by means of a concrete case study. Journalists are often reluctant or even prohibited to release the identity of suspects or victims in criminal cases. They do, however, often release initials and background (such as their age and residential location). Through a large scale study of over 132.000 news articles, this paper demonstrates that currently applied privacy measures are often insufficient and straightforward re-identification strategies can de-anonymize individuals.

Public datasets

Data scraping

yelp Google Maps

User Reviews

- Name
- Location
- Reviews
- Preferences

whitepages SEARCH. FIND. KNOW.

Phone Books

- Name
- Accurate Location
- Phone Nr

Data breaches

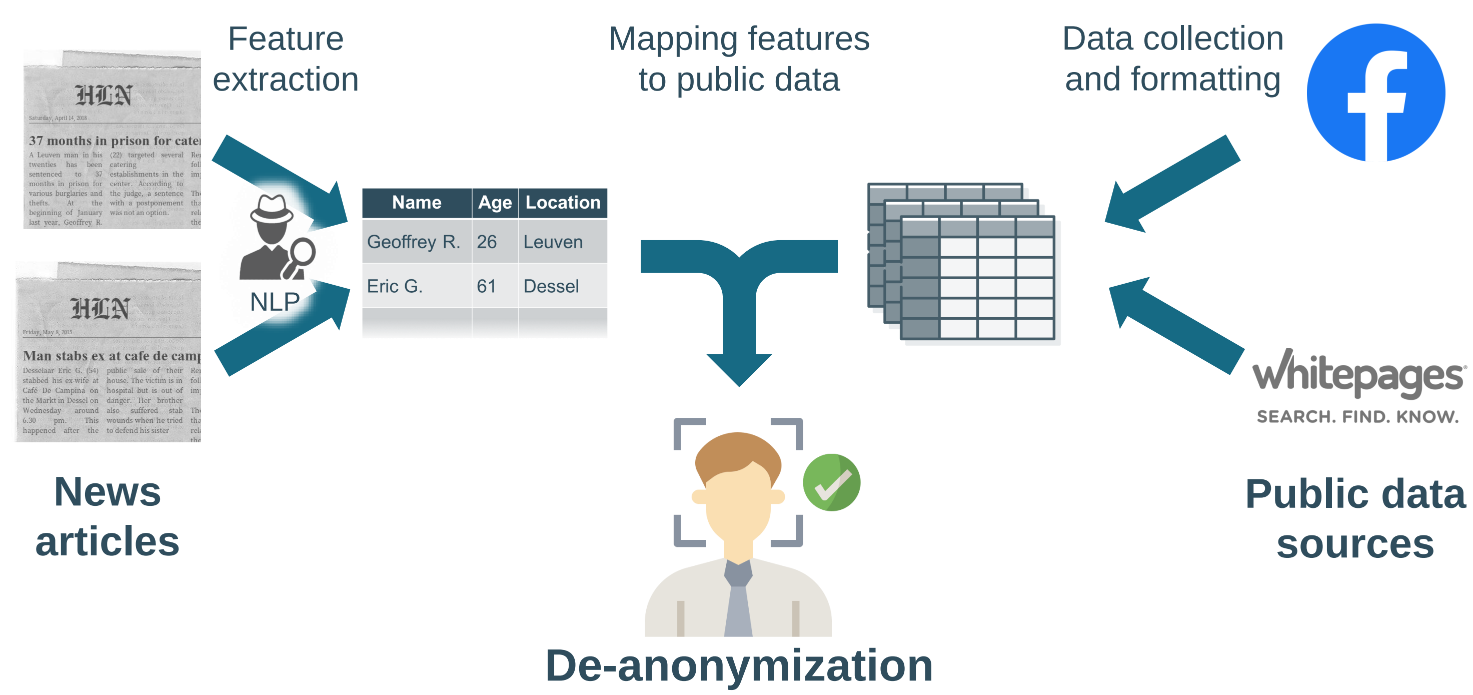
2021 Data Breach
~700M individuals

- Name
- Location
- Phone Nr
- Email
- Work information

2019 Data Breach
~533M individuals

- Name
- Date of Birth
- Location
- Phone Nr
- Email

Case study: News Articles


Feature extraction
Mapping features to public data
Data collection and formatting
NLP
News articles
Public data sources
whitepages SEARCH. FIND. KNOW.
De-anonymization

HLN
Saturday, April 14, 2018

37 months in prison for cater

A Leuven man in his twenties has been sentenced to 37 months in prison for various burglaries and thefts. At the beginning of January last year, Geoffrey R. (22) targeted several catering establishments in the center. According to the judge, a sentence with a postponement was not an option.

HLN
Friday, May 8, 2015

Man stabs ex at cafe de camp

Desselaar Eric G. (54) stabbed his ex-wife at Café De Campina on the Markt in Dessel on Wednesday around 6.30 pm. This happened after the public sale of their house. The victim is in hospital but is out of danger. Her brother also suffered stab wounds when he tried to defend his sister

Name	Age	Location
Geoffrey R.	26	Leuven
Eric G.	61	Dessel