TRANSACTIONS ON DATA PRIVACY 15 (2022) 41–85

How data brokers endanger privacy

Esma Aïmeur*, **Gilles Brassard***, **Muxue Guo***

*Université de Montréal, Département IRO, C.P. 6128, Succ. Centre-Ville, Montréal (QC), H3C 3J7, Canada.

E-mail: {aimeur,brassard}@iro.umontreal.ca, muxue.guo@umontreal.ca

Received 27 October 2021; received in revised form 31 March 2022; accepted 22 April 2022

Abstract. In the last decades, the information trading industry experienced important growth with the advent of Big Data. Information traders such as data brokers keep more and more detailed profiles of individuals, thus storing a variety of sensitive information. Those practices raise public concern, as leakage of personal data can harm data subjects and put the individual at risk of frauds or identity thefts. Even more worrisome is the fact that some data brokers, namely person search sites, deliver important amounts of personal data for free, by providing person registries on their public website. While a single data broker doing so may cause limited harm, when multiple person search sites provide different kinds of easily accessible personal information, those data can be linked to produce a complete profile containing a wide range of sensitive information. To provide the readers with an understanding of the current data broker industry, we conducted a survey on 75 data brokers and present the results in this paper. Furthermore, to show how easy it is to link data across different data brokers, we developed a system that automatically collects and links profiles from different person search sites. This system, named DROPLET, requires limited human intervention, but can produce linked profiles containing large amounts of personal sensitive information.

Keywords. Data privacy, data brokers, profile linking.

1 Introduction

Every day, as we proceed with our daily tasks or as we surf on the web, tremendous amounts of data are generated and some of them are collected, stored and analysed by data aggregators such as data brokers. It is estimated that on average, people generated at least 1.7 MB of data per second in 2020 [6]. Data brokers collect, analyse and combine data from multiple sources to produce a detailed profile of an individual and insights are derived from such profiles to produce additional value. Data brokers then sell those profiles to third-party data consumers who benefit from those data for marketing purposes, fraud detection during client authentication processes or personal use.

Such aggregation of personal data can lead to some potential threats such as surveillance, reputation harm or identity theft, especially when insufficient security measures lead to data breaches. A recent survey lead by the Aite Group reported that from 2019 to 2020, 47% of US consumers surveyed have experienced identity theft [15]. Also, the Identity Theft Resource Center® (ITRC) reported 1108 data breaches in 2020 in the United States, affecting 300.6 million individuals worldwide [15].

While some individuals may not worry about their data being sold to unknown corporations with the argument that "they have nothing to hide" [25], when personal informa-

tion is sold to individuals by person search sites, it also brings the potential of "relational control", defined as "the influence that a person can exert on another in their social or professional networks using covertly acquired private information" [21]. Indeed, people now not only need to worry about corporations leaking their data to ill-intended individuals, they also need to worry about how easy their personal information can be obtained by acquaintances they meet on a daily basis.

In response to growing public awareness about privacy issues, regulations were put in place to restrict data collection and dissemination or to limit usage of personal data for some purposes. For instance, in the United States, the Fair Credit Reporting Act (FCRA) regulates usage of credit information, the Health Insurance Portability and Accountability Act (HIPAA) regulates health information disclosure and the Children's Online Privacy Protection Rule (COPPA) limits collection of information about children. More recently, the General Data Protection Regulation, which came into effect in 2018, guarantees privacy rights of people residing in the European Economic Area (EEA) and is serving as model for other regulations worldwide. Some US states also grant further rights to their residents. For instance, the California Consumer Privacy Act (CCPA) protects California residents' privacy by providing them with diverse rights, including the right to know, the right to delete, the right to opt-out and the right to non-discrimination [26].

Despite the protections offered by current regulations around privacy measures, in practice, some people may still find their privacy threatened by different data sellers, partially because some types of data can be collected without the user's knowledge. While the right to erasure or the right to opt-out of data collection can be exerted, if a person is ignorant about the parties that possess his personal information, an opt-out request cannot be properly filled out and the parties who haven't received the opt-out request continue to gather personal information about the individual.

The difficulty to control one's own information is especially acute when the number of new data brokers increases every year and users have no ways of knowing if their personal data have been acquired by a new data broker. That is, when the number of leakage sources increases, the amount of information that can be retrieved about an individual also increases. Furthermore, while many may think that their personal data are well protected and are difficult to access by unrelated third parties, some person search sites often agree to provide many types of personal information for free as long as a potential data consumer has the "patience" to type in an individual's name in the website's search bar.

To provide a better view of the danger of such a service, this paper conducts an experimental research on the amount of information easily accessible on diverse person search sites. This is achieved by implementing a system for automatic person search that harnesses and combines data gathered from multiple person search sites. The data are combined by linking retrieved profiles to produce bigger profiles with the potential of containing more valuable information. The implemented system, named DROPLET (Data bROker Person Linkage idEnTification), performs a search based on a name provided as input and outputs resulting linked profiles for the requested person name. It is important to note that the system usually needs no further action from its user than providing the name used as search object and analysing the validity of the resulting linked profiles. Therefore, collection of personal information about individuals is made particularly easy.

After this introduction, Section 2 offers an overview of the data broker industry and Section 3 provides a brief review of related works. Section 4 explicates the methodology used by the DROPLET system to produce linked profiles. In Section 5, details about how the experiment was conducted are provided and results are analysed. Finally, Section 6 provides a discussion about the results and concludes this paper.

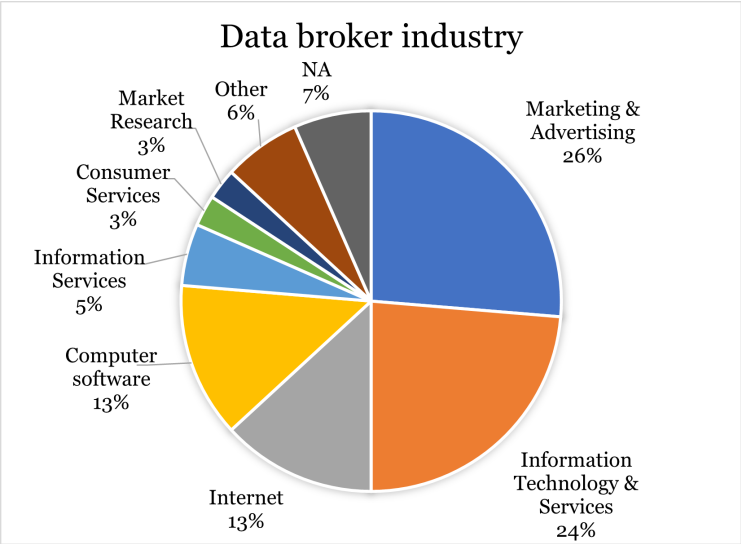


Figure 1: Data brokers industry

2 Data brokers

California law defines data brokers as "a business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a direct relationship" [1]. In other words, data brokers are businesses that profit from trading personal data without necessarily having the individual whose data are being traded benefit from the exchange. California law also requires data brokers to register themselves on the Attorney General's public website. As of the last time the authors consulted the website on August 25, 2021, 462 data brokers have completed their registration [27].

To provide an overview of the data broker industry, a survey was conducted on 75 data brokers registered on the Attorney General's website and multiple attributes were analysed. Of the selected sample, 10 data brokers were person search sites targeted at individuals, and the remaining ones were data brokers providing services to companies. The survey was conducted between May and August 2021 and used public data available for free at that moment on the data brokers' public pages. The following sections analyse data brokers according to the industry to which they belong, their year of foundation, the persuasion techniques they use to attract users and some characteristics of the privacy policies they provide.

2.1 Data broker industry

To determine the industry to which a data broker belongs, its LinkedIn page was consulted and the industry field of the page was retrieved. When no LinkedIn page was found, the information was considered Not Available (NA). Figure 1 presents the percentage of data brokers belonging to different industries, with the category "Other" containing industry types with only one entry in the surveyed sample.

From Figure 1, it can be seen that most data brokers orient their business around enhancing marketing efforts or supplying their clients with data to support their business

decisions. Some data brokers can also provide technologies or software to help their clients derive insights from their data. While most data brokers surveyed provide their services to businesses looking for "data solutions", some data brokers, described as "person search sites", are targeted at individuals looking for personal information about their acquaintances. Those data brokers sometimes prefer to describe their business as "consumer services", as their activities aim to aid individuals.

After analysing the current composition of the data broker market, it is also pertinent to evaluate the year of foundation of data brokers, and since when those information holders have been active in the market.

2.2 New data brokers by year interval

Following the advent of the era of Big Data, it would come as no surprise that the number of data brokers started to mushroom in the last decades. For this reason, we classified data brokers according to their year of creation and present resulting statistics in Figure 2. Instead of using standard decade classification, years were organised into groups of ten in such a way that 2020 could be included in the statistics. Data brokers from 2021 are not included in the figure since they belong to the ongoing year when the study was conducted.

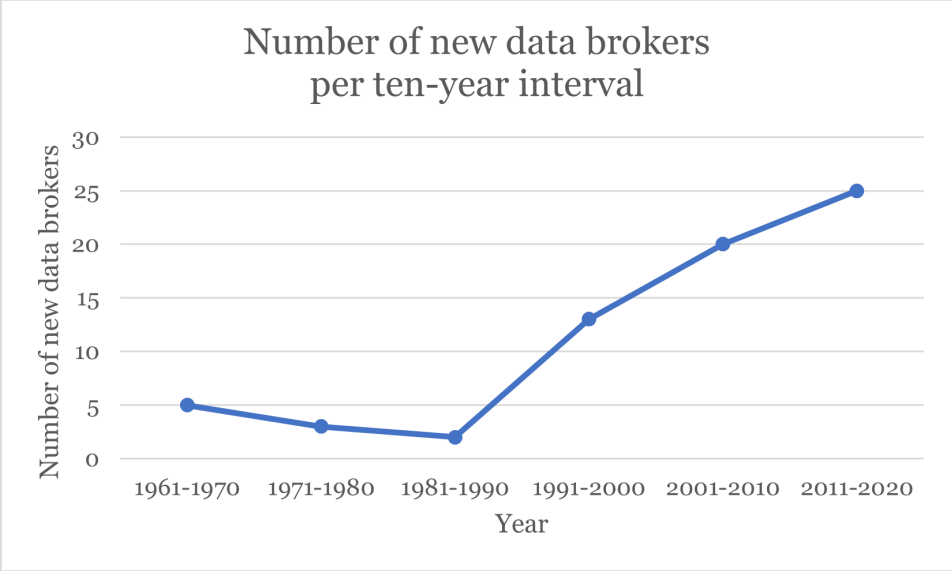


Figure 2: New data brokers by ten-year interval

We can notice an important increase in the number of data brokers after the 1990s, with one third of the data brokers from the surveyed sample beginning their business in the last ten years.

2.3 Persuasion techniques

Like any other businesses, data brokers need to attract clients to make a profit. For this reason, they use a number of persuasion techniques to attract their potential clients. We have analysed data brokers' strategies using the six persuasion principles proposed by Robert

Cialdini in his book "Influence: The Psychology of Persuasion" [9]. In his work, Cialdini presents six techniques marketers may use to manipulate their clients into buying their product, namely *reciprocity*, *consistency*, *social proof*, *liking*, *authority* and *scarcity*.

Reciprocity corresponds to the principle of "giving back what we take". When a person feels like he has received a service, he feels eager to return the favour. In the marketing context, it is often manifested in the form of "free samples" that are gifted to a potential buyer, who then feels obligated to buy the product in return.

Consistency, or commitment, results from the desire of appearing as a "consistent person". That is, when an individual starts performing some action or adopting some role, he feels more at ease when he continues what he has started, and is often reluctant to stop what he has undertaken. For this reason, many marketers will offer a "free trial" to potential users in hopes that the user will extend his membership even after the trial period. Other marketers will first try to hide the product's defects or boost its value to get an initial agreement from the client. Later, when the real (reduced) value of the product is presented, the client is reluctant to go back on his words and acknowledge that he has misjudged the value of the product. For this reason, the client often proceeds with the initial deal despite it not offering the promised return. The later technique is also called lowballing, since the customer is presented with a profitable deal, which is later revealed to be a bait.

Social proof is achieved by showing a hesitant client the product's success with other buyers. This technique is effective because humans tend to follow the crowd when they do not know the ideal choice to make. For this reason, marketers are happy to show in visible places how much another buyer liked the product. By doing so, they model the ideal consumer to be followed as example by other potential clients.

Liking corresponds to the art of evoking sympathy in the client's heart. To achieve this, the marketer has to give the consumer the illusion that they are on the same side. Many approaches can be used, as long as the marketer shows a friendly allure that creates trust and augmented compliance in the client.

Authority is effective because people will follow trusted sources to make better decisions. However, people rarely have the capacity to evaluate if a source is really as trustable as it seems. To appear as a credible business, marketers will try to pose as professionals or cite authoritative figures to convince a potential buyer that their advice is to be followed.

Finally, *scarcity* is, more often than not, an illusion forged by marketers to produce a sentiment of urgency in the mind of a buyer in order to encourage impulse spending. The marketer will insist that the product is limited in quantity or available for a restricted period of time, and that the client must buy the product "now". This technique is even more effective when a buying competitor is presented to the client, who must buy the product before the other buyer makes up his mind.

To provide a better picture of the differences in techniques used by data brokers when targeting companies versus when targeting individuals, person search sites were analysed separately. Figure 3 presents the number of data brokers using different persuasion techniques, with person search sites isolated from the other data brokers. It is important to understand that the figure only indicates the presence of one type of technique, and not to which extend the specific type of technique was exploited. In other words, even when different forms of a same technique are used by the same data broker, the technique is only counted once. For fairness, a data broker was marked as having used a persuasion technique only when an element on the data broker's home page clearly indicates the technique's usage, or when a user of a person search site inevitably encounters the use of a technique during a normal search process.

For the reciprocity technique, we considered the instances of a data broker offering a free

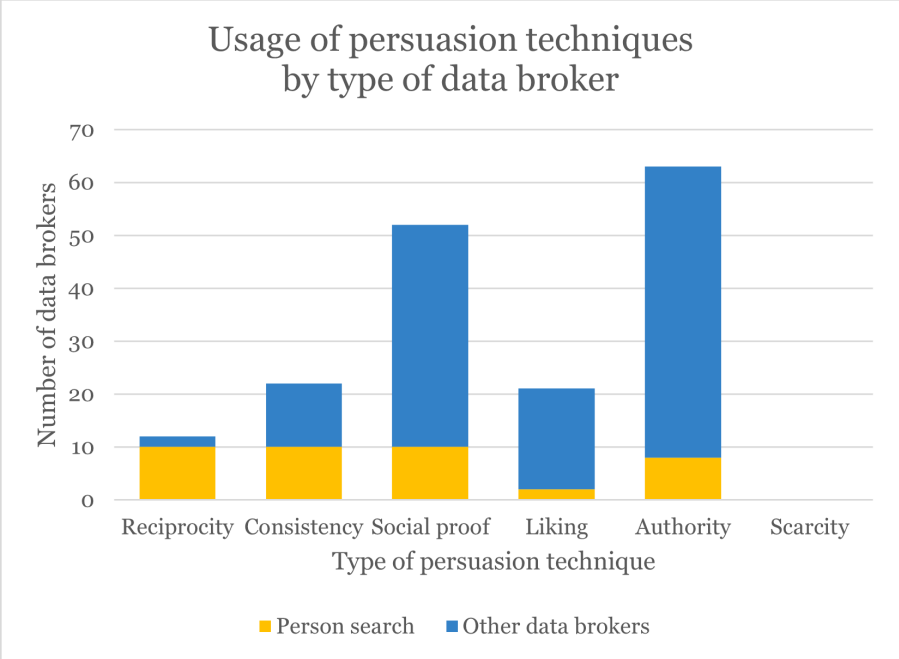


Figure 3: Usage of persuasion techniques

trial or partial person search services, without the user having to submit a request. For the consistency technique, we considered the occasions where free trials were offered, plus the occasions where a person search site employs a form of lowballing by reminding the user that "very sensitive information may be discovered" if they purchase an individual's full profile, even though such information is not present for the concerned profile. Indeed, this technique encourages customers to believe that they can discover an important amount of information for the price of one profile, while in reality the profile may not contain much information, thus making its content more expensive.

Social proof, liking and authority techniques are analysed and presented in more detail in Figure 4, Figure 5 and Figure 6. Data brokers are not showing signs of using the scarcity technique, likely due to the fact that data are not material and therefore are not naturally limited in quantity or in time. Also, to appear reliable, data brokers benefit more from guaranteeing to their clients that their data services remain available on demand.

From the data in Figure 3, it is interesting to note that person search sites are more eager to use reciprocity and consistency techniques than regular data brokers, likely due to the low cost of providing partial personal information compared to regular data brokers who often need to schedule a demonstration period with its client in order to demonstrate the power of their data solutions. Social proof and authority are frequently used by all data brokers, likely due to its ease of use. Indeed, compared to techniques requiring to offer a free trial to each potential client or to techniques demanding excessive effort in trying to be liked by clients with varying tastes, a "broadcast" of social acknowledgment is a far more efficient and convincing method.

Figure 4 lists the different applications of the social proof technique observed on data brokers' websites. We can notice that data brokers often publish other users' reviews as a

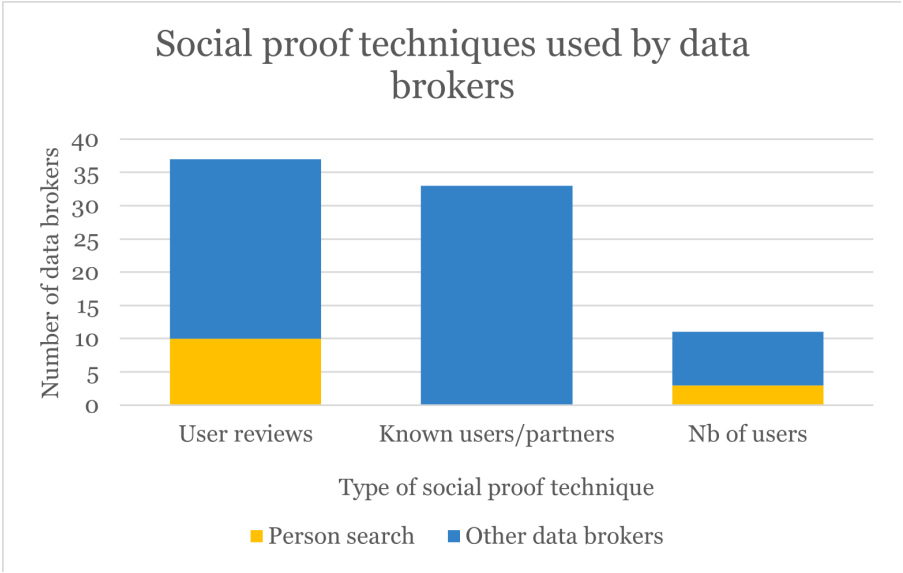


Figure 4: Usage of social proof techniques

mean of persuasion. When the review is accompanied by the reviewer's title (e.g. CEO of another business), or when a data broker simply lists some of his business partners, the data broker is also indicating to a potential client the social acknowledgment it has received by listing some renowned users/business partners. Occasionally, the data broker will indicate the number of users benefiting from his services. It is interesting to note that person search sites are not using names of renowned users/businesses as a mean of persuasion, possibly because their services are targeted at individuals and they want to appear closer to the public by removing the gap generated by too much professionalism. Furthermore, due to the promise of anonymity of the research process, person search sites ironically refuse to provide detailed information about their clients.

Liking techniques can take various forms, and Figure 5 merely lists the techniques that could be perceived consciously when navigating on the data broker's homepage. The most popular technique is the use of a chatbot to answer frequent questions. Occasionally, images of the data broker's staff will also be posted on the homepage. The two previously mentioned tactics both aim to make the data broker appear more "humane" and friendly in the eyes of the consumer. Data brokers can also increase a client's attachment to them by putting special effort on the navigation experience of his website with quality animations, by providing helpful resources that could help the client or by designing an attractive mascot to promote their services.

Numerous techniques used by data brokers to appear more professional and trustworthy are listed in Figure 6. Among those techniques, presenting names of renowned users/partners, listing the prizes and acknowledgments received, naming renowned journals that have published articles about them, and presenting their accreditation by the Better Business Bureau® (BBB) are techniques that rely on citing outside experts for authority. In a way, those techniques work similarly as social proof, with the difference that social proof rely more on quantity than quality. Even without outside acknowledgment, data brokers can appear more professional by listing statistics on the power of their

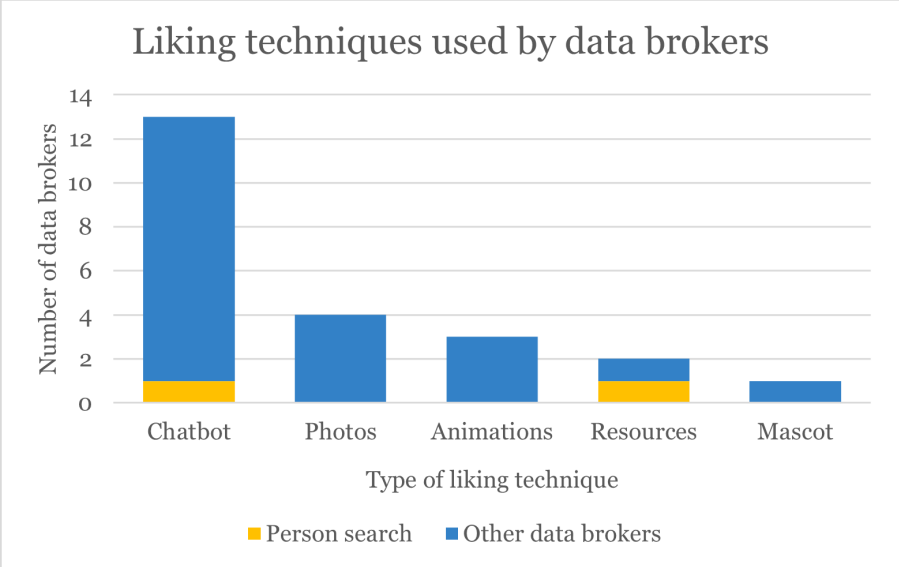


Figure 5: Usage of liking techniques

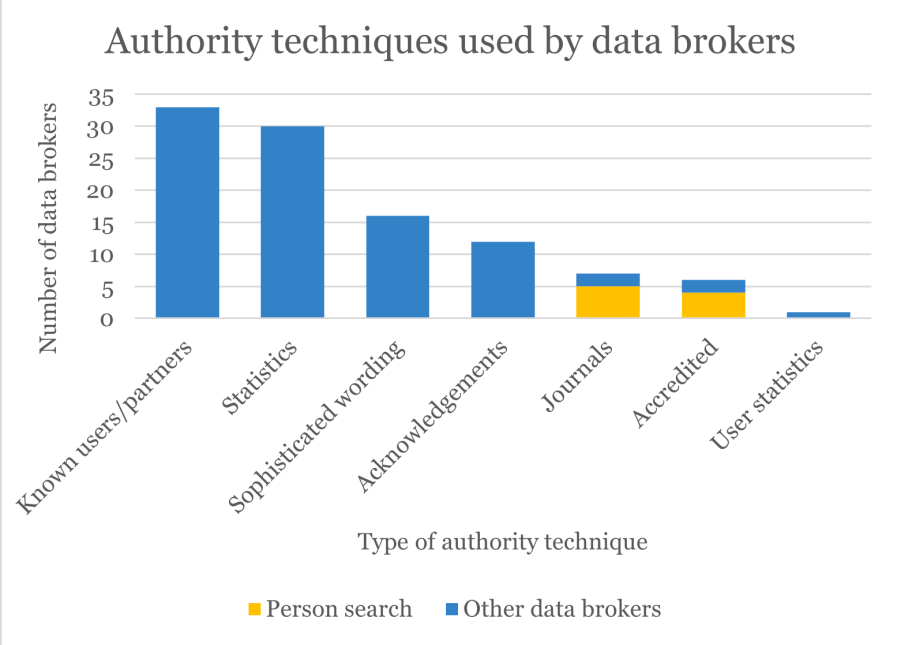


Figure 6: Usage of authority techniques

services, by voluntarily using more sophisticated wording (e.g. cutting-edge, data-driven, next-generation, ML, etc.), or by presenting some user statistics (e.g. a certain percentage of government officials are using their services). We can notice that person search sites are more inclined to cite journals or present their accreditation than using other authority tech-

niques, most likely because they do not want to pose as an authoritarian professional, but rather as an understanding friend eager to share others' information.

From the previous analysis, we can conclude that data brokers construct their online marketing techniques to fit the services they provide and their targeted audience. The form of their main product (intangible data) explains the absence of usage of scarcity techniques in all data brokers and the inclination to reciprocity and consistency techniques by person search sites. Furthermore, data brokers serving companies want to appear as trustworthy business partners by showing professionalism and presenting their popularity with other renowned businesses. On the other hand, person search sites want to attract individual users by posing as a friendly tool that is easy to use and does not rely on professional knowledge. Such an approach could help in not being marked as dishonest by partisans of anti-intellectualism, and therefore increase the number of potential clients. In both cases, the persuasion techniques chosen reflect the data broker's strategy to attract more clients to its business.

2.4 Privacy policies

From the viewpoint of a consumer, one important factor in deciding whether a company is trustworthy should be the presence of a privacy policy. Indeed, increased transparency helps users trust a business because they know what a business does with their data. Yet, only about one-in-five Americans respond that they often read the privacy policy before approving it, with about half of the remaining adults responding that they sometimes read the policy and the other half responding that they never read it [4].

Furthermore, among the adults who have ever read a privacy policy, only 22% attest to having read the entire policy till the end [4]. With privacy policies increasing in size to comply with new regulations such as the GDPR [18], we can worry that the proportion of adults reading the policy may drop even lower. In the context of data broker websites, since reading the privacy policy is not a requirement to use the services (the website assumes that the privacy policy was read), the real proportion of users who have truly read the privacy policy may be lower than the research numbers.

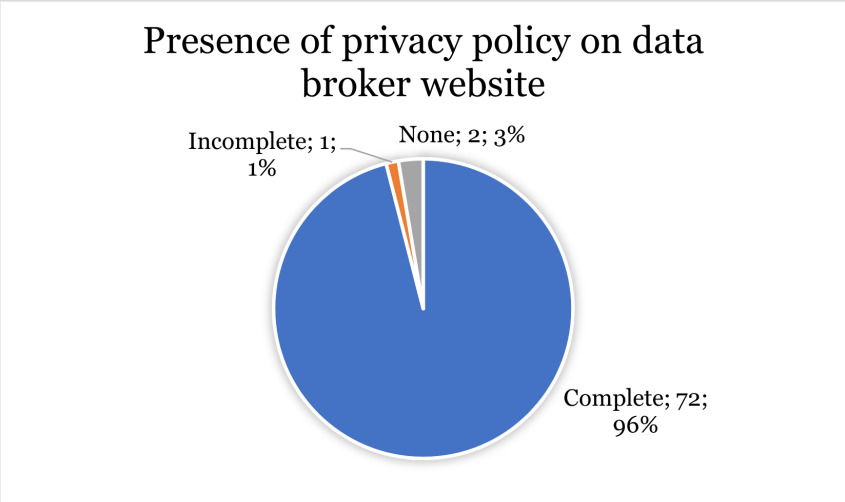


Figure 7: Presence of privacy policies

Another aspect to consider is the readability of a privacy policy. Indeed, 63% of Americans have recognized possessing very limited knowledge about the current laws and regulations aimed at protecting their privacy [4]. However, when asked about privacy policies, only 32% declare understanding none or little about what they read, with 13% affirming that they understand "a great deal" and 55% attesting that they understand some of it [4].

We have analysed data brokers' websites' privacy policies and inspected some key aspects such as the presence of a complete privacy policy (Figure 7), the length of the privacy policy (Figure 8) and the estimated reading level necessary to understand the policy's content (Figure 9).

The criteria for assessing a privacy policy as complete were set to be very loose and do not reflect its completeness in the sense of conformity to privacy regulations. As long as the privacy policy answered a user's main questions such as "what kind of data are collected?", "why are those data collected?" and "are those data shared with third parties?", the privacy policy was considered as complete. Figure 7 shows that the vast majority of data broker websites from the sample have a privacy policy that is complete enough to answer a user's most important questions.

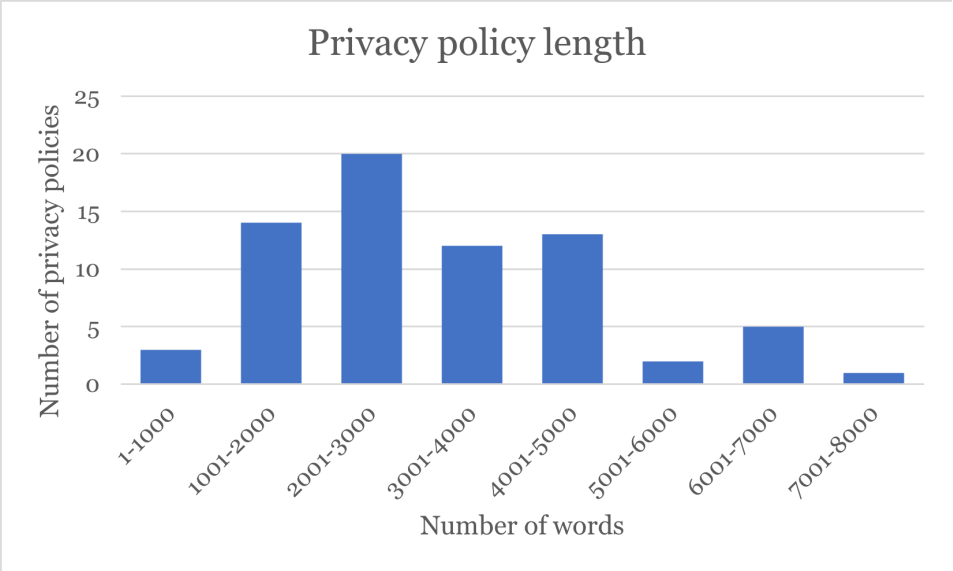


Figure 8: Length of privacy policies

Figure 8 presents the privacy policy's length in number of words. When a data broker presented multiple privacy policies for their different products by dividing the policy on multiple web pages, the entry was not considered for the graph. For fairness, sections of the privacy policy regarding specific regions (e.g. sections related to CCPA and California residents) weren't included when the name of the region is clearly indicated in a section's subtitle, unless the section is about transmission of data to foreign territories. When cookie policies were specified, they were also included when calculating the number of words. On average, privacy policies contained 3209 words, requiring 11m 40s of reading time, based on an average reading time of 275 words per minute. The longest privacy policy found in the sample had 7821 words, necessitating 28m 26s of reading time.

A privacy policy's reading level is estimated by wordcounter.net and represents the edu-

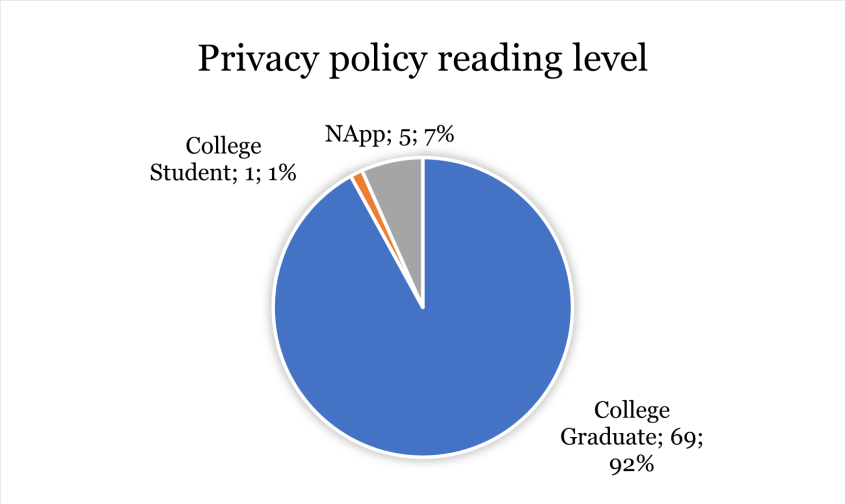


Figure 9: Reading level required to understand privacy policies

cation level required for a reader to understand the text's content. This estimation is based on the usage of non-frequent words in the text, which would require a certain level of literacy to be understood [34]. When the privacy policy was absent or when the data broker had multiple privacy policies for different products, we considered the measure as Non-Applicable (NApp). Figure 9 shows that most privacy policies require college graduate level of reading abilities to be able to understand its content.

In short, the vast majority of data brokers have complete enough privacy policies on their website, with some even having multiple privacy policies for their different products. However, those privacy policies require a certain amount of reading time (around 12 minutes) and some level of comprehension to be understood, so it is not guaranteed that users will take the time to learn their content.

3 Related work

Many researchers have studied the threat on personal privacy caused by data brokers or discussed how the data industry could be reshaped to give more power to the data providers whose personal information is being traded.

Governmental institutions have studied the activities of data brokers and made recommendations about regulations addressing issues with the current data industry. The Federal Trade Commission performed, in 2014, an in-depth study about data collection and usage practices of nine data brokers, by analysing the benefits and risks related to those practices and by making legislative recommendations to enforce the protection of customers' privacy [11]. In the same year, the Research Group of the Office of the Privacy Commissioner of Canada produced a report analysing data brokers in Canada and in the United States by comparing legislation in both countries [19].

Other works also proposed insights about the data broker industry and advocated for more transparency about their practices. Crain [10] warned that consumer-empowerment strategies based on transparency alone are not enough to counter the commodification of personal information, where people are treated as products instead of consumers. West

[33] examined the history and foundation of data capitalism, alongside its manifestation in the digital age. Green [13] questioned companies' honesty regarding the collection of audio data and proposed a framework where consumers' consent is required and businesses' usage of data are guaranteed to be non-harmful by the government. Yeh [35] analysed data brokers' practices and urged for more regulations on the data broker industry using the European Union (EU) data protection framework as model. This aligns with the comment made by Kuempel [17], who calls for comprehensive regulations similar to those in the EU. Geronimo [12] examined the ways and benefits of customer profiling, and the regulations around such practices. Helveston [14] argued that despite multiple regulations around the usage of purchased personal data, some forms of problematic abuse of data remains. Russell et al. [22] focused their study on student data markets, and called for more transparency. Baik [5] examined public discussions around CCPA from the view of different stakeholders and arrived at the conclusion that corporations consider privacy as a commodity while consumers view it as a right.

Studies were also conducted on specific consequences of data brokers' business in multiple spheres. For instance, Schneider [23] identified the relation between the data broker industry and scams. Rostow [21] analysed the risks of relational control resulting from data brokers selling information to individual buyers. Palk and Muralidhar [20] pointed out the "data inequality" faced by researchers, who need to pay data brokers in order to get complete and accurate data for their research. Venkatadri et al. [32] showed that advertising platforms, who possess large amounts of personal identifiable information (PII), can be exploited to infer a user's PII and activity.

Compared to the previous works, our study focuses on the specific market of personal information being sold to individuals, and analyses the personal information data brokers provide for free on their websites. This information can be accessed by any person, and when such information falls into the hands of ill-intended individuals, personal harm can be caused to data subjects. We also show experimentally how easy such information can be gathered and linked, by applying data matching techniques to profiles retrieved from data brokers' websites. Such techniques have been presented in numerous works, and other works have experimented their effectiveness by applying them in a real-world context.

Multiple publications described approaches used to link personal profiles from different data sources, in various contexts. Christen [8] detailed multiple aspects of record linkage, entity resolution and duplicate detection techniques, by conducting an in-depth analysis of the challenges and steps of a data matching process, and by providing multiple techniques used to compare and classify data. Shu et al. [24] reviewed identity linkage techniques used in the context of online social networks. Kruse et al. [16] provided an overview of data linking techniques for record linkage and entity linking by reviewing related papers. Those different data linkage techniques were applied by researchers in various contexts, for instance to reconstruct online profiles [2] or to link between records of the World Trade Center registries [3].

4 Methodology

This section explains the methodology used by the DROPLET system to collect and combine personal profiles fetched from ten person search sites in the United States. For this research, the person search sites analysed were BeenVerified, InfoTracer, InstantCheckmate, Intelius, MyLife, PeopleFinders, Spokeo, TruthFinder, WhitePages and USSearch. We selected 52 celebrity names and ran a search on each of those websites, without adding age

or location filters on the search results. For each person search site, the first 1000 profiles for a given name were retrieved and then combined with profiles from other person search sites to produce a more complete profile for an individual. In practice, only some common names require setting a limit of 1000 profiles and most names do not possess more than 200 profiles.

We only retrieved information that were available on those websites for free and did not analyse the full profile that would require payment to unlock. This allowed us to retrieve partial profile information such as full name, location (city + state), age, aliases, past locations, relatives' names (relations), partial or complete addresses, partial or complete phone numbers, or partial email addresses. More rarely, some person search sites will also provide information on the individual's gender, occupation, licenses, education or possible bankruptcies. Other types of information like social media accounts, property values or profile pictures can also be found, but they were not analysed in this study because they lacked information (only the name of the social media platform (e.g. Twitter) is provided and the profile picture is rarely present) or they weren't structured enough to be fetched automatically (property values). By abuse of language, we'll consider that location is presented in city + state format, since it is what data brokers suggest in their search functionality, although location is often presented in county + state format. Similarly, the field containing a list of potential relatives' names will be named "relations" throughout this research.

It is important to note that while free partial profiles already contain much information, full profiles could contain even more personal information such as criminal records, property records, complete addresses and phone numbers, photos, social media accounts, career and education history and financial information. Also, data brokers update their websites frequently, so the types of information available may vary depending on the time period. Our experiment was conducted in August 2021, and used information present on data brokers websites at that time.

4.1 General structure

The general process consists of three phases. First, profiles were retrieved from person search sites in the data collection phase. Each set of profiles from different person search sites corresponded to a database of profiles. Then, the profiles from different databases were linked together in the data linking phase. Finally, linked profiles were evaluated to estimate their truthfulness and usefulness. This procedure is illustrated in Figure 10.

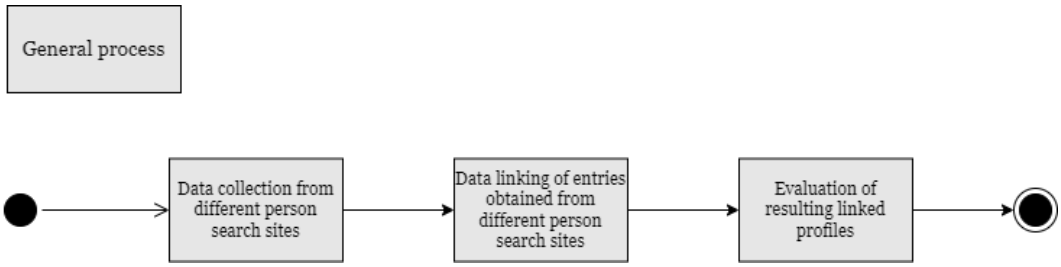


Figure 10: General process

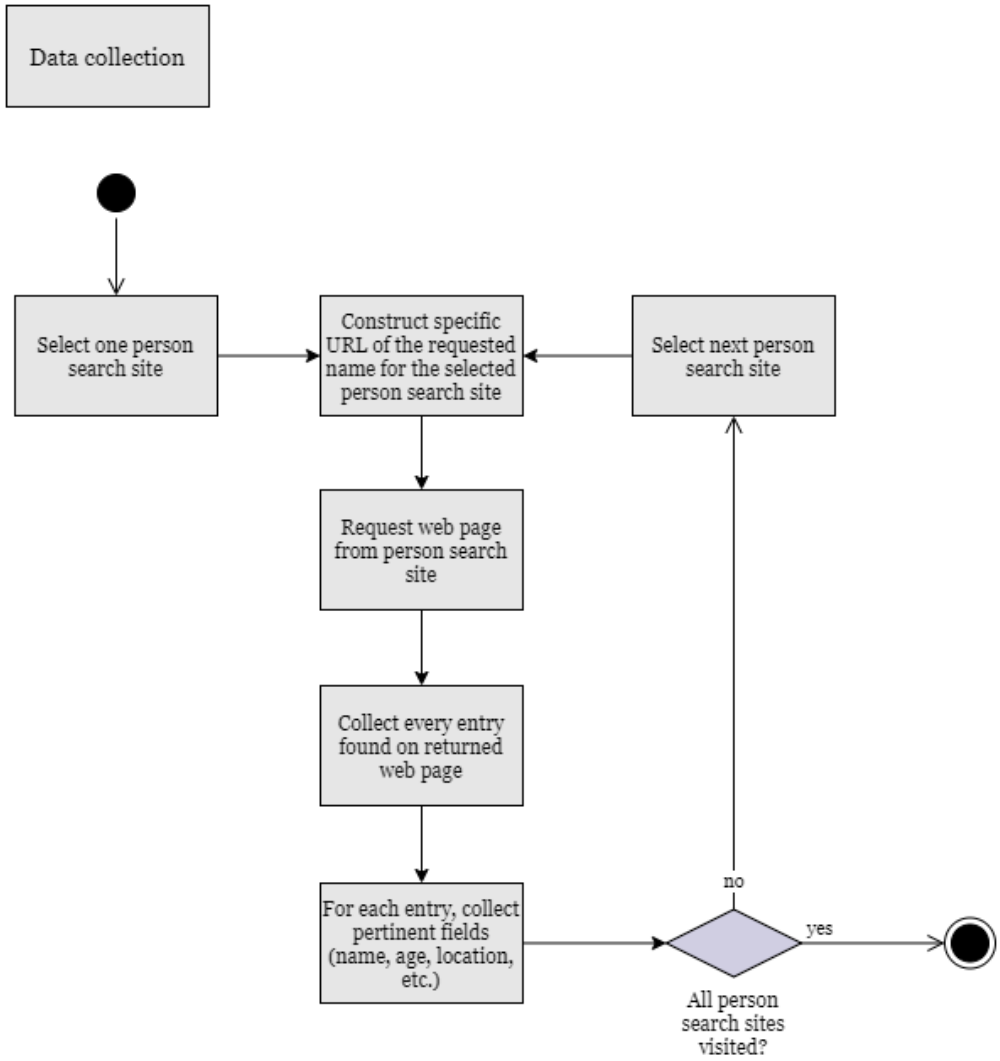


Figure 11: Data collection

4.2 Data collection

To collect profiles for a specific name, all person search sites were visited and the first 1000 profiles for each of those sites were retrieved. To do so, for each person search site, the URL for the requested name was constructed and the web page corresponding to the URL was retrieved. For instance, the base URL for a fictive person search site called namesearch may be www.namesearch.com/profiles/ and the specific URL for the name John Smith may be www.namesearch.com/profiles/John-Smith if the full name was requested in a single field, or www.namesearch.com/profiles/firstName=John&lastName=Smith when first and last names were requested separately.

On a side note, it is interesting to comment that while a normal person search procedure using the data broker website often forces users to wait several minutes before listing the resulting profiles, directly requesting the results with the constructed URL, which is the same as the final URL given by the data broker website after the time interval, enables us to view the profiles without the wait. Sometimes, merely adding "?loaded=1" after the URL allows to skip the wait process. This may correspond to a strategy used by person search sites to increase the sunk cost or to give the illusion to a user that much effort were put in by the data broker to produce the results, and the client should therefore pay a compensation for the generated outcome.

After profiles on the web page have been fetched, all pertinent fields from retrieved profiles are organized and stored to enable further analysis based on specific information fields. Figure 11 illustrates the data collection process.

4.3 Data linking

After profiles from different databases (person search sites) have been acquired, we proceeded to link profiles across databases. To do so, we first needed to clean and standardize some fields with variability in the data format. For instance, the locations can be written in multiple forms such as "City, State abbreviation", "City State abbreviation (no comma)", "City, State abbreviation, list of ZIP codes", etc.

We could also standardize phone and address fields, but the standardization was not performed because those fields were not used for comparison, due to the data fields being often partial or incomplete. Fields with names did not need further standardization since names were always presented in the same format across data broker websites. For the age field, only WhitePages differed from other data brokers by giving age approximates (e.g. 70s) instead of a precise number. Since this affected the comparison process, the differences where handled during the blocking step where comparison between ages were made.

The linking process aimed at linking all databases to a chosen reference database, as illustrated in Figure 12. We chose this methodology to avoid problems of transitivity that arise when profile "a" of an database A link to a profile "b" of database B which links to a profile "c" of database C, but profile "a" and profile "c" differ a lot and are not linked together. Always keeping the same reference database avoids such errors to propagate further when the number of databases is large. The major steps of the data linking process are summarized in Figure 13 and are detailed in following sections. The explanation of the process used to choose the reference database is also explained in later sections.

4.3.1 Data cleaning

As previously mentioned, location data from all profiles need to be standardized to enable proper comparison. This includes each profile's current location field, as well as all the locations in the past locations field. All locations were converted to lowercase and standardized to "city,state" format (no space after the comma). We also verified that locations were valid using cities data from the 2010 U.S. Census produced by the U.S. Census Bureau [30]. If the location appeared as an entry in the census data, it was considered valid because we had statistics on that location that would become necessary in later steps. If the location was not present, we assumed that the location corresponded either to a location outside of the United States, or to an unincorporated community for which demographic information was not detailed in the census data.

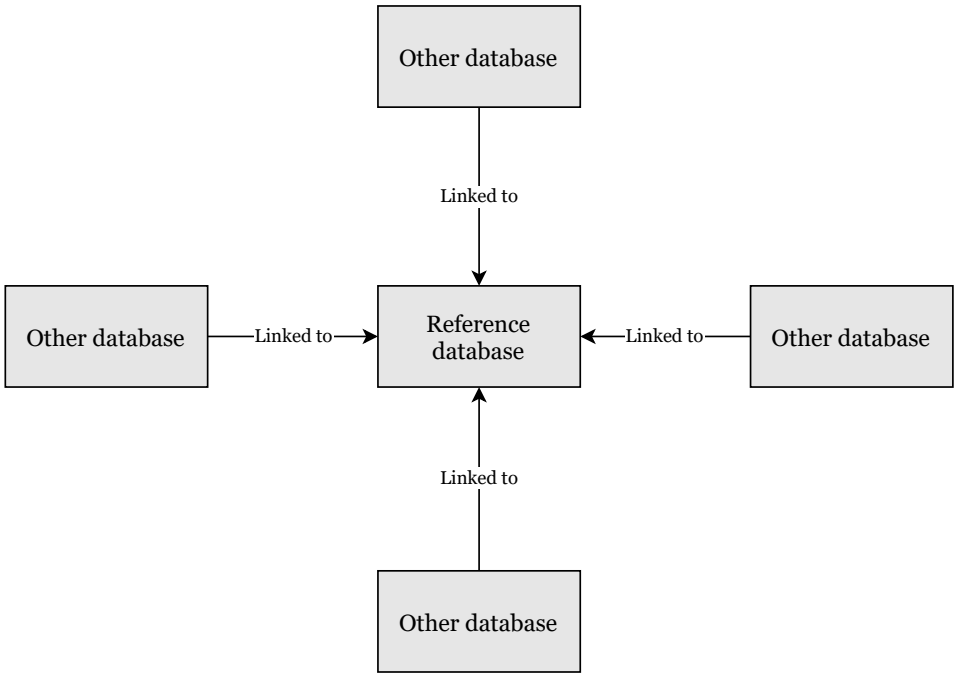


Figure 12: Linking all databases to reference database

In the first case, the state did not correspond to one of the 50 states of the United States or to the District of Columbia (e.g. a state code of AE, meaning "Armed Forces Europe"). Since no demographic information were present for those locations and such information would be needed in later steps of the process, the location was rejected and not accounted for in the linking process.

If the location corresponded to an unincorporated community in the United States, then the state would be valid, but the "city,state" combination may not be found in the census data. In that case, city information was simply removed and only the state information was considered during linking. Figure 14 summarizes the cleaning process for standardizing locations across profiles from different person search sites.

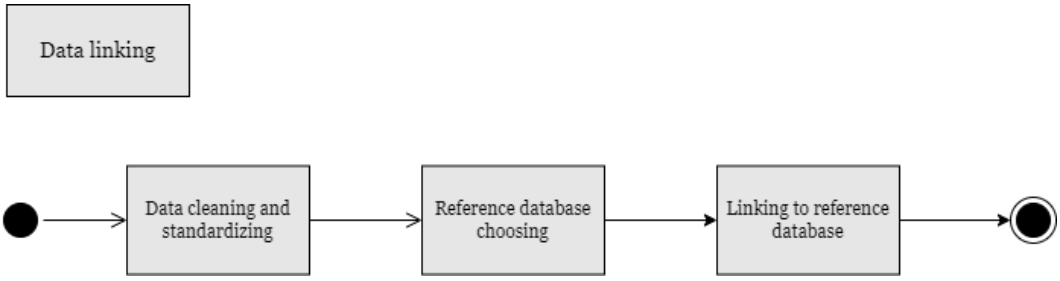


Figure 13: Data linking

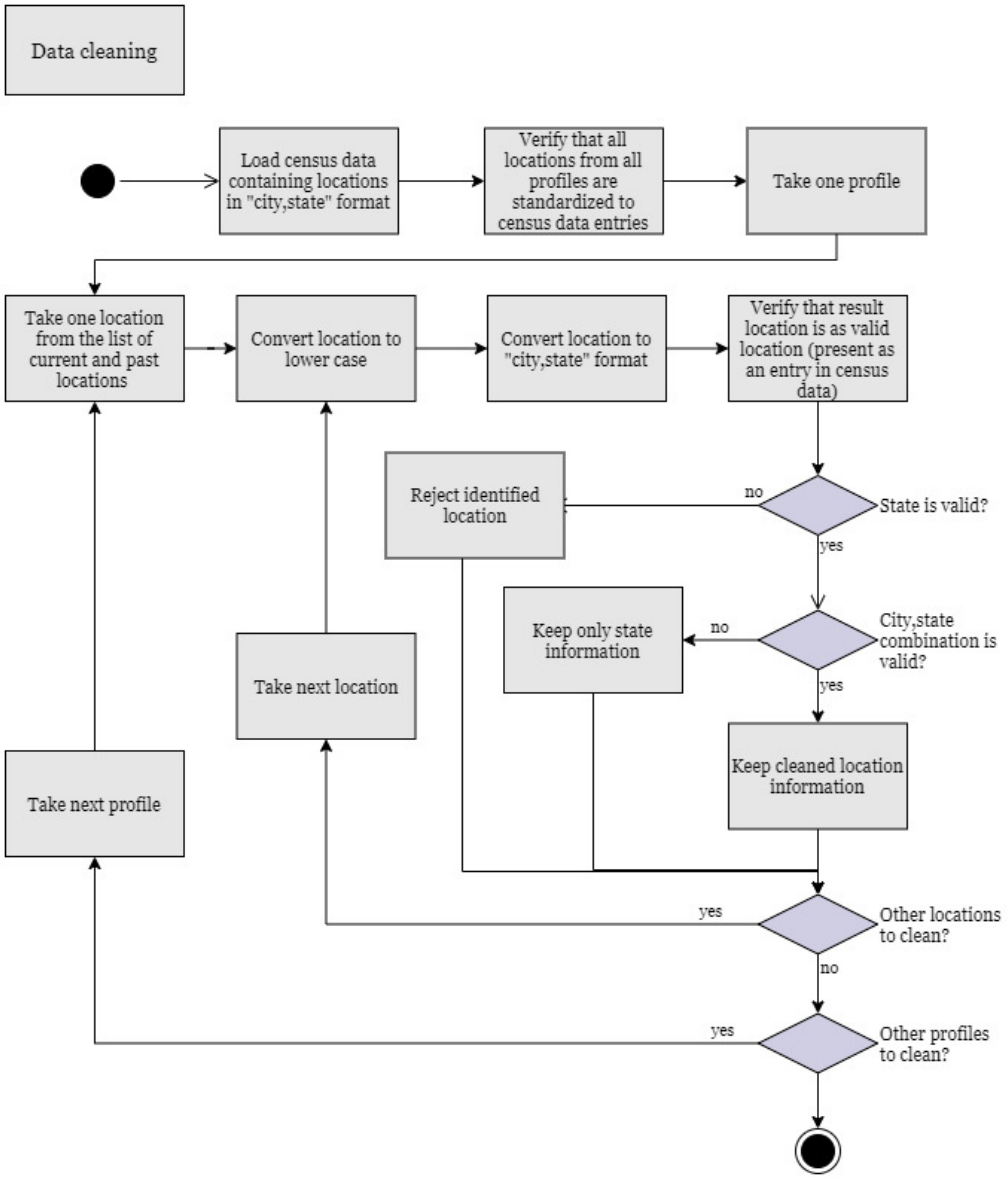


Figure 14: Data cleaning

4.3.2 Reference database choosing

The choice of the reference database can have an important impact on the number of profile links analysed and the size and number of successfully linked profiles. In order to maximize the number of linked profiles resulting from the linking process, we chose the database containing the greatest number of interesting profiles as reference database. A profile was considered interesting if it contained at least two past locations or at least one

relation. Those two fields were selected as criteria because they were later used to measure the similarity between two profiles. We decided to choose the reference database based on the number of interesting profiles instead of the plain number of profiles in order to avoid picking a database containing poor information in most of its profiles. This procedure is illustrated in Figure 15.

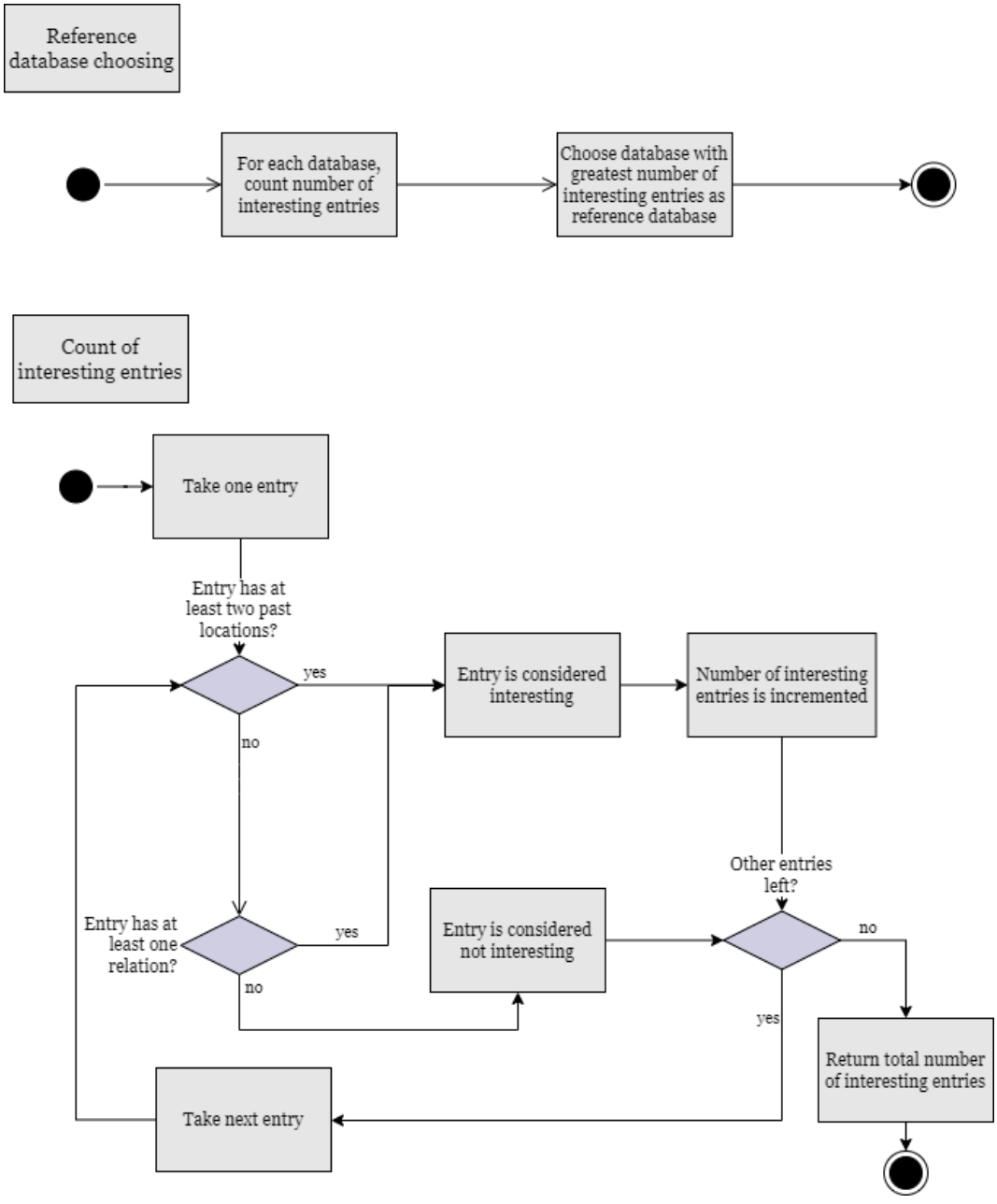


Figure 15: Reference database choosing

4.3.3 Linking pairs selection

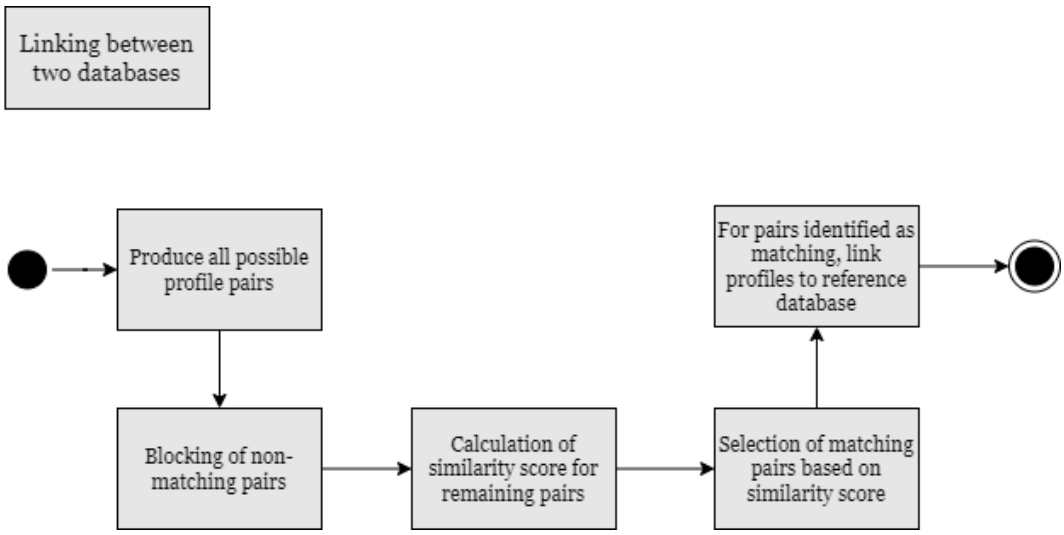


Figure 16: Database linking

To link profiles from different databases to the reference database, we needed to produce all potential matching pairs of profiles, and then decide for each pair if its profiles were considered to match. The production of the matching pairs corresponded to a cross product of the entries from the two databases. Then, to select the linking pairs, we used a procedure inspired by the general record linkage process presented by Christen [7]. That is, we first performed a "blocking" step to reduce the number of candidate pairs. Then, for the remaining candidate pairs, a similarity measure was computed and the matching decision of the pair depended on the value of the similarity measure. Figure 16 summarizes the process used to link profiles from two databases.

Blocking The blocking step aimed at removing the pairs that can be easily identified as non-matching, without having to compute a similarity measure. For achieving this, we considered the age and current location fields, which mustn't be in conflict between profiles from the same candidate pair. The name field was not compared because all profiles fetched should have similar names since the person search procedure used the same name when fetching profiles. If a conflict was identified between profiles from the same pair, the pair was rejected and no further analysis was done on the pair. When one of the fields was absent, we could not reject the pair because we could not be certain that the profiles did not match.

For the age field, a difference of three years was accepted, because we could not be sure that the profile of a specific person was up-to-date in a data broker's database. Though person search sites claim that their data are updated every day, it does not guarantee that every single profile from their database have indeed been updated with current information. When one of the profiles compared was a WhitePages profile, a difference of ten years was accepted in some circumstances due to WhitePages giving ages in intervals (e.g. accept 80s vs 87 but refuse 63 vs 70s).

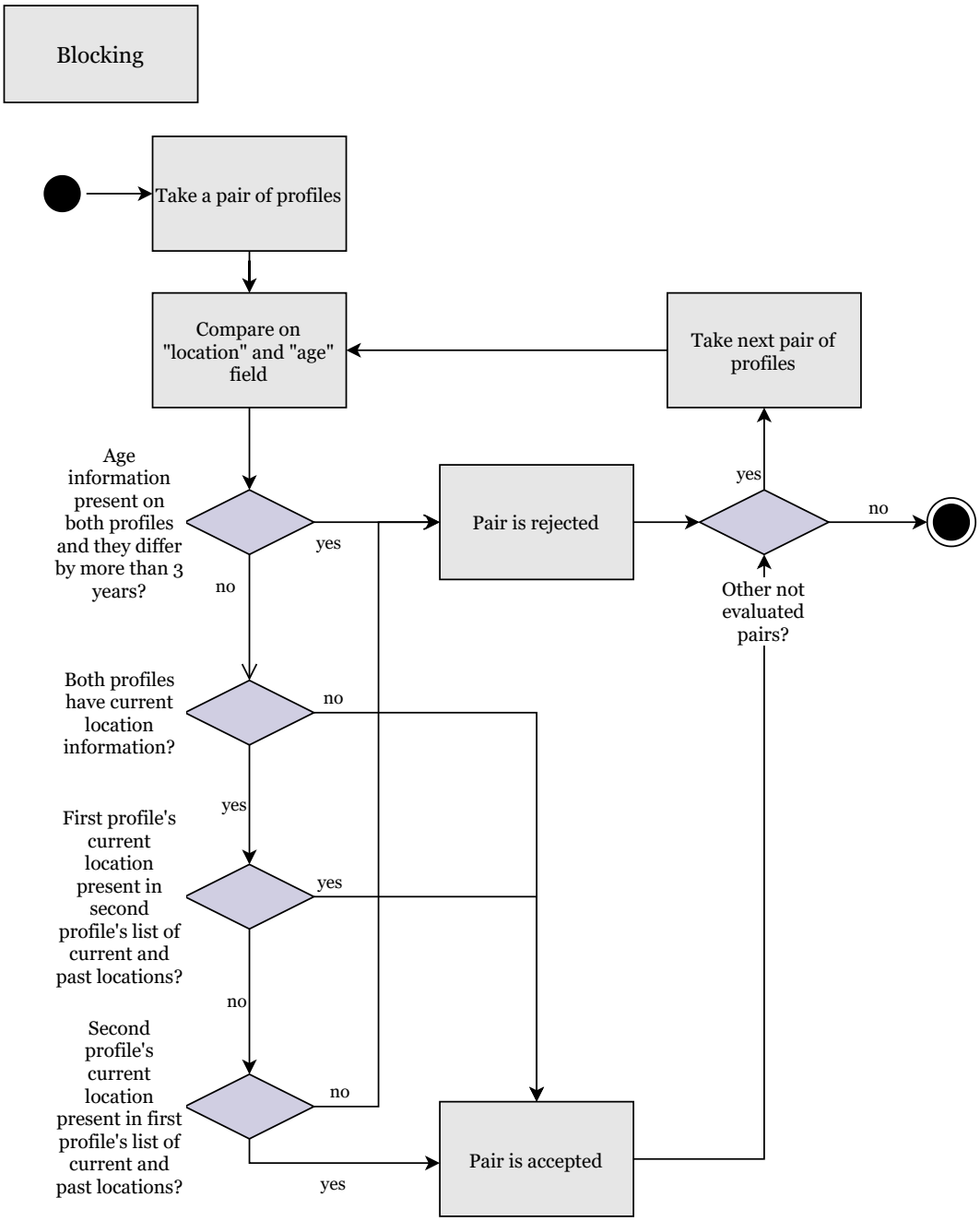


Figure 17: Blocking

For similar reasons, when comparing locations, the list of past locations was also considered. Since we assumed that one data broker may have more recent data than the other, it sufficed for one of the profiles to have its current location present in the other profile's list of past and current locations for the pair to be accepted. In that situation, it was assumed

that the second profile may contain more recent location information than the first profile, which would result in the second profile's location field not matching in the first profile, even though both profiles corresponded to the same individual.

Profile pairs that were not rejected in the blocking step required further analysis in the following steps to assess if they do match. The procedure for the blocking step is illustrated in Figure 17.

Similarity measure For remaining profile pairs that had not been rejected in the blocking step, a similarity score, or match weight, was computed to measure the similarity of the profiles in the pairs. The methodology used was inspired by the Fellegi-Sunter model and attributed, for each matching profile field between pairs, a positive score corresponding to $\lg(m/u)$, where $\lg$ denotes the binary logarithm function, m corresponds to the probability that fields do match when the profiles do correspond to the same individual, and u corresponds to the probability that the fields match by coincidence, even when the profiles do not correspond to the same individual.

Typically, m is close but not equal to 1 due to human errors when inputting the profile's data. The u probability vary according to the specific value of the field. For instance, the u probability will be bigger for a match on common names such as "Smith", but smaller for a rarer name. When fields do not match, a negative score of $\lg((1-m)/(1-u))$ is given for the field, using the larger u probability of the field values from the two profiles. We then take the sum of the match weight for each field as the global match weight for a profile pair.

To use the described method, we needed to decide which fields to compare, decide on the match condition for each compared field, and obtain the m and u values necessary for attributing a weight. Most profiles contained the fields full name, location (city + state), age, aliases, past locations, relatives' names (relations), partial or complete addresses, partial or complete phone numbers, and partial email addresses. Full names and aliases were not taken into account in the global match weight because we could expect those names to be very similar. Age was also not used because we could consider that ages match between pairs after the blocking step. Addresses, phone numbers and email addresses were not used either due to data incompleteness (e.g. when we only have some digits of a phone number or some letters of an email). Therefore, the fields used for comparison were the list of current and past locations (jointly called "locations"), plus the list of relatives. Those fields had a low chance of matching by luck, so they should produce a significant weight score allowing to identify true matches.

Since locations had been standardized in the data cleaning phase, comparison between two locations was direct and the locations were considered to match only when the exact same location text is presented. For relations, two relatives' names were considered to match if both first names and last names had a Jaro-Winkler similarity score exceeding 0.85. The middle name was not considered due to incompleteness of data and to absence of statistics on middle names frequency.

Since the fields used for comparison (locations and relations) had variable multiplicity across profiles (e.g. a profile with 3 past locations could be compared with a profile with 5 past locations). A decision must be made concerning how to treat list entries that did not have an entry in the other list with which to compare it. We attributed a null weight to those entries, but since this may boost match scores, when computing the match weight for non-matches, we chose locations with the smallest u probability because smaller u would decrease the match weight more.

Therefore, to compute the pair's similarity score for a field, for instance locations, we first identified the profile with the smaller number of locations. Then, for each location in this

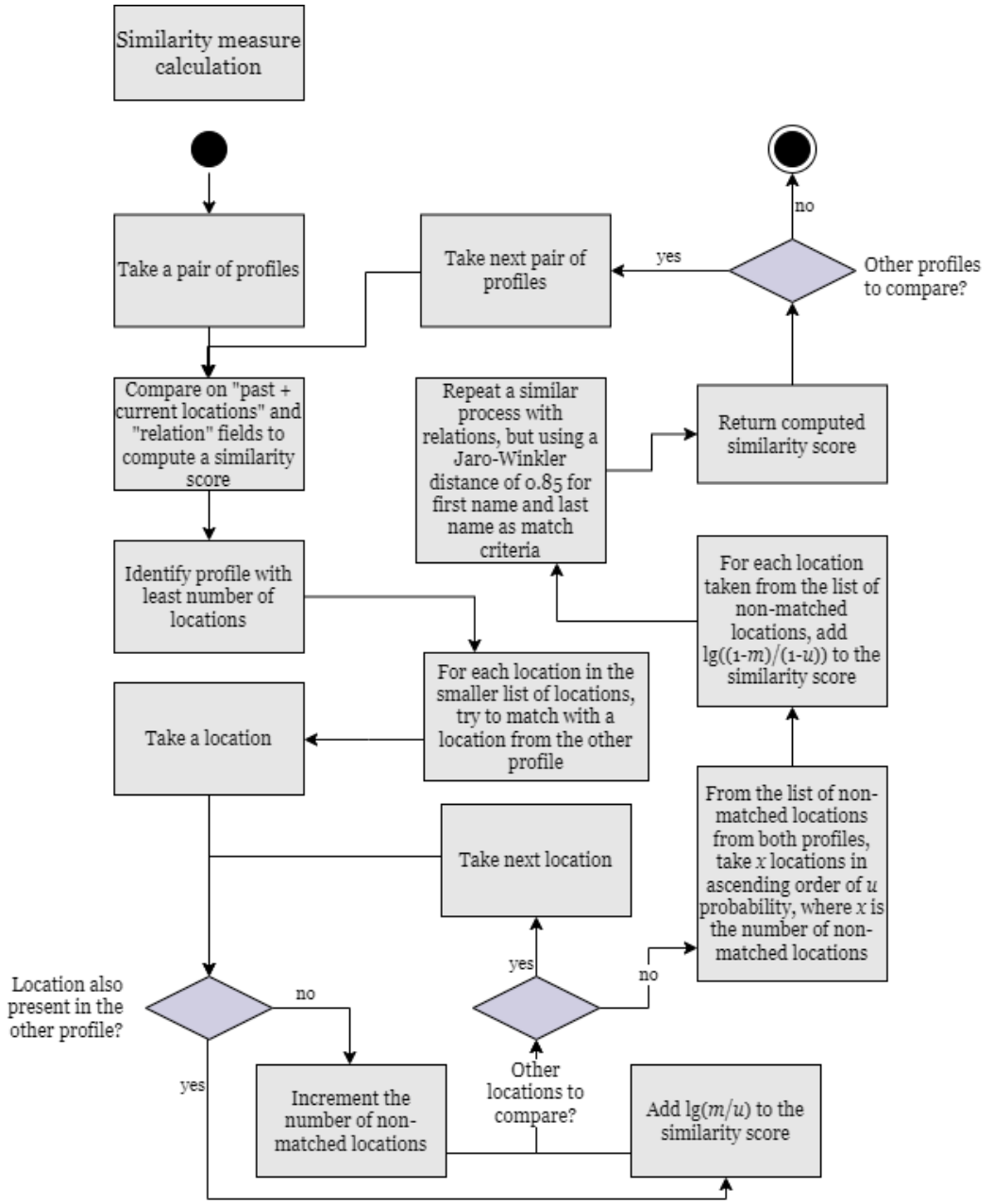


Figure 18: Calculation of similarity score

profile's locations list, we tried to find the same location in the other profile's location list. For each such matching location across profiles, we added $\lg(m/u)$ to the similarity score. Then, for each non-matched location in the smaller list, we added $\lg((1-m)/(1-u))$ to the similarity score, choosing the smallest u probability from the u probabilities of all non-matched locations across both lists.

The m value was set to 0.99 for locations, because we estimated that the probability of wrongly inputting a location should be low, since the list of possible location names is finite and a check had probably been made by person search sites when inputting a location. For names, the m value was set to 0.95, because more errors could be made when inputting a person's name.

The u probability for locations was estimated using U.S. Census Bureau's Subcounty resident population estimates for 2020 [30], when the location matched on city and state, and used the U.S. Census Bureau's 2020 Census Apportionment results' number of residents per state [29], when the location matched only on state due to the city having being removed in the data cleaning phase. After the population of the specific territory was obtained, it was divided by the U.S. total population in 2020 to estimate the u value.

For names, we could estimate both the probability of having a specific first name and the probability of having a specific last name, and then multiply both values to obtain the probability of a given full name. Middle names were not taken into account due to the lack of statistics on middle names frequencies. However, this methodology assumed that the probabilities of first names and last names were independent, which is not the case. Indeed, the probability of having a specific first name changes when the nationality of the last name is known. Yet, it was difficult to obtain official statistics of first names repartition depending on different last names, just as it was difficult to obtain full names statistics. Therefore, we restricted ourselves to performing an estimation of the u value for names by assuming that the probabilities were independent. To get statistics of names, websites such as howmanyofme.com could have been used, but to obtain up-to-date and official statistics, we preferred to use data released by official institutions, as we explain below.

The number of occurrences for different first names was computed using the U.S. Social Security Administration's "National Data on the relative frequency of given names in the population of U.S. births where the individual has a Social Security Number" [31] and we compiled the number of occurrences for each name entry from 1920 to 2020. The number of occurrences was then divided by the U.S. total population in 2020. It is important to note that the total number of names registered may not correspond to the total population of the U.S. in 2020, but the values should allow for a proper estimation. For first names who were not present in the list, we could deduce that the name occurred less than 5 times for each year between 1920 and 2020, and we assumed that it had 4 occurrences and used $4/[\text{2020 U.S. population}]$ as the u probability.

The number of occurrences for last names was computed using U.S. Census Bureau's "Frequently occurring surnames in the 2010 census" [28], which provided proportion of a given surname per 100 000 names. When a specific last name was not present in the list (grouped in "all other names"), we used the frequency of the least frequent last name present in the list (0.03 per 100 000).

For each pairs of profiles, a similarity score was produced using the previously explained method. The methodology is summarized in Figure 18, detailing the calculation of the similarity score for locations. Calculation of the similarity score was similar for names in the relations field, but this field only needs a Jaro-Winkler similarity for first and last names superior to 0.85, instead of an exact match, for the names to be considered belonging to the same individual.

Selection After a similarity score had been attributed to all profile pairs, we could perform a selection of profile pairs that would be linked between the two databases. We aimed at prioritizing links between pairs that contained profiles that are similar (higher similarity score).

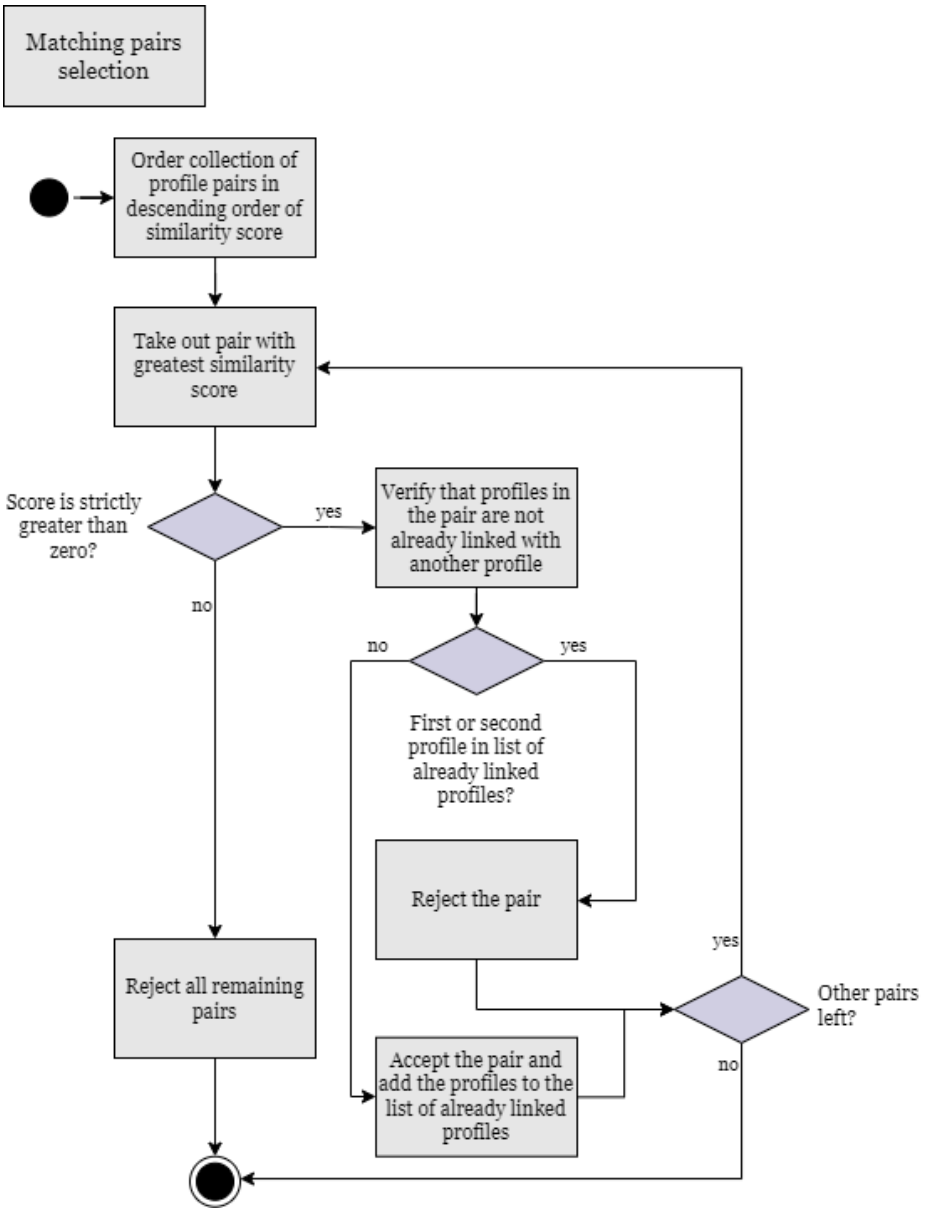


Figure 19: Matching pairs selection

Hence, we started by ordering profile pairs in descending order of similarity scores. Then, for each profile pair, starting from the pair with the greatest similarity score, we added it to a list of accepted links, if this pair did not contradict another pair already present in the list of accepted links.

For instance, if we were considering adding the pair (a_1, b_1) to the list of accepted links, but this list already contained (a_1, b_2), then we rejected (a_1, b_1) because a_1 was already linked to b_2, and the pair (a_1, b_2) had a similarity score greater or equal to the similarity

score of (a_1, b_1). In practice, to verify that a potential link did not contradict an accepted link, we could also maintain a set of already linked profiles and simply verify that both profiles in the potential link were not present in the set.

This process was repeated until all profile pairs had been analysed or until a pair with a similarity score equal or inferior to zero was encountered. In that case, profiles in the present and subsequent pairs did not have enough similarity to indicate that they may be linked, so it was not necessary to analyse remaining pairs.

The profiles of pairs in the list of accepted links were considered belonging the same linked profile. Therefore, the profile of the pair which belonged to the "other database" was linked to the profile from the "reference database". The process of selecting accepted links is outlined in Figure 19.

4.4 Efficiency

In this section, we analyse the time performance for the two main steps of the methodology: data collection and data linking. Only the execution time needed to produce the results is analysed, and the quality of the produced results is not considered in this section.

4.4.1 Data collection

Data collection is a process that relies heavily on the interface provided by data brokers and the turnaround time needed for data brokers to respond to the search query. Presentation of the resulted profiles can also influence the collection time. Profiles presented across multiple web pages will require more access time than when only one web page needs to be retrieved. Furthermore, some data brokers can detect that their web page is being requested by an automatic process (a bot) and will require solving a captcha to enable navigation.

Still, the time needed to complete collection of profiles is mostly linear in complexity and varies according to the number of retrieved profiles. Figure 20 presents the fetching time according to the number of retrieved profiles.

Variability in fetching time can be explained by some data broker websites, from which most profiles are retrieved, having a fetch time per profile different from the fetch time of other data brokers. For instance, when the total number of profiles is low, most data broker sites will provide a low number of profiles and the contribution of each data broker to the fetching time is similar. However, when requesting for a common name, some data brokers may limit the number of returned results, while other data brokers provide all the profiles corresponding to the request. If, among the data brokers returning all matching profiles, one of them needs navigating to each single profile page to retrieve complete information on the individual, the total fetching time can be boosted disproportionately.

Of the data brokers analysed, BeenVerified and MyLife seem to impose no restriction on the number of returned results, Spokeo limits their results at 300 profiles, InfoTracer limits their results at 200 profiles, and other data brokers never returned more than 100 profiles. Both BeenVerified and MyLife required navigating across multiple web pages to retrieve all results, which explains the higher fetching time when a high number of profiles need to be retrieved.

4.4.2 Data linking

Since profile pairs need to be produced by a cross product across databases and every candidate pair needs to be analysed, data linking is performed in quadratic complexity in

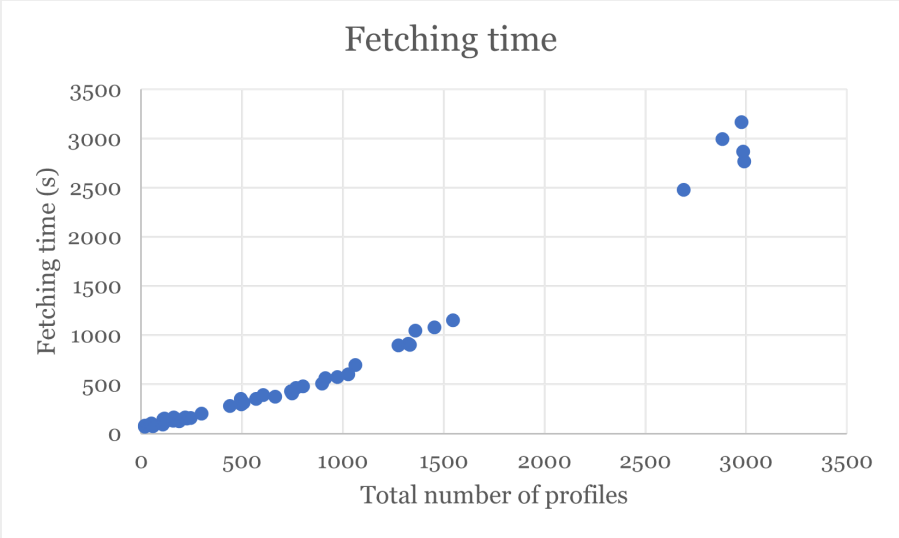


Figure 20: Fetching time

the worst case. However, due to different data brokers providing an unequal number of profiles, the execution time may vary a lot in practice. Figure 21 presents the linking time, depending on the total number of profiles from all the data brokers.

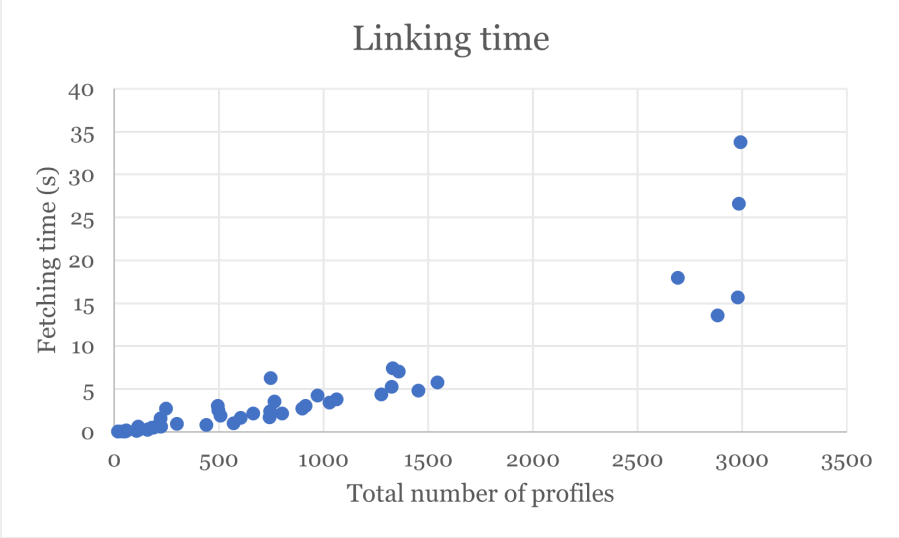


Figure 21: Linking time

Despite the linking process having a worse asymptotic complexity compared to the fetching process, in practice, for the number of profiles analysed, the fetching phase is by far more time-consuming than the linking phase. In fact, on average, it takes around 150 times longer to fetch and create the profiles than to link and produce linked profiles.

4.5 Limitations

Despite the efforts put into improving the DROPLET system to automatically fetch and produce quality linked profiles, its automaticity remains limited and some steps require human support.

The most important flaw of the system is that it cannot guarantee nor evaluate the quality of the produced results. Though we implemented strategies to help identify similar profiles across data brokers and we prioritize linking those profiles together, due to some limits of the system (e.g. fixed reference database, no global optimization of total similarity measure, etc.), the evaluation of the results require human judgment.

Also, some technical problems limit the automaticity of the system. For instance, data brokers put in place safety measures to limit access to their websites by a bot. For this reason, during the fetching process, we often have to manually complete a captcha to continue navigation. Moreover, the implementation of the fetching step heavily depends on the interface provided by data brokers. If a data broker changes the organization of his website, the implementation of the fetching phase would need to be modified to adapt to the new website.

Furthermore, as previously mentioned, some profile information presented on data brokers' websites could not be automatically retrieved due to some data being unstructured. Indeed, in order to retrieve information fields automatically and efficiently, data need to be organised and stored on separate web elements on the data broker's website. When such separation wasn't made, for instance if different types of information are combined in a short text, retrieval of pertinent information would need further text analysis that have not been implemented in this system.

Finally, the system does not estimate beforehand the number of resulting profiles from each data broker website, and does not take into account that some data brokers limit the number of profiles they return. Some data brokers like USSearch even refuse to return results if the number of matching profiles is too high. To reduce the number of profiles, we could add filters such as current location or age to the search request.

5 Experiment and results

This section presents in more details the implementation of the previously explained methodology using the Java programming language, and analyses the results obtained after experimenting the DROPLET system on 52 names.

5.1 Implementation techniques

The system was implemented in Java and classes were created for each data broker, since they presented different information in their profiles. Those data broker profiles all inherited from a Profile class, because they presented some fields in common (name, location, age, past locations, relations). The alias field was also added to the Profile class because it was often present in profiles. Profile fields, including age, were stored in text strings when a single entry was expected, and stored in HashSet or ArrayList of strings, if multiple entries were expected (e.g. past locations, phone numbers, etc.).

We also created classes for linked profiles and profile pairs, which contained pointers to corresponding profiles. For linked profiles, an additional field was added to indicate the profile that belonged to the reference database from the group of linked profiles.

A class was also created to handle interactions with the demographic statistics, namely first and last name statistics, city population statistics, and state population statistics. Before executing the system, methods from this class needed to be called to initialize the demographic information by loading those information tables into the system, in the form of a HashMap. This enabled us to fetch statistics from the table quicker than if those data needed to be retrieved from files every time.

Data collection from the data brokers was handled by a separate utility class, which established web connections and requested web pages from data brokers. We used the jsoup library version 1.14.1 and the selenium framework version 3.141.59 to interact with data broker's web pages. Both technologies were necessary because data brokers could block access for one of the technologies. Despite our efforts, for some data brokers, we still needed to solve a captcha in order to unblock access to the web page. When the system detected that a captcha was presented, it paused execution and requested human support for solving the captcha. Access to the web page was then granted for a period of time, before the data broker requested solving another captcha.

Finally, a class was responsible of linking profiles and calculating their similarity scores. This class used all other classes to store produced results or to collect necessary inputs and statistics.

5.2 Experiment

We performed multiple evaluations on collected results to identify a relation between the presence of personal information and the number of data brokers from which such information could be retrieved. In other words, we want to evaluate the influence of the size of the resulting linked profiles on the presence of personal information in those linked profiles, where the size of a linked profile is defined by the number of profiles from different data brokers that compose this linked profile.

Personal information were divided into two categories: sensitive information and non-sensitive information. Sensitive information included relatives' names (relations), phone numbers, addresses, email, work, licenses, bankruptcies and education. Non-sensitive information included age, current location, aliases, past locations and gender. For our purposes, information was considered sensitive if the concerned individual might be bothered if this information was known, either because it contained precise contact information enabling the possibility of spam, or because it contained potential harmful information that could damage an individual's reputation. Information was non-sensitive when, taken alone, it did not seem to produce much harm to the individual because the information was too vague. It could relate to many persons and the information should not cause any reputational harm to the individual.

5.2.1 Sensitive information

As previously mentioned, sensitive information included the fields relatives' names (relations), phone numbers, addresses, email, work, licenses, bankruptcies and education. Among those fields, many were provided for free by only some data brokers. For instance, work was found only on BeenVerified and Intelius, licenses and bankruptcies only on InstantCheckmate and education only on Intelius. For this reason, we could expect that the presence of those sensitive information would be significantly influenced by the size of the linked profile.

However, even when the data brokers accepted to provide some very sensitive information, its presence remained rare in the profiles returned by the data broker. This limited the growth of the quantity of information in a linked profile with the addition of a new profile in fields such as licenses, work, bankruptcies and education.

5.2.2 Non-sensitive information

Non-sensitive information included the fields age, current location, aliases, past locations and gender. Most data brokers would accept to provide those fields in the profiles they return, so we could expect that the presence of those fields would be high even for smaller linked profiles. Still, complementing the reference profile (from the reference database) with another profile could complete some missing fields. For bigger linked profiles, we could expect that all non-sensitive information fields would achieve high presence statistics.

Furthermore, though not evaluated in this paper, having multiple profiles corroborating the same information might increase the trustworthiness of those information in the profile. When different information contradict each another, their trustworthiness decreases, but the probability that the correct information is found in one of the profiles increases. For instance, if one profile indicates that an individual's phone number is a, and another profile gives the phone number b, it might be that one of those phone numbers is an old number that is not used anymore, but the probability that one of those numbers is still used increases.

The results of the evaluation are presented in the following sections. However, before discussing those results, we need to first analyse in more details the characteristics of profiles returned by different data brokers, because they will influence how results should be interpreted.

5.3 Comparison of data brokers

The data brokers analysed in this paper are BeenVerified, InfoTracer, InstantCheckmate, Intelius, MyLife, PeopleFinders, Spokeo, TruthFinder, USSearch and WhitePages. For the same queried person name, they could produce a very different amount of results. Figure 22 presents the average number of profiles returned by different data brokers for the sample of 52 names. Recall that for some data brokers, a limit on the number of results may have been imposed, thus lowering the average number of profiles. For USSearch, only the results for 47 names were considered, because results could not be retrieved for 5 of the names due to the number of results being too large.

From Figure 22, one would expect that data brokers like Spokeo, BeenVerified or MyLife would often be selected as reference database due to their high average number of profiles. However, in practice, the quality of the profiles was also analysed, and the profiles who may not produce good linked results were not taken into account when choosing the reference database. Figure 23 shows the number of times a specific data broker was selected as reference database during our experiment. We observe that despite Spokeo, BeenVerified and MyLife offering a high number of profiles, InfoTracer and WhitePages had more complete profiles for our purposes, and they were therefore chosen as reference database more often.

The choice of the reference database is very important as it not only influences the number of successful links, but also the information present in smaller-sized linked profiles. Indeed, if the reference profile taken from the reference database already contained a lot

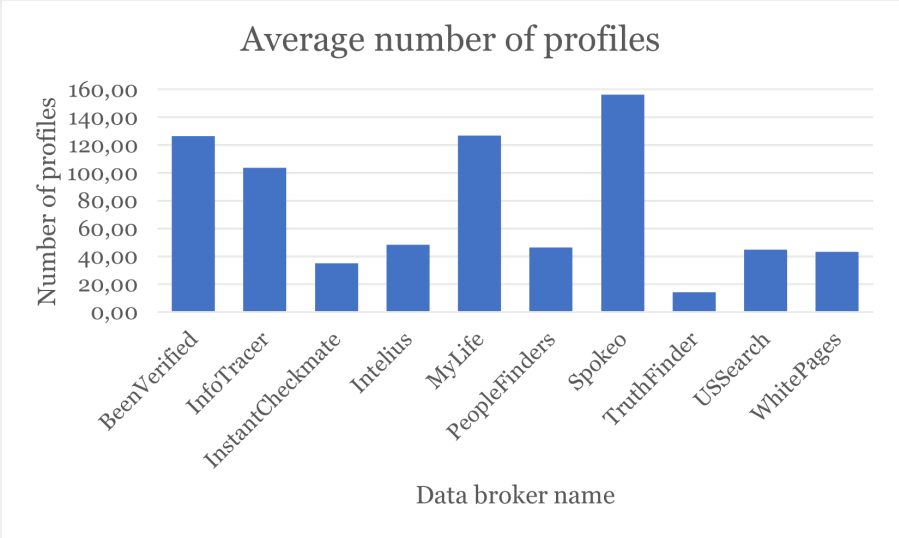


Figure 22: Average number of profiles provided by the data broker

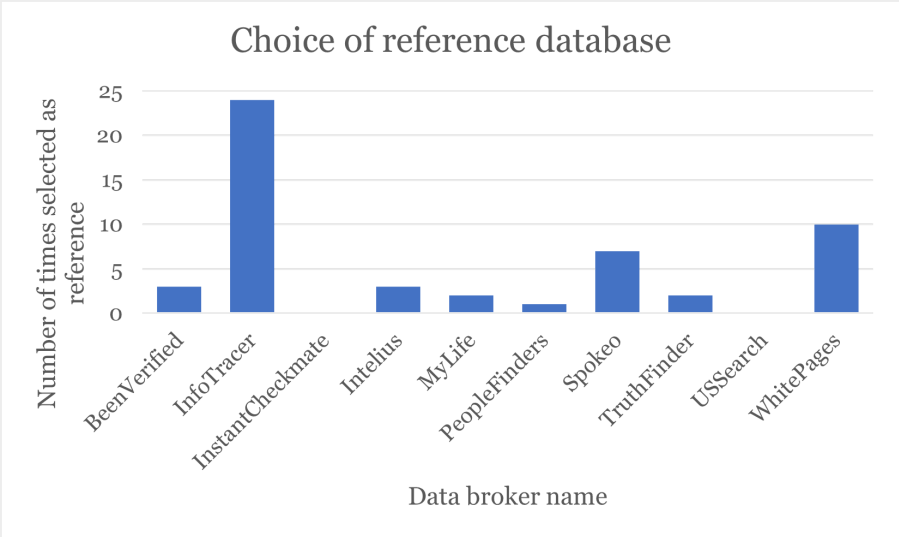


Figure 23: Number of times a data broker is chosen as reference database

of information, linking with other profiles with less information would not add much information to the linked profile. The growth of the size of a linked profile would therefore impact less on the quantity of information present than if the base profile contained limited quantity of information.

For this reason, it is important to evaluate the information that profiles from different data brokers tend to contain. Those statistics also help evaluate which types of personal information are easy to find on person search sites, and which types of personal information can be found only on some data broker websites. Figures 24 to 33 present the average presence

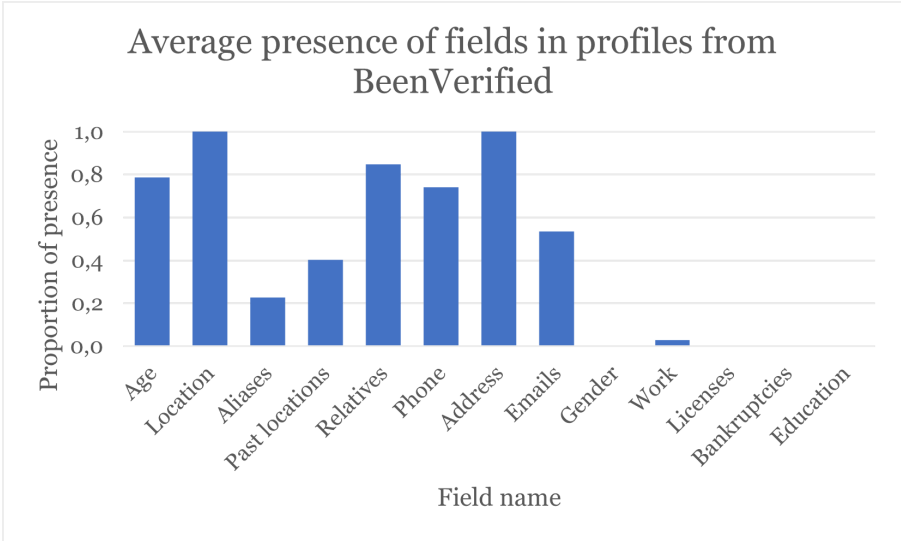


Figure 24: Information in BeenVerified profiles

of information fields in profiles returned by different data brokers, with a proportion of 1.0 meaning that the information was present on all profiles returned by the data broker.

As seen in Figure 24, BeenVerified profiles contained a lot of information, with 9 of the 13 analysed fields being potentially present in a BeenVerified profile. Furthermore, we notice that all BeenVerified profiles contained current location and address information, but past locations, a very important field, was not very present in BeenVerified profiles. Also, BeenVerified profiles sometimes provide work information, a sensitive information rarely present in profiles provided by other data brokers.

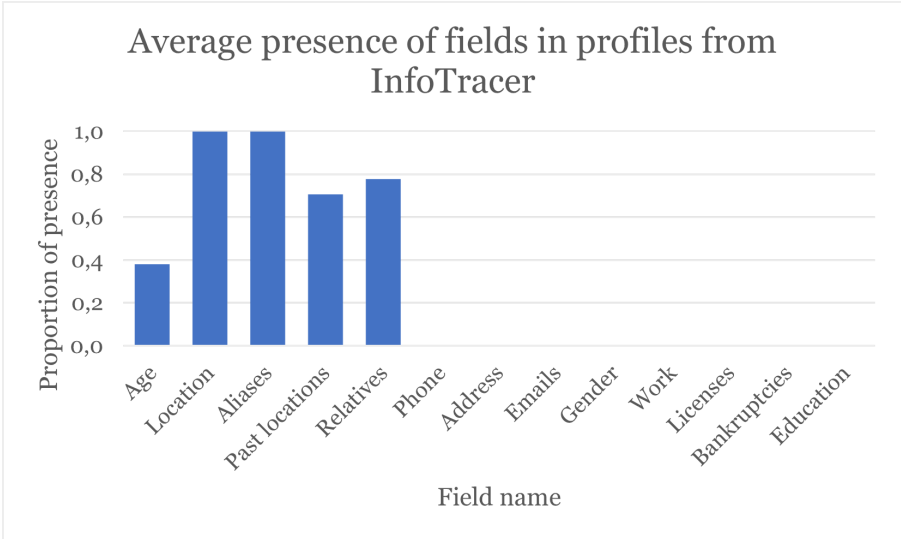


Figure 25: Information in InfoTracer profiles

Figure 25 shows that InfoTracer profiles did not contain many types of information, but the important fields for profile linkage (past locations and relations) were often present in those profiles. Also, alias and current location information were present in all InfoTracer profiles. The presence of the important fields for profile linkage and the absence of other sensitive information fields made InfoTracer an ideal candidate as reference database, as we could expect to fill the missing sensitive information fields with profiles linked to the reference InfoTracer profile.

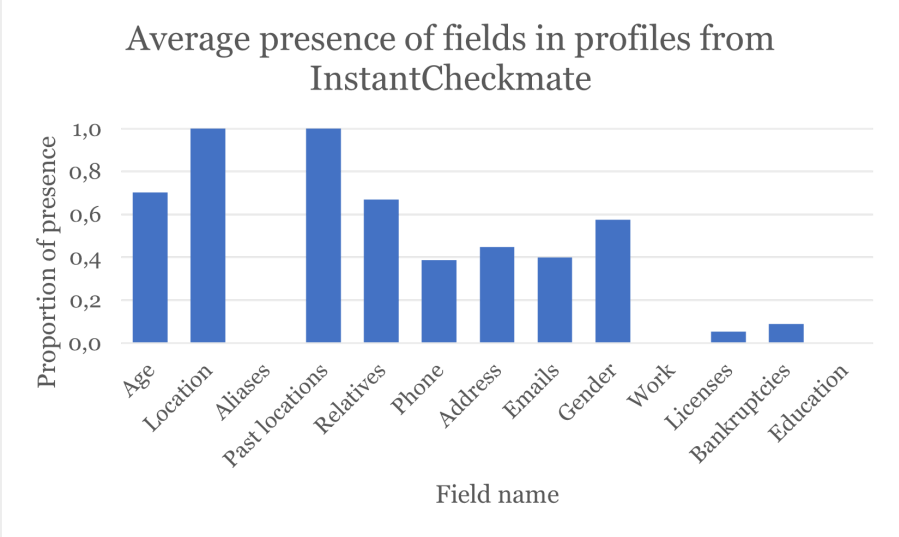


Figure 26: Information in InstantCheckmate profiles

We see from Figure 26 that InstantCheckmate profiles contained important amounts of information, as 10 of the 13 analysed fields could be found in an InstantCheckmate profile. Current and past location information were present in all InstantCheckmate profiles, but we notice that, curiously, alias information was not provided. Also, InstantCheckmate could provide very sensitive information such as licenses acquired by an individual and his possible bankruptcies.

From Figure 27, we can establish that Intelius profiles often contained only basic information such as age, location, past locations and relations. However, on rare occasions, sensitive information such as an individual's education record or current work could also be found.

Figure 28 indicates that MyLife profiles contained more than half of the fields analysed in this research. Moreover, though not analysed in this paper, MyLife profiles often contained more information on individuals than what we could retrieve, such as property records of ethnicity. Yet, this information was not presented in a way that could be easily retrieved and organized by an automatic process, so this information was not compiled in this figure. Despite those limitations, MyLife profiles still made available important amounts of information, with current location and addresses present on all MyLife profiles.

Compared to other data brokers, Figure 29 seems to indicate that PeopleFinders did not provide too much information on individuals. However, for fields that were available, information is often present. This suggests that most of its profiles were of high quality. In particular, current location information was present in all profiles.

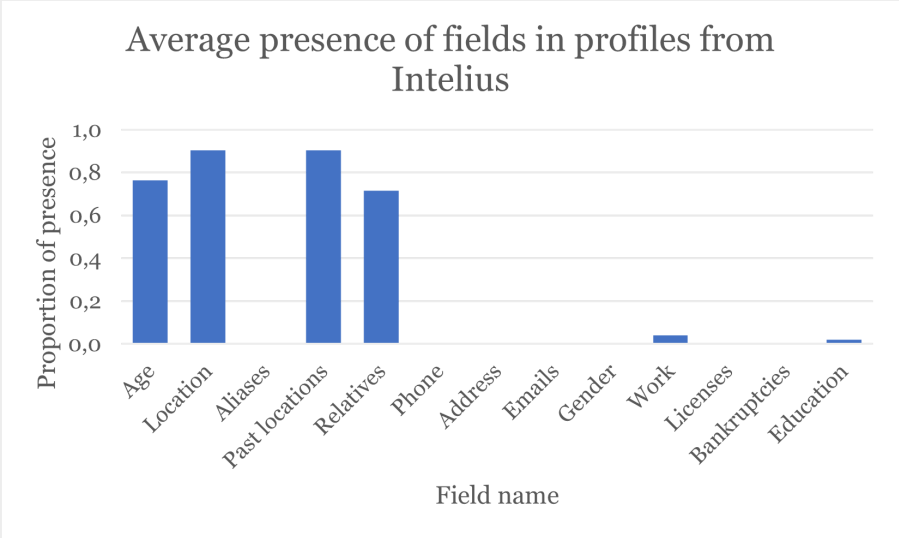


Figure 27: Information in Intelius profiles

Figure 30 shows that Spokeo profiles did not present much sensitive information aside from relative's names, just like all other data brokers, and address, which was present in all Spokeo profiles. However, even though current location and address were often present in Spokeo profiles, those profiles only occasionally contained other pertinent information. This explains why despite Spokeo returning the largest number of profiles on average, it was selected as reference database only on some occasions.

As shown in Figure 31, TruthFinder profiles presented a different information distribution

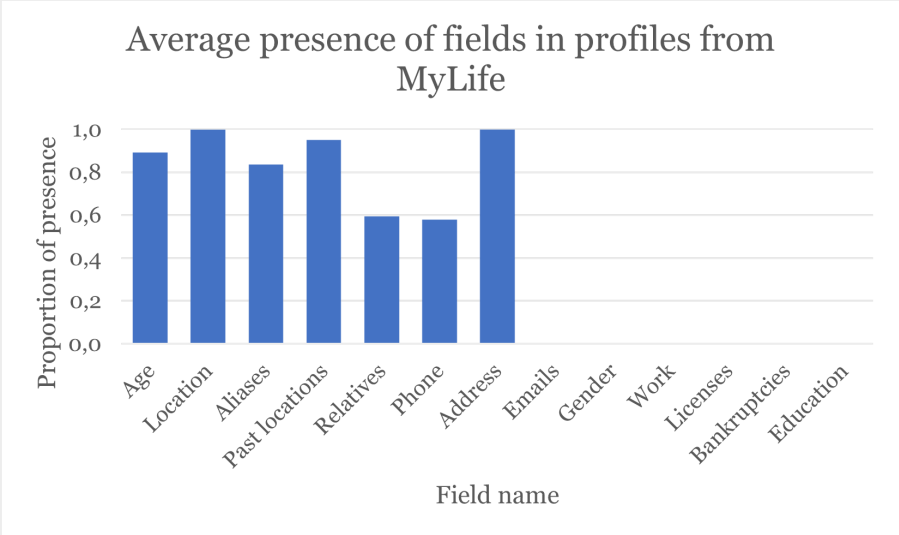


Figure 28: Information in MyLife profiles

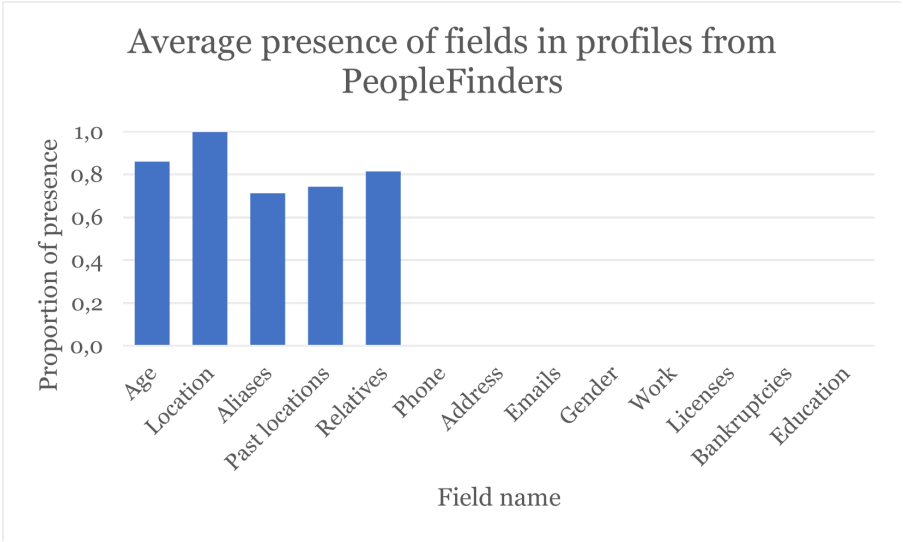


Figure 29: Information in PeopleFinders profiles

than other person search sites that we have analysed. Indeed, it did not provide current location, address and alias information, but would often give a person's email or gender, which other data brokers did not always provide. However, despite not indicating in which location an individual is currently at, TruthFinder would give a list of past locations, which may also contain a person's current location though not marked so. We can also notice that TruthFinder seldom presented profiles with missing information: most fields of a returned profile were complete.

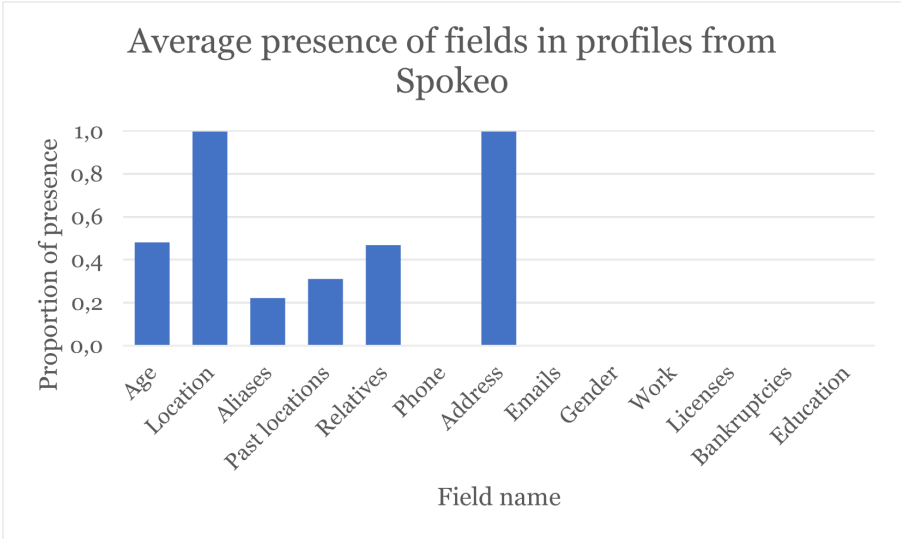


Figure 30: Information in Spokeo profiles

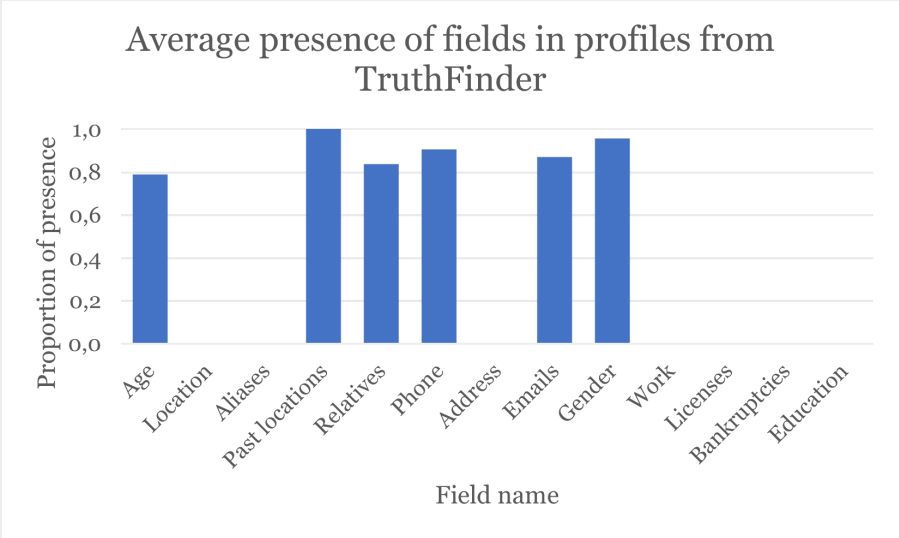


Figure 31: Information in TruthFinder profiles

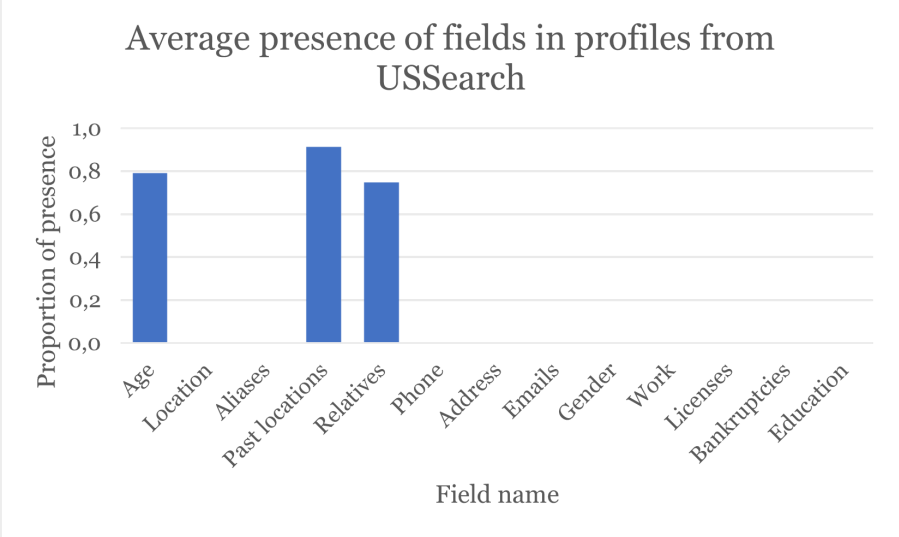


Figure 32: Information in USSearch profiles

From the data on Figure 32, we can observe that USSearch did not provide many types of personal information in its free person search service, nor did it always provide them. Indeed, it only provided age, past locations and relatives' names, not even indicating a person's current location.

As Figure 33 shows, WhitePages profiles had more than half of the fields analysed in this study, and did not contain many empty fields, except for aliases. Also, all WhitePages profiles contained current location and address, plus almost always having an age field on their profiles. The quality of WhitePages profiles justified WhitePages often being selected

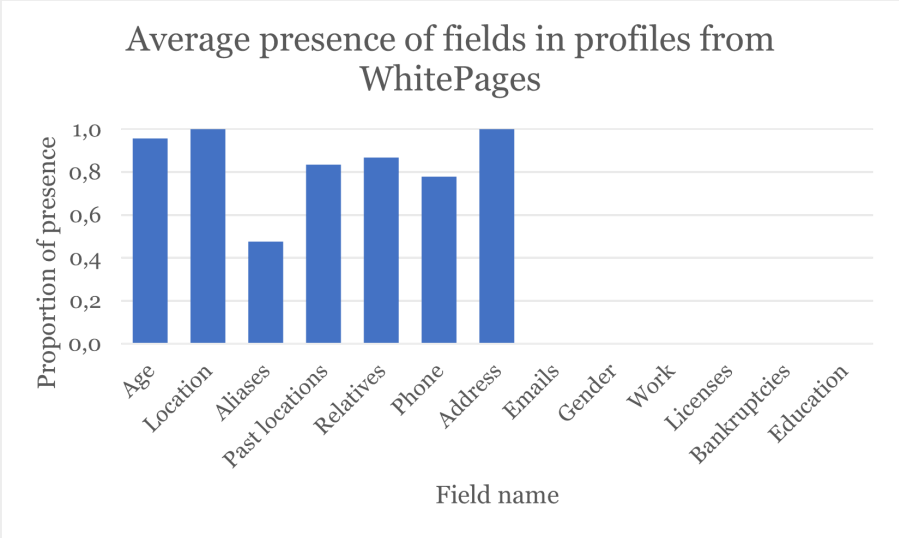


Figure 33: Information in WhitePages profiles

as reference database despite its relatively low number of result profiles compared to other data brokers.

From Figures 24 to 33, we can observe that some types of information were given by all data brokers (age, past locations and relations), whereas some other types of information were given by most data brokers (current location). Yet, other types were given by many data brokers (aliases, phone numbers, addresses and email) or given by only a few data brokers (gender, work, licenses, bankruptcies and education). Among the last category of information, gender was provided only by InstantCheckmate and TruthFinder, work only by BeenVerified and Intelius, licenses and bankruptcies only by InstantCheckmate and education solely by Intelius. This information present on only some data brokers' websites are more likely to be obtained if the size of the linked profile was large, since there would be a bigger probability that a profile containing this information has been linked.

Equipped with the previous insights about the distribution of information across different person search sites, we can proceed to the analysis of the influence of a linked profile's size on the presence of some information fields in its profile.

5.4 Influence of linked profile size

We ran the experiment of searching for a person's name and linking their profiles across person search websites 52 times, allowing us to analyse the proprieties of the linked profiles. On the average, 736 profiles from different websites were retrieved for each name, with a median profile number of 497. With those profiles, we produced linked profiles with size varying from 2 to 10, with 10 representing a profile present on all data broker websites. On the average, for each name, the biggest linked profile we could construct from the profiles fetched with that name contained information from 7.38 person search sites. Figure 34 presents the average number of linked profiles of different size produced for each searched name. We can observe that smaller linked profiles were much easier to obtain than larger linked profiles.

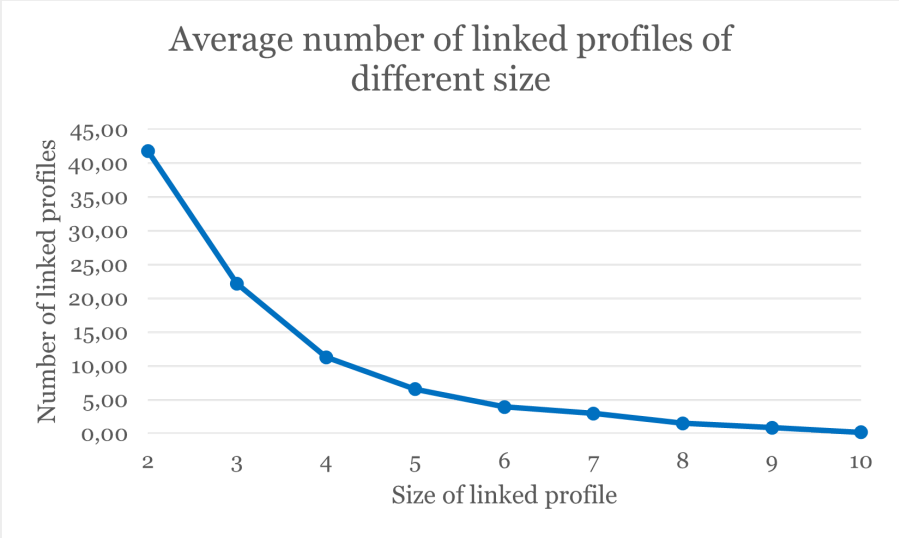


Figure 34: Average number of linked profiles of different size for a single name search

5.4.1 Sensitive information

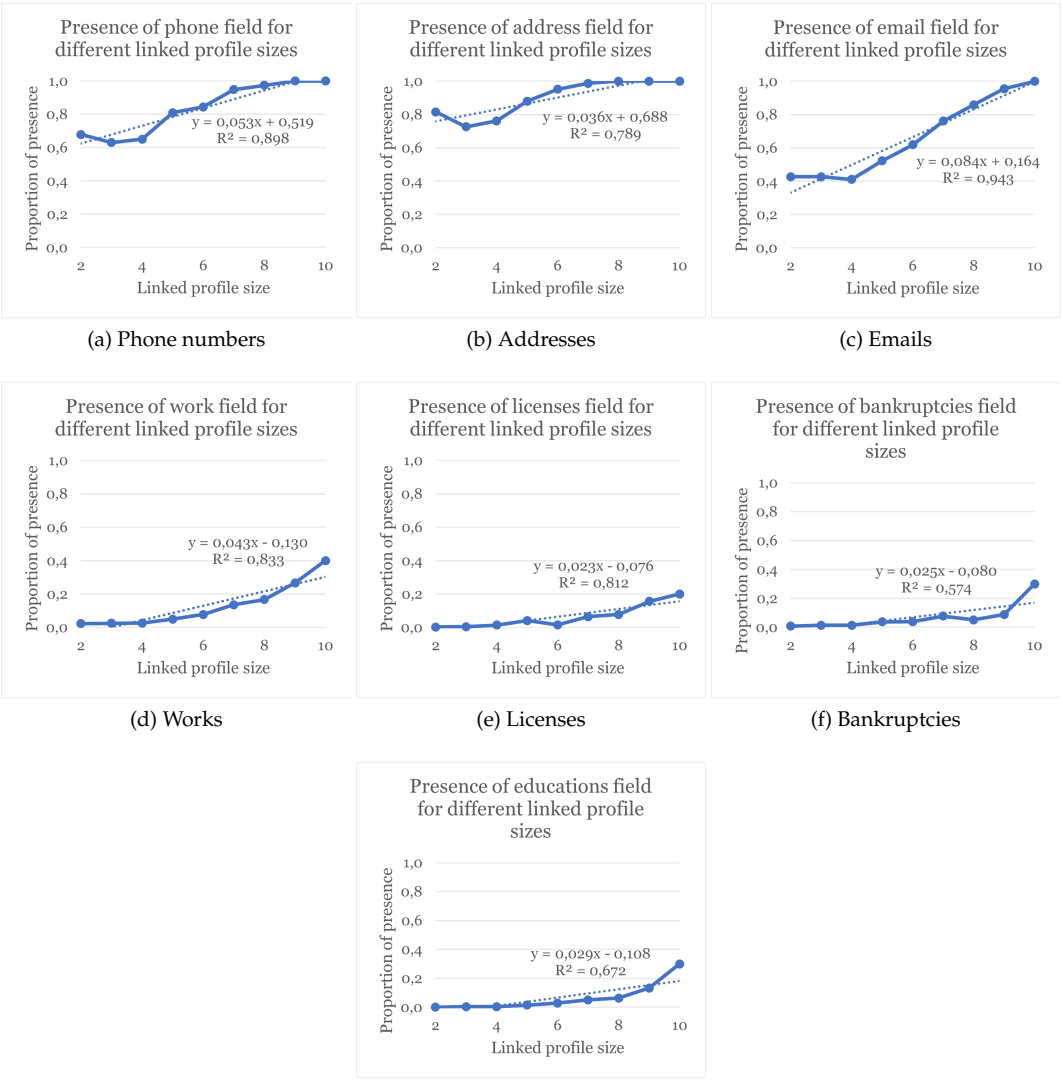
The presence of sensitive information in different linked profile sizes is analysed in Figures 35a to 35g. For the relations field, we preferred to analyse the number of identified relatives, and present the results in Figure 36.

To estimate the impact of a bigger linked profile size, we used the linear regression model, which informs us about the impact of the increase in the value of one variable (the linked profile's size) to the value of the other variable (presence of information fields), by the slope of its line. We can also get an estimate of the credibility of this slope with its R^2 score, which should be close to 1 when the data points fit the linear model.

The email field presented the biggest growth, with a regression slope of 0.08, followed by the phone field with a regression slope of 0.05. This value indicates that on average, the probability of finding a person's email increases by 8% for each additional profile we were able to link to this person. Address and work fields had a slope of 0.04, bankruptcies and educations a slope of 0.03 and licenses a slope of 0.02. However, for bankruptcies and educations, the low R^2 score hints that the higher slope may be mostly due to the high number of the last value (linked profile of size 10), and did not indicate a stable growth for other profile sizes. Also, to compare slopes, we assumed that the growth followed a linear regression, which may not be valid for all kinds of information.

We can also note that some information were almost always present in a linked profile with sufficient size, while other information were still rarely present despite linking profiles. Phone numbers, addresses and email all had relatively high presence even with low linked profile sizes, but their presence was even more important when a high linked profile size was reached. Other fields did not demonstrate such remarkable growth, with information presence on those fields remaining low in big linked profiles.

For the relations field, Figure 36 shows that on average, the number of relatives increased by 0.90 when the linked profile's size increased by 1. Also, from a profile of size 2 to a profile of size 10, the quantity of identified relatives almost doubled. However, this tendency may not follow for bigger profile sizes exceeding 10, as the majority of relatives of an individual



(a) Phone numbers (b) Addresses (c) Emails

(d) Works (e) Licenses (f) Bankruptcies

(g) Educations

Figure 35: Presence of sensitive information for different linked profile sizes

would most likely already have been identified. Furthermore, it is important to note that some relatives' names may point to the same individual, but we were unable to identify this case due to a woman having changed her last name after marriage. Indeed, some data brokers may present an individual's maiden name, and other data brokers will provide the name after marriage. Though, in such cases, a new relative was not identified, we still identified new information on a specific relative, so we accepted that those cases increased the relations' information slope.

Our results show that the email field had a bigger growth than the other fields. In our case, this can be explained by the fact that WhitePages, a data broker often selected as reference

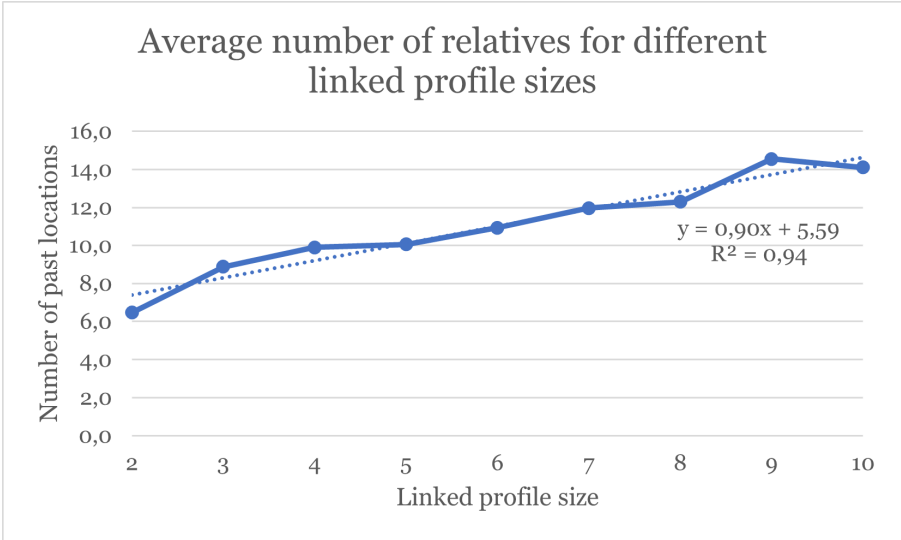


Figure 36: Number of relatives for different linked profile sizes

database, did not provide email information. Thus, when WhitePages was selected as reference database, the email field could experience a bigger growth, since this information would be absent for low linked profile sizes. In general, we can see that presence of information for most of the sensitive information fields had important growth. The relations field also showed a significant growth as more relatives were identified.

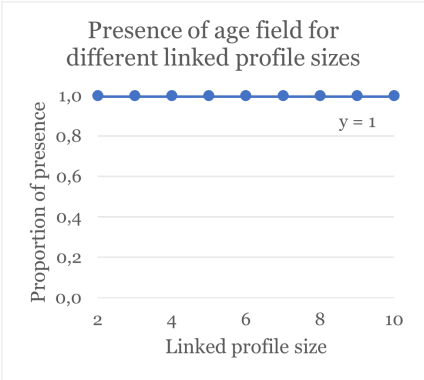
5.4.2 Non-sensitive information

Figures 37a to 37d illustrate the presence of non-sensitive information in different linked profile sizes. Similarly to the relation field, we preferred to analyse past locations by analysing the quantity of identified past locations, and present the results on Figure 38.

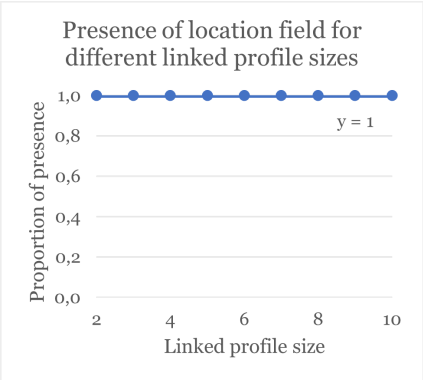
Figures 37a to 37d show that a significant growth in information presence was only observed for the gender field, with a regression slope of 0.12. Other fields presented much smaller regression slopes, with the aliases field having a regression slope of 0.02 and both age and location fields being already present in linked profiles of size 2, therefore showing no growth.

Concerning past locations, Figure 38 indicates a regression slope of 0.83, which is lower than the regression slope for relations (0.90). However, the quantity of past locations almost quadrupled from a profile of size 2 to a profile of size 10, compared to the quantity of relatives, which only doubled.

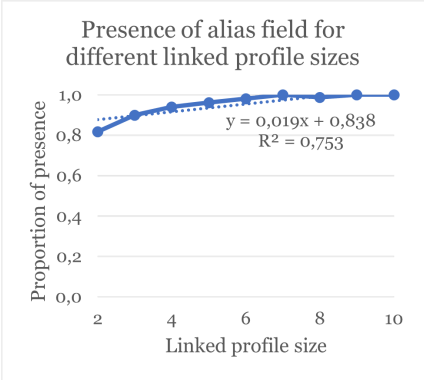
We observe that since most non-sensitive information was already present in smaller linked profiles, it was not more present in bigger sized ones. The only exception is in the gender field, which was considered as non-sensitive information in our study due to this information often being easy to deduce from a person's name, even though it could be considered as sensitive information in other contexts. This field was less frequent in smaller linked profiles, but was almost always present in bigger linked profiles. This was due to person search sites not considering gender as an important field (most people search for persons of whom they know the gender, which is therefore not a good discriminant to identify the right person to search for), but when a data broker does decide to provide this



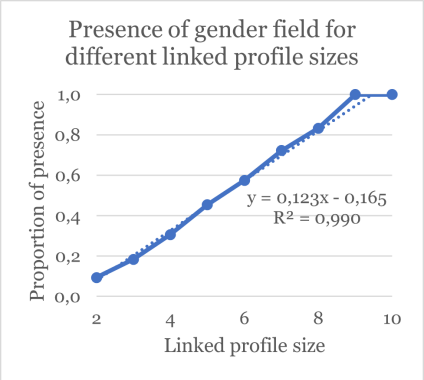
(a) Age



(b) Location



(c) Aliases



(d) Gender

Figure 37: Presence of non-sensitive information for different linked profile sizes

information, the field was often present. Still, on average, non-sensitive information was more present in linked profiles of all sizes.

5.4.3 Comparison between sensitive and non-sensitive information

The average slope for sensitive information was 0.042, compared to an average slope of 0.036 for non-sensitive information. Furthermore, the median slope for sensitive information was 0.036, while the median slope for non-sensitive information was 0.019. Therefore, even with the gender field boosting the average slope for non-sensitive information, we can still conclude that an increase in the size of linked profiles impacts more the presence of sensitive information than the presence of non-sensitive information.

5.5 Findings

We notice that although data brokers share some fields with high presence of information in common, the presence of other fields, and the presence of information in those fields, vary across data brokers. This makes data brokers different regarding the profile information that they are expected to provide. Therefore, different data brokers with different

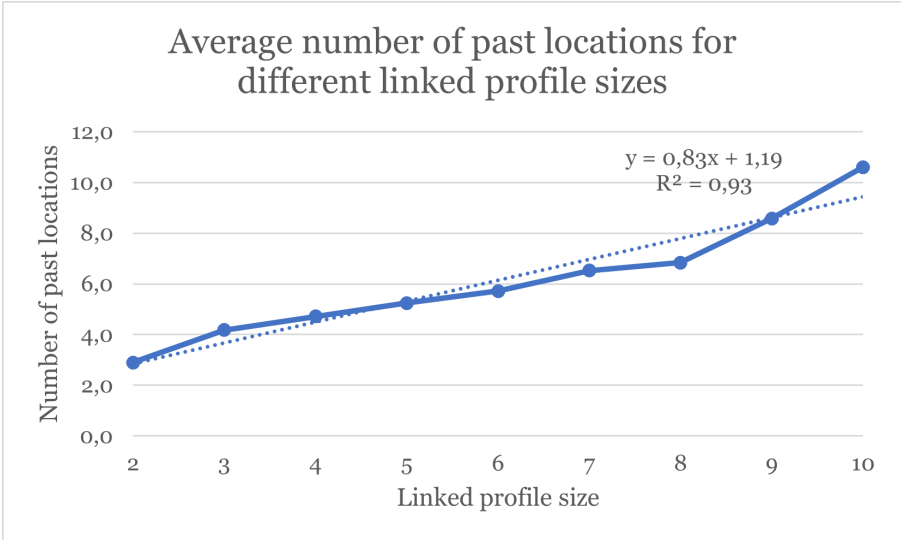


Figure 38: Number of past locations for different linked profile sizes

information can complement each other and fill in the absence of some fields when profiles are linked across data brokers.

In general, the presence of information fields will always grow if the linked profile size is bigger. For different names, we also computed the presence of fields for the individual with the biggest linked profile in Figure 39. We can observe that many common fields (age, location, aliases, phones, addresses, email and gender) achieve high presence in the biggest linked profile, and very sensitive information (work, licenses, bankruptcies and education) have lower but still non-negligible presence. On average, 7.4 past locations and 11.6 relatives' names were identified for the biggest linked profile. Therefore, we can conclude that a person search using a name can uncover detailed profiles about individuals, whether or not we know the person whose name was requested. This could provide opportunities for phishing, harpooning and scams, as knowledge of personal information cannot guarantee an individual's identity anymore.

Also, we observe that sensitive information was more impacted by the increase in size of a linked profile than non-sensitive information, with contact information (phone numbers, addresses and email) being the most affected fields. This is due to the variation in availability of this information across different data broker websites, which makes this information less easy to find in smaller linked profiles and more present in bigger linked profiles. Sensitive information from which discrimination could be made (work, licenses, bankruptcies and education) were not highly present, but their presence remain non-negligible as divulgation of those information can greatly affect an individual's reputation. Also, though we were not able to retrieve this information automatically, some person search sites provided very sensitive information such as religion, ethnicity or political beliefs, which should be handled cautiously due to the numerous potential prejudices associated with this information. In general, for many fields, information was often present in the biggest linked profile for a queried name. Therefore, we should worry about how easily personal information can be acquired by strangers.

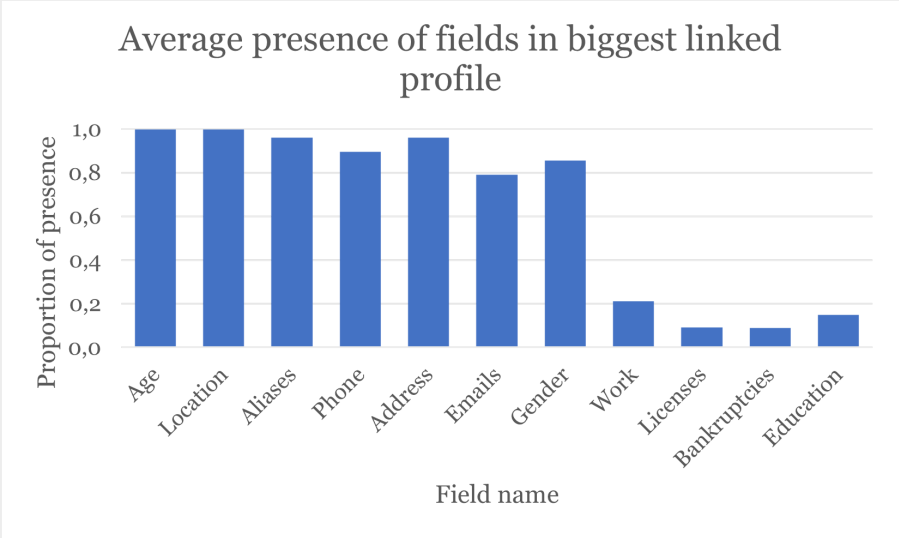


Figure 39: Presence of fields in biggest linked profile

6 Discussion and conclusions

This work examined the industry of data brokers by exploring its diverse characteristics and focused on person search sites to provide readers insights about how easily their personal information can be found on the internet. This was achieved by implementing the DROPLET system, which enabled us to explore the limits of the amount of personal information that could be easily and automatically gathered through exploitation of person search sites. The experiment showed that by using the DROPLET system, an action as easy as inputting a person's name can return a phenomenal quantity of results that grow faster than ripples propagating on a water surface. The sea of user information that can be found online is very deep and free services provided by person search sites only scratch the surface of the amount of personal information strangers can access without the concerned individual's knowledge. This information can be used for ill intentions such as fraud or identity theft, causing harm to the individual whose information is being divulged.

Furthermore, our research shows that important amounts of information can be retrieved from person search sites at almost no cost, and that the amount of sensitive information present in linked profiles are particularly impacted by the number of profiles grouped in the linked profile.

To mitigate personal data being divulged online, individuals can avoid voluntarily providing personal information to online services, or add noise to their data to make profile linking more difficult. Providing information about relatives should also be avoided, as this not only put their own personal data in danger, but also personal data of other people. Also, opting out of data brokers' lists can erase the individual's personal profile from search results, guaranteeing that no stranger with ill intentions will find those data easily.

However, there is a limit to what individuals can do. As data are often collected without their knowledge, it is hard for people to identify new data brokers of which they have not opted out. Furthermore, data brokers also collect data from public records, which cannot be erased. Therefore, governments should work with data brokers on protecting

personal privacy by proposing regulations that ensure privacy rights, and by following existent regulations by providing adequate data control measures to customers. In the future, we should expect further regulations providing individuals additional rights and control on their data, for current rights only protects against data leakage from data brokers we know of, but not against new data brokers who are appearing faster than our data erasure process can follow. That is, though we can request erasure by specifically contacting one data broker, our data would not be erased for new data brokers we haven't heard of.

In order to limit the collection of personal data without a user's consent, technologies and regulations should be developed to allow users to provide their consent or refusal to data collection at the same time as those information are accessed by a data aggregator. For instance, most browsers support sending a Do Not Track (DNT) signal when requesting websites, informing companies about the user's choices regarding privacy. However, many data brokers websites do not honour DNT signals, as specified in their privacy policy, due to the absence of legal regulations regarding how those signals should be treated.

For this reason, the Global Privacy Control (GPC) signal was developed with the intent of being supported by regulations such as GDPR or CCPA and providing clear specifications as to how companies should respond to this signal. Development of the GPC specification is still in progress, but it is expected to receive more and more support in the future.

Individuals, governments and industries should pay further attention to the divulgation of sensitive personal information online, and how easy those information can be accessed by ill-intentioned individuals. Also, personal information should be handled with care, and further measures to protect its safety are expected.

Acknowledgements

The authors of this paper have benefited from scientific discussions with Rim Ben Salem, Zakaria Sahnoune and Jian-Yun Nie, providing ideas used in the methodology employed in the current paper.

References

[1] AB-1202 Privacy: Data brokers 2019 (Cal) (U.S.). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB1202

[2] Aïmeur, E., Brassard, G., & Molins, P. (2012). Reconstructing profiles from information disseminated on the Internet. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 875-883. https://doi.org/10.1109/SocialCom-PASSAT.2012.38

[3] Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, *17*(18), 6937. https://doi.org/10.3390/ijerph17186937

[4] Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019, November 15). *Americans and Privacy: Concerned, Confused and Feeling Lack of Control over their Personal Information*. Pew Research Center. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

[5] Baik, J. (2020). Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA). *Telematics and Informatics*, *52*, 101431. https://doi.org/10.1016/j.tele.2020.101431

[6] Bulao, J. (2021, August 6). *How Much Data Is Created Every Day in 2021?* TechJury. https://techjury.net/blog/how-much-data-is-created-every-day/

[7] Christen, P. (2007). A two-step classification approach to unsupervised record linkage. *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, *70*, 111-119.

[8] Christen, P. (2012). *Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2

[9] Cialdini, R. (2009). *Influence: The Psychology of Persuasion*. Harper Collins.

[10] Crain, M. (2016). The limits of transparency: Data brokers and commodification. *New Media & Society*, *20*(1), 88-104. https://doi.org/10.1177/1461444816657096

[11] Federal Trade Commission. (2014). *Data Brokers : A Call for Transparency and Accountability*. https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

[12] Geronimo, M. (2017). Online browsing: Can, should, and may companies combine online and offline data to learn about you. *Hastings Science and Technology Law Journal*, *9*(2), 211-246.

[13] Green, D. (2017-2018). Big Brother is listening to you: Digital eavesdropping in the advertising industry. *Duke Law & Technology Review*, *16*, 352-392.

[14] Helveston, M. N. (2019). Reining in commercial exploitation of consumer data. *Penn State Law Review*, *123*(3), 667-702.

[15] Insurance Information Institute. (n.d.). *Facts + Statistics: Identity Theft and Cybercrime*. Retrieved August 24, 2021, from https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime

[16] Kruse, F., Hassan, A. P., Awick, J., & Gómez, J. M. (2020). A qualitative literature review on linkage techniques for data integration. *Hawaii International Conference on System Sciences*. http://hdl.handle.net/10125/63871

[17] Kuempel, A. (2016). The invisible middlemen: critique and call for reform of the data broker industry. *Northwestern Journal of International Law & Business*, *36*(1), 207-234.

[18] Linden, T., Khandelwal, R., Harkous, H. & Fawaz, K. (2020). The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies*, *2020*(1), 47-64. https://doi.org/10.2478/popets-2020-0004

[19] Office of the Privacy Commissioner of Canada. (2014, October 10). *Data Brokers: A Look at the Canadian and American Landscape*. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2014/db_201409/

[20] Palk, L., & Muralidhar, K. (2018). A free ride: Data brokers' rent-seeking behavior and the future of data inequality. *Vanderbilt Journal of Entertainment & Technology Law*, *20*(3), 779-838.

[21] Rostow, T. (2017). What happens when an acquaintance buys your data: New privacy harm in the age of data brokers. *Yale Journal on Regulation*, *34*(2), 667-708.

[22] Russell, N., Reidenberg, J. R., Martin, E., & Norton, T. B. (2019). Transparency and the marketplace for student data. *Virginia Journal of Law & Technology*, *22*(2), 107-157.

[23] Schneider, A. (2015). How could they know that behind the data that facilitates scams against vulnerable Americans. *Virginia Journal of Law & Technology*, *19*(3), 716-769.

[24] Shu, K., Wang, S., Tang, J., Zafarani, R., Liu, H. (2016). User identity linkage across online social networks: A review. *ACM SIGKDD Explorations Newsletter*, *18*(2), 5-17. https://doi.org/10.1145/3068777.3068781

[25] Solove, D. (2011). *Nothing to Hide: The False Tradeoff Between Privacy and Security*. Yale University Press.

[26] State of California - Department of Justice - Office of the Attorney General. (2021, July 14). *California Consumer Privacy Act (CCPA)*. https://oag.ca.gov/privacy/ccpa

[27] State of California - Department of Justice - Office of the Attorney General. (n.d.). *Data Broker Registry*. Retrieved August 25, 2021, from https://oag.ca.gov/data-brokers

[28] U.S. Census Bureau (2016). *Frequently Occurring Surnames in the 2010 Census* [Data set]. U.S. Census Bureau. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[29] U.S. Census Bureau (2020). *Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 2020 Census* [Data set]. U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/data/apportionment/apportionment-2020-table02.xlsx

[30] U.S. Census Bureau (2021). *Subcounty Resident Population Estimates: April 1, 2010 to July 1, 2019; April 1, 2020; and July 1, 2020* [Data set]. U.S. Census Bureau. https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/cities/

[31] U.S. Social Security Administration (2021). *National Data on the Relative Frequency of Given Names in the Population of U.S. Births where the Individual Has a Social Security Number* [Data set]. U.S. Social Security Administration. https://www.ssa.gov/oact/babynames/limits.html

[32] Venkatadri, G., Andreou, A., Liu, Y., Mislove, A., Gummadi, K. P., Loiseau, P. & Goga, O. (2018). Privacy risks with Facebook's PII-based targeting: Auditing a data broker's advertising interface. *2018 IEEE Symposium on Security and Privacy (SP)*, 89-107. https://www.doi.org/10.1109/SP.2018.00014

[33] West, S. M. (2017). Data capitalism: Redefining the logics of surveillance and privacy. *Business & Society*, *58*(1), 20-41. https://doi.org/10.1177/0007650317718185

[34] WordCounter (2015, November 5). *Word Counter Reading Level Feature*. Retrieved August 29, 2021, from https://wordcounter.net/blog/2015/11/05/10805_writing-reading-level-tool.html

[35] Yeh, C. (2018). Pursuing consumer empowerment in the age of big data: A comprehensive regulatory framework for data brokers. *Telecommunications Policy*, *42*(4), 282-292. https://doi.org/10.1016/j.telpol.2017.12.001